Genomic DNA transposition induced by human PGBD5

Anton G. Henssen[a], Elizabeth Henaff[b], Eileen Jiang[a], Amy R. Eisenberg[a], Julianne R. Carson[a], Camila M. Villasante[a], Mondira Ray[a], Eric Still[a], Melissa Burns[c], Jorge Gandara[b], Cedric Feschotte[d], Christopher E. Mason[b], Alex Kentsis [a,e,1]

[a] Molecular Pharmacology Program, Sloan Kettering Institute, New York, NY;

[b] Institute for Computational Biomedicine, Weill Cornell Medical College, New York, NY;

[c] Boston Children's Hospital, Boston, MA;

[d] Department of Human Genetics, University of Utah School of Medicine, Salt Lake City, UT;

[e] Department of Pediatrics, Memorial Sloan Kettering Cancer Center, Weill Cornell Medical College, New York, NY 10065.

[1] To whom correspondence may be addressed: Email: kentsisresearchgroup@gmail.com

Author contributions: A.H. and A.K. designed the research, A.H., E.H., A.E., E.J., M.B., J.R.C, C.V. performed the research, A.H., E.H., C.F., C.E.M., A.K. analyzed data, A.H. and A.K. wrote the manuscript in consultation with all authors.

The authors declare no conflict of interest.

Data deposition: The data reported in this manuscript have been deposited to the Sequence Read Archive (SRA), http://www.ncbi.nlm.nih.gov/sra/ (accession number SRP061649).

This article contains supporting information online.

Abstract

Transposons are mobile genetic elements that are found in nearly all organisms, including humans. Mobilization of DNA transposons by transposase enzymes can cause genomic rearrangements, but our knowledge of human genes derived from transposases is limited. Here, we find that the protein encoded by human *PGBD5*, the most evolutionarily conserved transposable element-derived gene in chordates, can induce stereotypical cut-and-paste DNA transposition in human cells. Genomic integration activity of PGBD5 requires distinct aspartic acid residues in its transposase domain, and specific DNA sequences with inverted terminal repeats with similarity to *piggyBac* transposons. DNA transposition catalyzed by PGBD5 in human cells occurs genome-wide, with precise transposon excision and preference for insertion at TTAA sites. The apparent conservation of DNA transposition activity by PGBD5 raises the possibility that genomic remodeling may contribute to its biological function.

Introduction

Transposons are genetic elements that are found in nearly all living organisms (1). They can contribute to the developmental and adaptive regulation of gene expression and are a major source of genetic variation that drives genome evolution (2). In humans and other mammals, they comprise about half of the nuclear genome (3). The majority of primate-specific sequences that regulate gene expression are derived from transposons (4), and transposons are a major source of structural genetic variation in human populations (5).

While the majority of genes that encode transposase enzymes tend to become catalytically inactive and their transposon substrates tend to become immobile in the course of organismal evolution, some can maintain their transposition activities (6, 7). In humans, at least one hundred L1 long interspersed repeated sequences (LINEs) actively transpose in human genomes and induce structural variation (8), including somatic rearrangements in neurons that may contribute to neuronal plasticity (9). The human *Transib*-like transposase RAG1 catalyzes somatic recombination of the V(D)J receptor genes in lymphocytes, and is essential for adaptive immunity (10). The *Mariner*-derived transposase SETMAR functions in single-stranded DNA resection during DNA repair and replication in human cells (11).

Among transposase enzymes that can catalyze excision and insertion of transposon sequences, DNA transposases are distinct in their dependence only on the availability of competent genomic substrates and cellular repair enzymes that ligate and repair excision sites, as compared to retrotransposons which require transcription of the mobilized sequences (12). Most DNA transposases utilize an RNase H-like domain with three aspartate or glutamate residues (so-called DDD or DDE motif) that catalyze magnesium-dependent hydrolysis of phosphodiester bonds and strand exchange (13-15). The IS4 transposase family, which includes piggyBac

transposases, is additionally distinguished by precise excisions without modifications of the transposon flanking sequences (16). The piggyBac transposase and its transposon were originally identified as an insertion in lepidopteran *Trichoplusia ni* cells (17). The *piggyBac* transposon consists of 13-bp inverted terminal repeats (ITR) and 19-bp subterminal inverted repeats located 3 and 31 base pairs from the 5' and 3' ITRs, respectively (18). PiggyBac transposase can mobilize a variety of ITR-flanked sequences and has a preference for integration at TTAA target sites in the host genome (15, 18-23).

Members of the piggyBac superfamily of transposons have colonized a wide range of organisms (24), including a recent and likely ongoing invasion of the bat *M. lucifugus* (25). The human genome contains 5 paralogous genes derived from *piggyBac* transposases, *PGBD1-5* (24, 26). *PGBD1* and *PGBD2* invaded the common mammalian ancestor, and *PGBD3* and *PGBD4* are restricted to primates, but are all contained as single coding exons, fused in frame with endogenous host genes, such as the Cockayne Syndrome B gene (CSB-PGBD3)-PGBD3 fusion (24, 27). Thus far, only the function of *PGBD3* has been investigated in some detail. CSB-PGBD3 is capable of binding DNA, including endogenous *piggyBac*-like transposons in the human genome, but has no known catalytic activity, though biochemical and genetic evidence indicates that it may participate in DNA damage response (28, 29). *PGBD5* is distinct from other human *piggyBac*-derived genes by having been domesticated much earlier in vertebrate evolution approximately 500 million years (My) ago, in the common ancestor of cephalochordates and vertebrates (24, 30). *PGBD5* is transcribed as a multi-intronic but non-chimeric transcript predicted to encode a full-length transposase (30). Furthermore, *PGBD5* expression in both human and mouse appears largely restricted to the early embryo and certain

areas of the embryonic and adult brain (24, 30). These intriguing features prompted us to investigate whether human PGBD5 has retained the enzymatic capability of mobilizing DNA.

Results

Human *PGBD5* contains a C-terminal RNase H-like domain that has approximately 20% sequence identity and 45% similarity to the active lepidopteran *piggyBac*, ciliate *piggyMac*, and mammalian *piggyBat* transposases (Fig. 1A and S1) (24, 25, 31). We reasoned that even though the ancestral transposon substrates of PGBD5 cannot be predicted due to its very ancient evolutionary origin (~500 My), preservation of its transposase activities should confer residual ability to mobilize distantly related *piggyBac*-like transposons. To test this hypothesis, we used a synthetic transposon reporter PB-EF1-NEO comprised of a neomycin resistance gene flanked by *T. ni piggyBac* ITRs (Fig. 1B) (20, 32). We transiently transfected human embryonic kidney (HEK) 293 cells, which lack endogenous *PGBD5* expression with the PB-EF1-NEO transposon reporter plasmid in the presence of a plasmid expressing PGBD5, and assessed genomic integration of the reporter using clonogenic assays in the presence of G418 to select cells with genomic integration conferring neomycin resistance (Fig. 1C, Fig. S2). Given the absence of suitable antibodies to monitor PGBD5 expression, we expressed PGBD5 as an N-terminal fusion with the green fluorescent protein (GFP). We observed significant rates of neomycin resistance of cells conferred by the transposon reporter with GFP-PGBD5, but not in cells expressing control GFP or mutant GFP-PGBD5 lacking the transposase domain (Fig. 1C), despite equal expression of all transgenes (Fig. S3). The efficiency of neomycin resistance conferred by the transposon reporter with GFP-PGBD5 was approximately 4.5-fold less than that of the *T. ni* piggyBac-derived transposase (Fig. 1D), consistent with their evolutionary divergence. These results suggest that human PGBD5 can promote genomic integration of a *piggyBac*-like transposon.

If neomycin resistance conferred by the PGBD5 and the transposon reporter is due to genomic integration and DNA transposition, then this should require specific activity on the transposon ITRs. To test this hypothesis, we generated transposon reporters with mutant ITRs and assayed them for genomic integration (Fig. 1B & S4). DNA transposition by the piggyBac family transposases involves hairpin intermediates with a conserved 5'-GGGTTAACCC-3' sequence that is required for target site phosphodiester hydrolysis (15). Thus, we generated reporter plasmids lacking ITRs entirely or containing complete ITRs with 5'-ATATTAACCC-3' mutations predicted to disrupt the formation of productive hairpin intermediates (15). To enable precise quantitation of mobilization activity, we developed a quantitative genomic PCR assay using primers specific for the transposon reporter and the endogenous human *TK1* gene for normalization (Fig. S5, S6). In agreement with the results of the clonogenic neomycin resistance assays, we observed efficient genomic integration of the donor transposons in cells transfected by GFP-PGBD5 as compared to the minimal signal observed in cells expressing GFP control (Fig. 1E). Deletion of transposon ITRs from the reporter reduced genomic integration to background levels (Fig. 1E). Consistent with the specific function of *piggyBac* family ITRs in genomic transposition, mutation of the terminal GGG sequence in the ITR significantly reduced the integration efficiency (Fig. 1E). These results indicate that specific transposon ITR sequences are required for PGBD5-mediated DNA transposition.

DNA transposition by piggyBac superfamily transposases is distinguished from most other DNA transposon superfamilies by the precise excision of the transposon from the donor site and preference for insertion in TTAA sites (20, 32). To determine the structure of the donor sites of transposon reporters mobilized by PGBD5, we isolated plasmid DNA from cells two days after transfection, amplified the transposon reporter using PCR, and determined its

sequence using capillary Sanger sequencing (Fig. S7). Similar to the hyperactive *T. ni* piggyBac, cells expressing GFP-PGBD5, but not those expressing GFP control vector, exhibited robust excision of ITR-flanked transposon with apparently precise repair of the donor plasmid (Fig. 2A, 2B & S7). These results suggest that PGBD5 is an active cut-and-paste DNA transposase.

To validate chromosomal integration and determine the location and precise structure of the insertion of the reporter transposons in the human genome, we isolated genomic DNA from G418-resistant HEK293 cells following transfection with PGBD5 and PB-EF1-NEO, and amplified the genomic sites of transposon insertions using flanking-sequence exponential anchored (FLEA) PCR, a technique originally developed for high-efficiency analysis of retroviral integrations (33). We adapted FLEA-PCR for the analysis of genomic DNA transposition by using unique reporter sequence to prime polymerase extension upstream of the transposon ITR into the flanking human genome, followed by reverse linear extension using degenerate primers, and exponential amplification using specific nested primers to generate chimeric amplicons suitable for massively parallel single-molecule Illumina DNA sequencing (Fig. S8) (34). This method enabled us to isolate specific portions of the human genome flanking transposon insertions, as evidenced by the reduced yield of amplicons isolated from control cells lacking transposase vectors or expressing GFP (Fig. S9). To identify the sequences of the transposon genomic insertions at single base pair resolution, we aligned reads obtained from FLEA-PCR Illumina sequencing to the human hg19 reference genome and synthetic transposon reporter, and identified split reads that specifically span both (Fig. S8). These data have been deposited to the Sequence Read Archive (SRA), http://www.ncbi.nlm.nih.gov/sra/ (accession number SRP061649).

To infer the mechanism of genomic integration of transposon reporters, we analyzed the sequences of the insertion loci to determine integration preferences at base pair resolution and identify potential sequence preferences. We found that transposon amplicons isolated from cells expressing GFP-PGBD5, but not those isolated from GFP control cells, were significantly enriched for TTAA sequences, as determined by sequence entropy analysis (35) (Fig. 2C). To discriminate between potential DNA transposition at TTAA target sites and alternative mechanisms of chromosomal integration, we classified genomic insertions based on target sites containing TTAA and those containing other sequence motifs (Table 1). Consistent with the DNA transposition activity of PGBD5, we observed significant induction of TTAA-containing insertions in cells expressing GFP-PGBD5 and transposons with intact ITRs, as compared to control cells expressing GFP, or to cells transfected with GFP-PGBD5 and mutant ITR transposons (Table 1). Sequence analysis of split reads containing transposon-human junction at TTAA sites revealed that, in each case examined (n = 65), joining between TTAA host and transposon DNA occurred precisely at the GGG/CCC terminal motif of the donor transposon ITR (Fig. 2E), in agreement with its requirement for efficient DNA transposition (Fig. 1E). Consistent with the genome-wide transposition induced by PGBD5, we identified transposition events in all human chromosomes, including both genic and intergenic loci (Fig. 2D). Thus, PGBD5 can mediate canonical cut-and-paste DNA transposition of *piggyBac* transposons in human cells.

Requirement for transposon substrates with specific ITRs, their precise excision and preferential insertion into TTAA-containing genomic locations are all consistent with the preservation of PGBD5's DNA transposase activity in human cells. Like other cut-and-paste transposases, PiggyBac superfamily transposases are thought to utilize a triad of aspartate or

glutamate residues to catalyze phosphodiester bond hydrolysis, but the catalytic triad of aspartates previously proposed for *T. ni* piggyBac is apparently not conserved in the primary sequence of PGBD5 (Fig. S1) (14, 15, 24, 36). Thus, we hypothesized that distinct aspartic or glutamic acid residues may be required for DNA transposition mediated by PGBD5. To test this hypothesis, we used alanine scanning mutagenesis and assessed transposition activity of GFP-PGBD5 mutants using quantitative genomic PCR (Fig. 3 & S10). This analysis indicated that simultaneous alanine mutations of D168, D194, and D386 reduced apparent transposition activity to background levels, similar to that of GFP control (Fig. 3). We confirmed that the mutant GFP-PGBD5 proteins have equivalent stability and expression as the wild-type protein in cells by immunoblotting against GFP (Fig. 3B). These results suggest that PGBD5 represents a distinct member of the piggyBac family of DNA transposases.

Discussion

Our current findings indicate that human PGBD5 is an active *piggyBac* transposase that can catalyze DNA transposition in human cells. DNA transposition by PGBD5 requires its C-terminal transposase domain, and depends on specific inverted terminal repeats derived from the lepidopteran *piggyBac* transposons (Fig. 1). DNA transposition involves trans-esterification reactions mediated by DNA hairpin intermediates (15). Consistent with the requirement of intact termini of the piggyBac, Tn10, and Mu transposons (18), elimination or mutation of the terminal GGG nucleotides from the transposon substrates also abolishes the transposition activity of PGBD5 (Fig. 1). PGBD5-induced DNA transposition is precise with preference for insertions at TTAA genomic sites (Fig. 2). Since our analysis was limited to ectopically expressed PGBD5 fused to GFP and episomal substrates derived from lepidopteran *piggyBac* transposons, it is

possible that endogenous PGBD5 may exhibit different activities on chromatinized substrates in the human genome.

Current structure-function analysis indicates that PGBD5 requires three aspartate residues to mediate DNA transposition (Fig. 3), but its DDD domain appears to be distinct from other piggyBac transposase enzymes with respect to primary sequence (Fig. S1) (14). Thus, the three aspartate residues required for efficient DNA transposition by PGBD5 may form a catalytic triad that functions in phosphodiester bond hydrolysis, similar to the DDD motif in other *piggyBac* family transposases, or alternatively may contribute to other steps in the transposition reaction, such as synaptic complex formation, hairpin opening, or strand exchange (14, 15, 18). In addition, we find that alanine mutations of the three required aspartate residues in the PGBD5 transposase domain significantly reduce but do not completely eliminate genomic integration of the transposon reporters (Fig. 3). This could reflect residual catalytic activity despite these mutations, or that PGBD5 expression may affect other mechanisms of DNA integration in human cells.

The evolutionary conservation of the transposition activity of PGBD5 suggests that it may have hitherto unknown biologic functions among vertebrate organisms. DNA transposition is a major source of genetic variation that drives genome evolution, with some DNA transposases becoming extinct and others domesticated to evolve exapted functions. The evolution of transposons' activities can be highly variable, with some organisms such as *Z. mays* undergoing continuous genome remodeling and recent two-fold expansion through endogenous retrotransposition, *Drosophila* and *Saccharomyces* owing over half of their known spontaneous mutations to transposons, and primate species including humans exhibiting relative extinction of transposons (1).

Evolutionary conservation of transposase genes is generally interpreted as evidence of their biological function. However, these functions can undergo exaptation, with biochemical activities of transposase genes and their transposon substrates evolving to have endogenous functions other than genomic transposition *per se*. For example, human RAG1 is a domesticated *Transib* transposase that has retained its active transposase domain, and can transpose ITR-containing transposons *in vitro*, but catalyzes somatic recombination of immunoglobulin and T-cell receptor genes in lymphocytes across signal sequences that might be derived from related transposons (37, 38). Human SETMAR is a *Mariner*-derived transposase with a divergent DDN transposase domain that has retained its endonuclease but not transposition activity, and functions in double strand DNA repair by non-homologous end joining (7). The human genome encodes over 40 other genes derived from DNA transposases (1, 3), including *THAP9* that was recently found to mobilize transposons in human cells with as of yet unknown function (39). RAG1, THAP9 and PGBD5 are, to our knowledge, the only human proteins with demonstrated transposase activity.

The distinct biochemical and structural features of PGBD5 indicated by our findings are consistent with its unique evolution and function among human *piggyBac* derived transposase genes (24, 30). *PGBD5* exhibits deep evolutionary conservation predating the origin of vertebrates, including a preservation of genomic synteny across lancelet, lamprey, teleosts, and amniotes (30). This suggests that while PGBD5 likely derived from an autonomous mobile element, this ancestral copy was immobilized early in evolution and PGBD5 can probably no longer mobilize its own genomic locus, at least in germline cells. The human genome contains several thousands of miniature inverted repeat transposable elements (MITE) with similarity to *piggyBac* transposons (1, 24). CSB-PGBD3 can bind to the *piggyBac*-derived *MER85* elements

in the human genome (28, 29). Similarly, it is possible that PGBD5 can act *in trans* to recognize and mobilize one or several related MITEs in the human genome. Recently, single-molecular maps of the human genome have predicted thousands of mobile element insertions, and the activity of PGBD5 or other endogenous transposases may explain some of these novel variants (40, 41). *PGBD5* localizes to the cell nucleus, and is expressed during embryogenesis and neurogenesis, but its physiological function is not known (30).

Given that both human RAG1 and ciliate piggyMac domesticated transposases catalyze the elimination of specific genomic DNA sequences (10, 31), it is reasonable to hypothesize that PGBD5's biological function may similarly involve the excision of as of yet unknown ITR-flanked sequences in the human genome or another form of DNA recombination. Since DNA transposition by *piggyBac* family transposases requires substrate chromatin accessibility and DNA repair, we anticipate that additional cellular factors are required for and regulate PGBD5 functions in cells. Likewise, just as RAG1-mediated DNA recombination of immunoglobulin loci is restricted to B lymphocytes, and rearrangements of T-cell receptor genes to T lymphocytes, potential DNA rearrangements mediated by PGBD5 may be restricted to specific cell types and developmental periods. Generation of molecular diversity through DNA recombination during nervous system development has been a long-standing hypothesis (42, 43). The recent discovery of somatic retrotransposition in human neurons (44-46), combined with our finding of DNA transposition activity by human PGBD5, which is highly expressed in neurons, suggest that additional mechanisms of somatic genomic diversification may contribute to vertebrate nervous system development.

Because DNA transposition is inherently topological and orientation of transposons can affect the arrangements of reaction products (47), potential activities of PGBD5 can depend on

the arrangements of accessible genomic substrates, leading to both conservative DNA transposition involving excision and insertion of transposon elements, as well as irreversible reactions such as DNA elimination and chromosomal breakage-fusion-bridge cycles, as originally described by McClintock (48). Finally, given the potentially mutagenic activity of active DNA transposases, we anticipate that unlicensed activity of PGBD5 and other domesticated transposases can be pathogenic in specific disease states, particularly in cases of aberrant chromatin accessibility, such as cancer.

Materials and Methods

Reagents

All reagents were obtained from Sigma Aldrich if not otherwise specified. Synthetic oligonucleotides were obtained from Eurofins MWG Operon (Huntsville, AL, USA) and purified by HPLC.

Cell culture

HEK293 and HEK293T were obtained from the American Type Culture Collection (ATCC, Manassas, Virginia, USA). The identity of all cell lines was verified by STR analysis and lack of *Mycoplasma* contamination was confirmed by Genetica DNA Laboratories (Burlington, NC, USA). Cell lines were cultured in DMEM supplemented with 10% fetal bovine serum and 100 U / ml penicillin and 100 µg / ml streptomycin in a humidified atmosphere at 37 ˚C and 5% CO_2.

Plasmid constructs

Human PGBD5 cDNA (Refseq ID: NM_024554.3) was cloned as a GFP fusion into the lentiviral vector pReceiver-Lv103-E3156 (GeneCopoeia, Rockville, MD, USA). *piggyBac* inverted terminal repeats (5'TTAACCCTAGAAAGATAATCATATTGTGACGTACGTTAAAGATAATCATGTGTAA AATTGACGCATG3' and 5'CATGCGTCAATTTTACGCAGACTATCTTTCTAGGGTTAA3'), as originally cloned by Malcolm Fraser and colleagues (18, 23), were cloned into PB-EF1-NEO to flank IRES-driven neomycin resistance gene, as obtained from System Biosciences (Mountain View, CA, USA). Plasmid encoding the hyperactive *T. ni* piggyBac transposase, as originally generated by Nancy

Craig and colleagues (49), was obtained from System Biosciences. Site-directed PCR mutagenesis was used to generate mutants of PGBD5 and *piggyBac*, according to manufacturer's instructions (Agilent, Santa Clara, CA, USA). Plasmids were verified by restriction endonuclease mapping and Sanger sequencing, and deposited in Addgene. Lentivirus packaging vectors psPAX2 and pMD2.G were obtained from Addgene (50).

Cell transfection

HEK293 cells were seeded at a density of 100,000 cells per well in a 6-well plate, and transfected with 2 μg of total plasmid DNA, containing 1 μg of transposon reporter (PB-EF1-NEO or mutants) and 1 μg of transposase cDNA (pRecLV103-GFP-PGBD5 or mutants) using Lipofectamine 2000, according to manufacturer's instructions (Life Technologies, CA, USA). After 24 hours, transfected cells were trypsinized and re-plated for functional assays.

Quantitative RT-PCR

Upon transfection, cells were cultured for 48 hours and total RNA was isolated using the RNeasy Mini Kit, according to manufacturer's instructions (Qiagen, Venlo, Netherlands). cDNA was synthesized using the SuperScript III First-Strand Synthesis System (Invitrogen, Waltham, MA, USA). Quantitative real-time PCR was performed using the KAPA SYBR FAST PCR polymerase with 20 ng template and 200 nM primers, according to the manufacturer's instructions (Kapa Biosystems, Wilmington, MA, USA). PCR primers are listed in Supp. Table 1. Ct values were calculated using ROX normalization using the ViiA 7 software (Applied Biosystems).

Neomycin resistance colony formation assay

Upon transfection, cells were seeded at a density of 1,000 cells per 10 cm dish and selected with G418 sulfate (2 mg/ml) for 2 weeks. Resultant colonies were fixed with methanol and stained with Crystal Violet.

Transposon excision assay

Upon transfection, cells were cultured for 48 hours and DNA was isolated using the PureLink Genomic DNA Mini Kit, according to manufacturer's instructions (Life technologies, Carlsbad, CA, USA). Reporter plasmid sequences flanking the neomycin resistance cassette transposons were amplified using hot start PCR with an annealing temperature of 57 ˚C and extension time of 2 minutes, according to the manufacturer's instructions (New England Biolabs, Beverly, MA, USA) using the Mastercycler Pro thermocycler (Eppendorf, Hamburg, Germany). PCR primers are listed in Supp. Table 1. The PCR products were resolved using agarose gel electrophoresis and visualized by ethidium bromide staining. Identified gel bands were extracted using the PureLink Quick Gel Extraction Kit (Invitrogen) and Sanger sequenced to identify excision products.

Quantitative PCR assay of genomic transposon integration

Upon transfection, cells were selected with puromycin (5 μg/ml) for 2 days to eliminate non-transfected cells. After selection, cells were expanded for 10 days without selection and genomic DNA isolated using PureLink Genomic DNA Mini Kit (Life technologies, Carlsbad, CA, USA). Quantitative real-time PCR was performed using the KAPA SYBR FAST PCR polymerase with 20 ng template and 200 nM primers, according to the manufacturer's instructions (Kapa Biosystems, Wilmington, MA, USA). PCR primers are listed in Supp. Table 1. Ct values were calculated using ROX normalization using the ViiA 7 software (Applied

Biosystems). We determined the quantitative accuracy of this assay using analysis of serial dilution PB-E1-NEO plasmid as reference (Fig. S6).

Flanking sequence exponential anchored (FLEA) PCR

To amplify genomic transposon integration sites, we modified flanking sequence exponential anchored (FLEA) PCR (33), as described in Supp. Fig. S8 (34). First, linear extension PCR was performed using 2 μg of genomic DNA and 100 nM biotinylated linear primer using the Platinum HiFidelity PCR mix, according to manufacturer's instructions (Invitrogen Corp., Carlsbad, CA, USA). Linear extension parameters for PCR were: 95 ˚C (45 sec), 62 ˚C (45 sec), 72 ˚C (3 min) for 30 cycles. Reaction products were purified by diluting the samples in a total volume of 200 μl of nuclease-free water and centrifugation using the Amicon Ultra 0.5 ml 100K at 12,000 *g* for 10 minutes at room temperature (EMD Millipore, Billerica, MA, USA) purification. Retentate was bound to streptavidin ferromagnetic beads on a shaker at room temperature overnight (Dynal, Oslo, Norway). Beads were washed with 40 μl of washing buffer (Kilobase binder kit; Dynal), then water, then 0.1 N NaOH and finally with water again.

To anneal the anchor primer, washed beads were resuspended in a total volume of 20 μl containing 5 μM anchor primer, 500 nM dNTP, and T7 DNA polymerase buffer (New England Biolabs, Beverly, MA, USA). Samples were placed in a heating block pre-heated to 85 ˚C, and allowed to passively cool to 37 ˚C. Once annealed, 10 units of T7 DNA polymerase (New England Biolabs) was added and the mixtures were incubated for 1 h at 37 °C. Next, the beads washed 5 times in water.

To exponentially amplify the purified products, beads were resuspended in a total volume of 50 μl containing 500 nM of exponential and Transposon1 primers, and the Platinum

HiFidelity PCR mix. PCR was performed with the following parameters: 95 °C for 5 min, followed by 35 cycles of 95 °C for 45 sec, 62 °C for 45 sec and 72 °C for 3 min. PCR products were purified using the Invitrogen PCR purification kit (Invitrogen Corp., Carlsbad, CA, USA). Second nested PCR was performed using $1/50^{th}$ of the first exponential PCR product as template using the Platinum HiFidelity PCR with 500 nM of exponential and Transposon2 primers. PCR was performed with the following parameters: 35 cycles of 95 °C for 45 sec, 62 °C for 45 sec and 72 °C for 3 min. Final PCR products were purified using the Invitrogen PCR purification kit, according to the manufacturer's instructions (Invitrogen Corp., Carlsbad, CA, USA).

Sequencing of transposon reporter integration sites

Equimolar amounts of purified FLEA-PCR amplicons were pooled, as measured using fluorometry with the Qubit instrument (Invitrogen Carlsbad, CA) and sized on a 2100 BioAanalyzer instrument (Agilent Technologies, Santa Clara, CA). The sequencing library construction was performed using the KAPA Hyper Prep Kit (KAPA biosystems, Wilmington, MA) and 12 indexed Illumina adaptors from IDT (Coralville, IO), according to the manufacturer's instructions.

After quantification and sizing, libraries were pooled for sequencing on a MiSeq (pooled library input at 10pM) on a 300/300 paired end run (Illumina, San Diego, CA). An average of 575,565 paired reads were generated per sample. The duplication rate varied between 56 and 87%. Because of the use of FLEA-PCR amplicons for DNA sequencing, preparation of Illumina sequencing libraries is associated with the formation of adapter dimers (51). We used cutadapt to first trim reads to retain bases with quality score > 20, then identify reads containing adapter dimers and exclude them from further analyses (parameters –q 20 -b P7=<P7_index> -B P5=<P5_index> –discard ; where <P7_index> is the P7 primer adapter with the specific barcode

for each library, and <P5_index> is the generic P5 adapter sequence: GATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCAT T (52). Anchor primer sequences were then trimmed from the reads retained using cutadapt (-g ^GTGGCACGGACTGATCNNNNNN). Filtered and trimmed reads were mapped to a hybrid reference genome consisting of the hg19 full chromosome sequences and the PB-EF1-NEO plasmid sequence using bwa-mem using standard parameters (53). Mapped reads were then analyzed with LUMPY using split read signatures (54), and insertion loci were identified using the called variants flagged as interchromosomal translocations (BND) between the plasmid sequence and the human genome. Breakpoints were resolved to base-pair accuracy using split read signatures when possible. Insertion loci were taken with 10 flanking base pairs and aligned with MUSCLE to establish consensus sequence (54). Genomic distribution of insertion loci were plotted using ChromoViz (https://github.com/elzbth/ChromoViz). All analysis scripts are available from Zenodo (http://dx.doi.org/10.5281/zenodo.22206).

Lentivirus production and cell transduction

Lentivirus production was carried out as described in (55). Briefly, HEK293T cells were transfected using TransIT with 2:1:1 ratio of the pRecLV103 lentiviral vector, and psPAX2 and pMD2.G packaging plasmids, according to manufacturer's instructions (TransIT-LT1, Mirus, Madison, WI). Virus supernatant was collected at 48 and 72 hours post-transfection, pooled, filtered and stored at -80 °C. HEK293T cells were transduced with virus particles at a multiplicity of infection of 5 in the presence of 8 μg/ml hexadimethrine bromide. Transduced cells were selected for 2 days with puromycin (5 μg/ml).

Western blotting

To analyze protein expression by Western immunoblotting, 1 million transduced cells were suspended in 80 μl of lysis buffer (4% sodium dodecyl sulfate, 7% glycerol, 1.25% beta-mercaptoethanol, 0.2 mg/ml Bromophenol Blue, 30 mM Tris-HCl, pH 6.8). Cells suspensions were lysed using Covaris S220 adaptive focused sonicator, according to the manufacturer's instructions (Covaris, Woburn, CA). Lysates were cleared by centrifugation at 16,000 g for 10 minutes at 4 ˚C. Clarified lysates (30 μl) were resolved using sodium dodecyl sulfate-polyacrylamide gel electrophoresis, and electroeluted using the Immobilon FL PVDF membranes (Millipore, Billerica, MA, USA). Membranes were blocked using the Odyssey Blocking buffer (Li-Cor), and blotted using the mouse and rabbit antibodies against GFP (1:500, clone 4B10) and β-actin (1:5000, clone 13E5), respectively, both obtained from Cell Signaling Technology (Beverly, MA). Blotted membranes were visualized using goat secondary antibodies conjugated to IRDye 800CW or IRDye 680RD and the Odyssey CLx fluorescence scanner, according to manufacturer's instructions (Li-Cor, Lincoln, Nebraska).

Statistical analysis

Statistical significance values were determined using two-tailed non-parametric Mann-Whitney tests for continuous variables, and two-tailed Fisher exact test for discrete variables.

Acknowledgments

We are grateful to Alejandro Gutierrez, Marc Mansour, Thomas Look, Charles Roberts, Daniel Bauer, Leo Wang, Hao Zhu and Michael Kharas for critical discussions, Nahum Meller for cloning assistance, and Alan Chramiec for technical support. This work was supported by the University of Essen Pediatric Oncology Research Program (A.H.), NIH K08 CA160660, Burroughs Wellcome Fund, CureSearch for Children's Cancer, and Hyundai Hope on Wheels

(A.K.), and the Irma T. Hirschl and Monique Weill-Caulier Charitable Trusts, the STARR Consortium, the Bert L and N Kuggie Vallee Foundation, and the WorldQuant Foundation (C.E.M.). We thank the MSKCC Integrated Genomics Core Facility and Bioinformatics Core Facility for assistance with DNA sequencing and analysis (NIH P30 CA008748).

References

1. Feschotte C & Pritham EJ (2007) DNA transposons and the evolution of eukaryotic genomes. *Annu Rev Genet* 41:331-368.
2. Cordaux R & Batzer MA (2009) The impact of retrotransposons on human genome evolution. *Nat Rev Genet* 10(10):691-703.
3. Smit AF (1999) Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr Opin Genet Dev* 9(6):657-663.
4. Jacques PE, Jeyakani J, & Bourque G (2013) The majority of primate-specific regulatory sequences are derived from transposable elements. *PLoS Genet* 9(5):e1003504.
5. Stewart C, *et al.* (2011) A comprehensive map of mobile element insertion polymorphisms in humans. *PLoS Genet* 7(8):e1002236.
6. Munoz-Lopez M & Garcia-Perez JL (2010) DNA transposons: nature and applications in genomics. *Curr Genomics* 11(2):115-128.
7. Liu D, *et al.* (2007) The human SETMAR protein preserves most of the activities of the ancestral Hsmar1 transposase. *Mol Cell Biol* 27(3):1125-1132.
8. Kazazian HH, Jr. (2004) Mobile elements: drivers of genome evolution. *Science* 303(5664):1626-1632.
9. Erwin JA, Marchetto MC, & Gage FH (2014) Mobile DNA elements in the generation of diversity and complexity in the brain. *Nat Rev Neurosci* 15(8):497-506.
10. Hiom K, Melek M, & Gellert M (1998) DNA transposition by the RAG1 and RAG2 proteins: a possible source of oncogenic translocations. *Cell* 94(4):463-470.
11. Shaheen M, Williamson E, Nickoloff J, Lee SH, & Hromas R (2010) Metnase/SETMAR: a domesticated primate transposase that enhances DNA repair, replication, and decatenation. *Genetica* 138(5):559-566.
12. Berg DE & Howe MM (1989) *Mobile DNA* (Am. Soc. Microbiol, Washington, DC).
13. Dyda F, Chandler M, & Hickman AB (2012) The emerging diversity of transpososome architectures. *Q Rev Biophys* 45(4):493-521.
14. Keith JH, Schaeper CA, Fraser TS, & Fraser MJ, Jr. (2008) Mutational analysis of highly conserved aspartate residues essential to the catalytic core of the piggyBac transposase. *BMC Mol Biol* 9:73.
15. Mitra R, Fain-Thornton J, & Craig NL (2008) piggyBac can bypass DNA synthesis during cut and paste transposition. *The EMBO journal* 27(7):1097-1109.
16. De Palmenaer D, Siguier P, & Mahillon J (2008) IS4 family goes genomic. *BMC Evol Biol* 8:18.
17. Fraser MJ, Brusca JS, Smith GE, & Summers MD (1985) Transposon-mediated mutagenesis of a baculovirus. *Virology* 145(2):356-361.
18. Elick TA, Lobo N, & Fraser MJ, Jr. (1997) Analysis of the cis-acting DNA elements required for piggyBac transposable element excision. *Mol Gen Genet* 255(6):605-610.
19. Beames B & Summers MD (1990) Sequence comparison of cellular and viral copies of host cell DNA insertions found in Autographa californica nuclear polyhedrosis virus. *Virology* 174(2):354-363.
20. Fraser MJ, Cary L, Boonvisudhi K, & Wang HG (1995) Assay for movement of Lepidopteran transposon IFP2 in insect cells using a baculovirus genome as a target DNA. *Virology* 211(2):397-407.

21. Wang HG & Fraser MJ (1993) TTAA serves as the target site for TFP3 lepidopteran transposon insertions in both nuclear polyhedrosis virus and Trichoplusia ni genomes. *Insect Mol Biol* 1(3):109-116.
22. Fraser MJ, Smith GE, & Summers MD (1983) Acquisition of Host Cell DNA Sequences by Baculoviruses: Relationship Between Host DNA Insertions and FP Mutants of Autographa californica and Galleria mellonella Nuclear Polyhedrosis Viruses. *J Virol* 47(2):287-300.
23. Handler AM, McCombs SD, Fraser MJ, & Saul SH (1998) The lepidopteran transposon vector, piggyBac, mediates germ-line transformation in the Mediterranean fruit fly. *Proc Natl Acad Sci U S A* 95(13):7520-7525.
24. Sarkar A, *et al.* (2003) Molecular evolutionary analysis of the widespread piggyBac transposon family and related "domesticated" sequences. *Mol Genet Genomics* 270(2):173-180.
25. Mitra R, *et al.* (2013) Functional characterization of piggyBat from the bat Myotis lucifugus unveils an active mammalian DNA transposon. *Proc Natl Acad Sci U S A* 110(1):234-239.
26. Smit AF & Riggs AD (1996) Tiggers and DNA transposon fossils in the human genome. *Proc Natl Acad Sci U S A* 93(4):1443-1448.
27. Newman JC, Bailey AD, Fan HY, Pavelitz T, & Weiner AM (2008) An abundant evolutionarily conserved CSB-PiggyBac fusion protein expressed in Cockayne syndrome. *PLoS Genet* 4(3):e1000031.
28. Bailey AD, *et al.* (2012) The conserved Cockayne syndrome B-piggyBac fusion protein (CSB-PGBD3) affects DNA repair and induces both interferon-like and innate antiviral responses in CSB-null cells. *DNA Repair (Amst)* 11(5):488-501.
29. Gray LT, Fong KK, Pavelitz T, & Weiner AM (2012) Tethering of the conserved piggyBac transposase fusion protein CSB-PGBD3 to chromosomal AP-1 proteins regulates expression of nearby genes in humans. *PLoS Genet* 8(9):e1002972.
30. Pavelitz T, Gray LT, Padilla SL, Bailey AD, & Weiner AM (2013) PGBD5: a neural-specific intron-containing piggyBac transposase domesticated over 500 million years ago and conserved from cephalochordates to humans. *Mob DNA* 4(1):23.
31. Baudry C, *et al.* (2009) PiggyMac, a domesticated piggyBac transposase involved in programmed genome rearrangements in the ciliate Paramecium tetraurelia. *Genes Dev* 23(21):2478-2483.
32. Cary LC, *et al.* (1989) Transposon mutagenesis of baculoviruses: analysis of Trichoplusia ni transposon IFP2 insertions within the FP-locus of nuclear polyhedrosis viruses. *Virology* 172(1):156-169.
33. Pule MA, *et al.* (2008) Flanking-sequence exponential anchored-polymerase chain reaction amplification: a sensitive and highly specific method for detecting retroviral integrant-host-junction sequences. *Cytotherapy* 10(5):526-539.
34. Henssen A, Carson J, & Kentsis A (2015) Transposon mapping using flanking sequence exponential anchored (FLEA) PCR.
35. Crooks GE, Hon G, Chandonia JM, & Brenner SE (2004) WebLogo: a sequence logo generator. *Genome Res* 14(6):1188-1190.
36. Nesmelova IV & Hackett PB (2010) DDE transposases: Structural similarity and diversity. *Adv Drug Deliv Rev* 62(12):1187-1195.

37. Fugmann SD, Lee AI, Shockett PE, Villey IJ, & Schatz DG (2000) The RAG proteins and V(D)J recombination: complexes, ends, and transposition. *Annu Rev Immunol* 18:495-527.
38. Landree MA, Wibbenmeyer JA, & Roth DB (1999) Mutational analysis of RAG1 and RAG2 identifies three catalytic amino acids in RAG1 critical for both cleavage steps of V(D)J recombination. *Genes Dev* 13(23):3059-3069.
39. Majumdar S, Singh A, & Rio DC (2013) The human THAP9 gene encodes an active P-element DNA transposase. *Science* 339(6118):446-448.
40. Pendleton M, *et al.* (2015) Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nature methods* 12(8):780-786.
41. Chaisson MJ, *et al.* (2015) Resolving the complexity of the human genome using single-molecule sequencing. *Nature* 517(7536):608-611.
42. Dreyer WJ, Gray WR, & Hood L (1967) The Genetics, Molecular, and Cellular Basis of Antibody Formation: Some Facts and a Unifying Hypothesis. *Cold Spring Harb Symp Quant Biol* 32:353-367.
43. Wu Q & Maniatis T (1999) A striking organization of a large family of human neural cadherin-like cell adhesion genes. *Cell* 97(6):779-790.
44. Coufal NG, *et al.* (2009) L1 retrotransposition in human neural progenitor cells. *Nature* 460(7259):1127-1131.
45. Evrony GD, *et al.* (2012) Single-neuron sequencing analysis of L1 retrotransposition and somatic mutation in the human brain. *Cell* 151(3):483-496.
46. Upton KR, *et al.* (2015) Ubiquitous L1 mosaicism in hippocampal neurons. *Cell* 161(2):228-239.
47. Claeys Bouuaert C, Liu D, & Chalmers R (2011) A simple topological filter in a eukaryotic transposon as a mechanism to suppress genome instability. *Mol Cell Biol* 31(2):317-327.
48. McClintock B (1942) The Fusion of Broken Ends of Chromosomes Following Nuclear Fusion. *Proc Natl Acad Sci U S A* 28(11):458-463.
49. Li X, *et al.* (2013) piggyBac transposase tools for genome engineering. *Proc Natl Acad Sci U S A* 110(25):E2279-2287.
50. Cudre-Mauroux C, *et al.* (2003) Lentivector-mediated transfer of Bmi-1 and telomerase in muscle satellite cells yields a duchenne myoblast cell line with long-term genotypic and phenotypic stability. *Hum Gene Ther* 14(16):1525-1533.
51. ILLUMINA (2015) Paired-End Sample Preparation Guide.
52. Lindgreen S (2012) AdapterRemoval: easy cleaning of next-generation sequencing reads. *BMC Res Notes* 5:337.
53. Li H & Durbin R (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* 26(5):589-595.
54. Layer RM, Chiang C, Quinlan AR, & Hall IM (2014) LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol* 15(6):R84.
55. Kentsis A, *et al.* (2012) Autocrine activation of the MET receptor tyrosine kinase in acute myeloid leukemia. *Nat Med* 18(7):1118-1122.

Figure 1

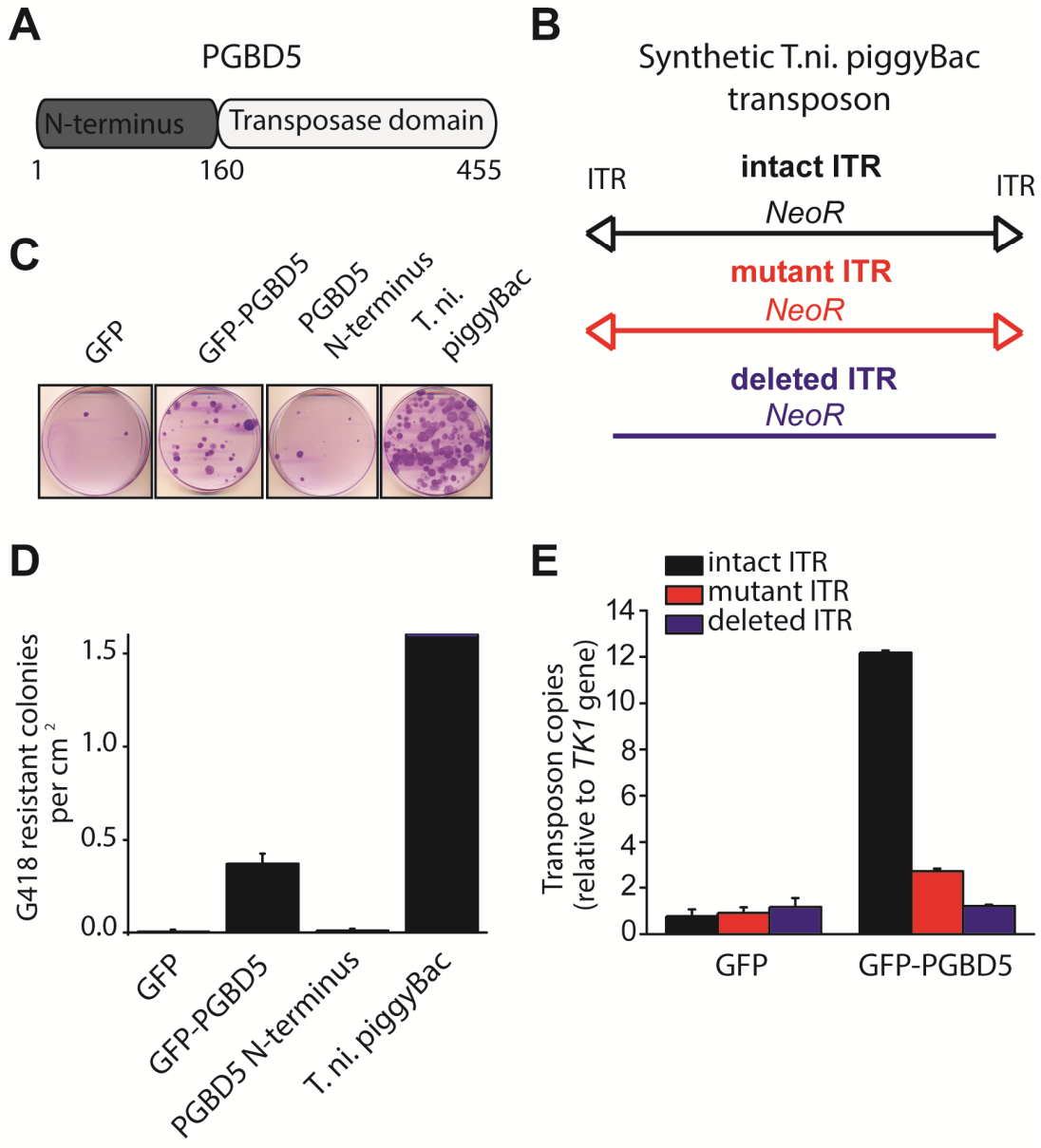

Figure 1. PGBD5 induces genomic integration of synthetic *piggyBac* transposons in human cells. (A) Schematic of the human PGBD5 protein with its C-terminal transposase homology domain, as indicated. **(B)** Schematic of synthetic transposon substrates used for DNA transposition assays, including transposons with mutant inverted terminal repeat (ITR) marked by triangles in red, and transposons lacking ITRs marked in blue. **(C)** Representative photographs of Crystal Violet-stained colonies obtained after G418 selection of HEK293 cells co-transfected with the transposon reporter plasmid along with transposase cDNA expression vectors. **(D)** Quantification of G418-selection clonogenic assays, demonstrating the integration activities of GFP-PGBD5, PGBD5 N-terminus, T.ni. piggyBac and GFP control (GFP-PGBD5 vs. GFP; $p = 0.00031$). **(E)** Quantification of genomic transposon integration using quantitative PCR of GFP-PGBD5 and GFP expressing cells using intact (black), mutant (red), and deleted (blue) ITR-containing transposon reporters (intact vs. mutant ITR; $p = 0.00011$). Error bars represent standard errors of 3 biologic replicates.

Figure 2

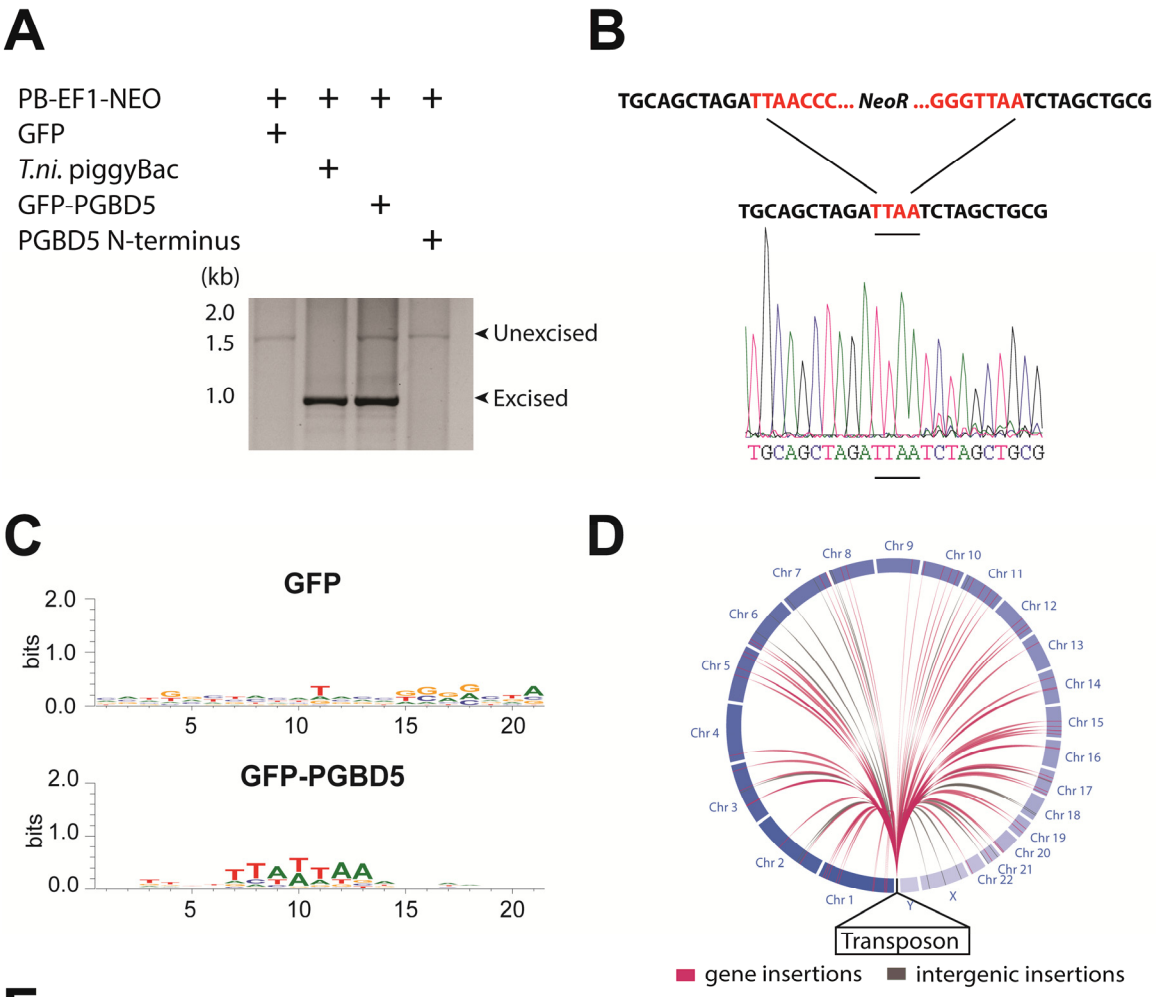

Figure 2. PGBD5 induces DNA transposition in human cells. (A) Representative agarose electrophoresis analysis of PCR-amplified PB-EF1-NEO transposon reporter plasmid from transposase-expressing cells, demonstrating efficient excision of the ITR-containing transposon by PGBD5, but not GFP or PGBD5 N-terminus mutant lacking the transposase domain. *T.ni* piggyBac serves as positive control. (B) Representative Sanger sequencing fluorogram of the excised transposon, demonstrating precise excision of the ITR and associated duplicated TTAA sequence, marked in red. (C) Analysis of the transposon integration sequences, demonstrating TTAA preferences in integrations in cells expressing GFP-PGBD5, but not GFP control. X-axis denotes nucleotide sequence logo position, and y-axis denotes information content in bits. (D) Circos plot of the genomic locations PGBD5-mobilized transposons plotted as a function of chromosome number and transposition into genes (red) and intergenic regions (gray). (E) Alignment of representative DNA sequences of identified genomic integration sites, demonstrating integrations of transposons (green) into human genome (blue) with TTAA insertion sites and genomic coordinates, as marked.

	Intact Transposon		Mutant Transposon	
	TTAA ITR	Non-ITR	TTAA ITR	Non-ITR
Transposase				
GFP-PGBD5	82% (65) †	18% (14)	11% (4) ‡	89% (33)
GFP Control	17% (2)	83% (10)	40% (27)	60% (40)

Table 1. Analysis of transposon integration sequences in human genomes induced by PGBD5. Cells expressing GFP-PGBD5 and intact transposons exhibit significantly higher frequency of genomic integration as compared to either GFP control, or GFP-PGBD5 with mutant transposons, with 87% (69 out of 79) of sequences demonstrating DNA transposition of ITR transposons into TTAA sites († $p = 1.8 \times 10^{-5}$). Mutation of the transposon ITR significantly reduces ITR-mediated integration, with only 11% (4 out of 37) of sequences (‡ $p = 0.0016$). Numbers in parentheses denote absolute numbers of identified insertion sites.

Figure 3

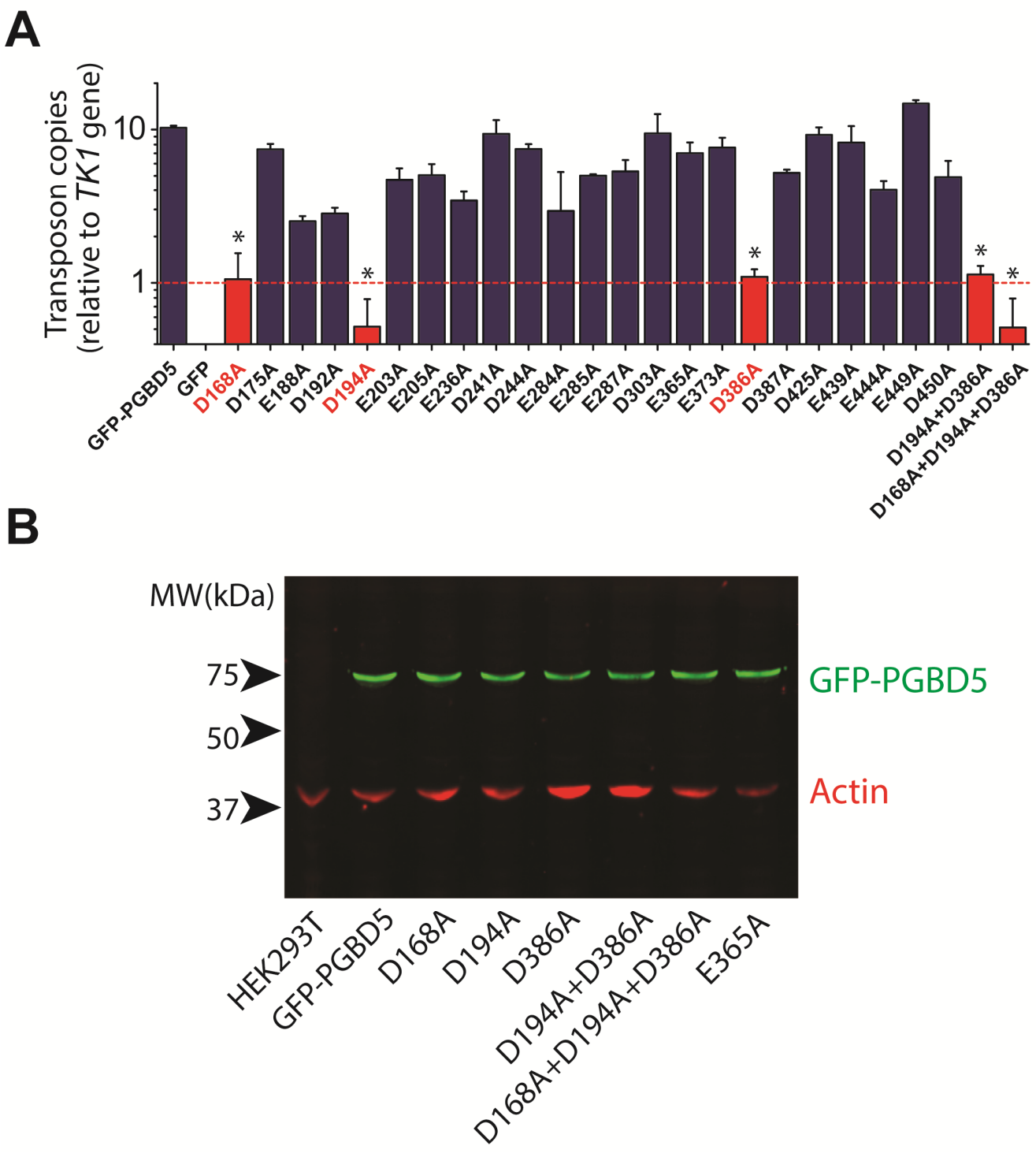

Figure 3. Structure-function analysis of PGBD5-induced DNA transposition using alanine scanning mutagenesis. **(A)** Quantitative PCR analysis of genomic integration activity of alanine point mutants of GFP-PGBD5, as compared to wild-type and GFP control-expressing cells. D168A, D194A, and D386A mutants (red) exhibit significant reduction in apparent activity ($p = 0.00011$, $p = 0.000021$, $p = 0.000013$ vs. GFP-PGBD5, respectively). Dotted line marks threshold at which less than 1 transposon copy was detected per haploid human genome. Error bars represent standard errors of 3 biological replicates. **(B)** Western immunoblot showing equal expression of GFP-PGBD5 mutants, as compared to wild-type GFP-PGBD5 (green). β-actin (red) serves as loading control.

SI Figures

Henssen et al. Genomic DNA transposition induced by human PGBD5

```
Uribo2          ------------MAKRFYSAEEAAAHCMASSSEEFSGSDSEYVPPASESDSSTEESWCSS 48
PiggyBat        --------------MSQHSDYSDDEFCADKLSNYSCDSDLENASTSDE-DSSDDEVMVRP 45
PiggyBac        ------------MGSSLDDEHILSALLQSDDELVGEDSDSEISDHVSEDDVQSDTEEAFI 48
PiggyMac        MFYNDEEEEDWGDKNEKDAEDEFQDLNPSEHKKRIPNNRPKLQPKPQKKKKEVQECEFAI 60
PGBD5           -----------------MPDYRTRALFVQSTLGDSAGPELQLLSIVPGRDLQPSDSFTGP 43
                                        .      .   :        . . .

Uribo2          STVSALEEPMEVDEDVDDLEDQEAGDRADAA---------------------------- 79
PiggyBat        RTLRRRRISSSSSDSESDIE-------------------------------------- 65
PiggyBac        DEVHEVQPTSSGSEILDEQNVIEQPGSSLASN-----------------------RILT 84
PiggyMac        DDPIKRRQLANYQHVPVRLEAGEKMNINVQYDDGQVRKFGQQKAFEEKLIPISKSGTFIY 120
PGBD5           TRKMPPSAS------------------------------------------------- 52

Uribo2          -------AGGEPAWGPPCNFPP----EIPPFT---------------------------- 100
PiggyBat        --------GGREEWS-HVDNPP----VLEDFL---------------------------- 84
PiggyBac        LPQRTIRGKNKHCWSTSKSTRRSRVSALNIVR---------------------------- 116
PiggyMac        EASTGKLERKNPGRPREDTRSLFDDPKLQKMRNNYVPKLLQRITTNNTKEVDEEEGVQDP 180
PGBD5           ------------------------------------------------------------

Uribo2          -------------------------TVPG-VKVDTS-NFEPINFFQLFMTEAILQDMVLY 133
PiggyBat        -------------------------GHQG-LNTDAV-INNIEDAVKLFIGDDFFEFLVEE 117
PiggyBac        -------------------------SQRGPTRMCRN-IYDPLLCFKLFFTDEIISEIVKW 150
PiggyMac        GVSWKPVKSVHEVPESFYMRQVFDKSASGPRSIDKSKIKSEYDAFRLFFDNDIYNTIIKH 240
PGBD5           ---------------------------------------AVDFFQLFVPDNVLKNMVVQ 72
                                                        .:**. : . . ::

Uribo2          TNVYAEQYLTQNP-----------LPRYARAHAWHPTDIAEMKRFVGLTLAMGLIKANSL 182
PiggyBat        SNRYYNQ--NRNN-----------FKLSKKSLKWKDITPQEMKKFLGLIVLMGQVRKDRR 164
PiggyBac        TNAEISLK---------------RRESMTGATFRDTNEDEIYAFFGILVMT-AVRKDNH 193
PiggyMac        TRERYQQKVEEQIYSYIHGMVHMGIRAKKPTLMQWEFTEYELEAYFAVQIFFGIVRLSNQ 300
PGBD5           TNMYAKKF---------------QERFGSDGAWVEVTLTEMKAFLGYMISTSISHCESV 116
                :.   .                                 *:  :..  :     : .

Uribo2          ESYWDTTTVL--------------------------SIPVFSATMSRNRYQLLLRFLHF 215
PiggyBat        DDYWTTEPWT--------------------------ETPYFGKTMTRDRFRQIWKAWHF 197
PiggyBac        MSTDDLFDRS--------------------------LSMVYVSVMSRDRFDFLIRCLRM 226
PiggyMac        RDYWKSSARQKPIKKAETGRRKLRELAQEKMDRYAHWVTQRMSSIVSYEKFKTIRNCLNI 360
PGBD5           LSIWSGGFYS--------------------------NRSLALVMSQARFEKILKYFHV 148
                 .                                          ::  ::  : .  ..

Uribo2          NNNATAVPPDQPGHDRLHKLRPLIDSLSERFAAVYTPCQNICIDESLLL----FKGRLQF 271
PiggyBat        NNNADIVNES----DRLCKVRPVLDYFVPKFINIYKPHQQLSLDEGIVP----WRGRLFF 249
PiggyBac        DD--KSIRPTLRENDVFTPVRKIWDLFIHQCIQNYTPGAHLTIDEQLLG----FRGRCPF 280
PiggyMac        SG---AEALKLKGRDPIWKIRDFLNQMNMRFAKYYYPGEFITIDEGMIP----FAGKVQF 413
PGBD5           VAFRSSQTTHG-----LYKVQPFLDSLQNSFDSAFRPSQTQVLHEPLIDEDPVFIATCTE 203
                               :  :: . : :       : *     :.* ::     : .

Uribo2          RQYIPSKRARYGIKFYKLCESSSGYTSYFLIYEGKDSKLDPPGCPPDLTV------SGKI 325
PiggyBat        RVYNAGKIVKYGILVRLLCESDTGYICNMEIYCGEGKRLLET------------------ 291
PiggyBac        RMYIPNKPSKYGIKILMMCDSGTKYMINGMPYLGRGTQTNGVPLG-----------EYY 328
PiggyMac        KVYNPDKPTKWGIKEYLLCDASNTYTFQLRLYHGQTMWNNDFKQTMFVNEEDTQHRTMEL 473
PGBD5           RELRKRKKRKFSLWVRQCSSTGFIIQIYVHLKEGGGPDGLDALKNKPQLHS-------MV 256
                :     *  ::.:     ..:.           *

Uribo2          VWELISPLLGQGFHLYVDNFYSSIPLFTALYCL--DTPACGTINRNRKGLPRALLDKK-- 381
PiggyBat        IQTVVSPYTDSWYHIYMDNYYNSVANCEALMKN--KFRICGTIRKNR-GIPKDFQTIS-- 346
PiggyBac        VKELSKPVHGSCRNITCDNWFTSIPLAKNLLQEPYKLTIVGTVRSNKREIPEVLKNSR-- 386
PiggyMac        VLQMCKDYEHKAHKVVMDNYYSSWMLFRELRNR--GIGAVGTIRHNRTGLTKKDLTSKHF 531
PGBD5           ARSLCRNAAGKNYIIFTGPSITSLTLFEEFEKQ--GIYCCGLLRARKSDCTGLPLSMLTN 314
                   :      .   :  .   .*      :          * :. .:   .

Uribo2          ----LNRGETYALRKNELLAIKFFD---KKNVFMLTSIHDESVIREQRVGRPPKN----- 429
PiggyBat        ----LKKGETKFIRKNDILLQVWQS---KKPVYLISSIHSAEMEESQNIDRTSKKKIV-- 397
PiggyBac        ----SRPVGTSMFCFDGPLTLVSYK---PKPAKMVYLLS--SCDEDASINESTGK----- 432
```

```
PiggyMac        QQIYNQYHYAYYLNQSNELMLMYFQGTSEKEIALISNFLDNSLNEQHMWDISKQHYYVPH 591
PGBD5           PATPPARGQYQIKMKGNMSLICWYN----KGHFRFLTNAYSPVQQGVIIKRKSGEIP--- 367
                              .        .    *    .         .         .

Uribo2          --KPLCSKEYSKYMGGVDRTDQLQHYYNATRKTRAWYKKVGIYLIQMALRNSYIVYKAAV 487
PiggyBat        --KPNALIDYNKHMKGVDRADQYLSYYSILRRTVKWTKRLAMYMINCALFNSYAVYKSVR 455
PiggyBac        ---PQMVMYYNQTKGGVDTLDQMCSVMTCSRKTNRWPMALLYGMINIACINSFIIYSHNV 489
PiggyMac        LKAPYMMYVYNKYKGGVDRRNSYVVKYRSRFPAKKWWQSVFERLFETAILNAYLIFRSYN 651
PGBD5           --CPLAVEAFAAHLSYICRYDDKYSKYFISHKPNKTWQQVFWFAISIAINNAYILYKMSD 425
                   *     :      :   :.          .      :    :. *  *:: ::

Uribo2          PGPK------------------------------------------------------- 491
PiggyBat        QRKMG------------------------------------------------------ 460
PiggyBac        SSKG------------------------------------------------------- 493
PiggyMac        PESSYRNKGQMRDFRINLMYQFAERYKSYEHQQEENGKNRFSYFAKIQPHTFIEGEEIVK 711
PGBD5           AYHVKR----------------------------------------------------- 431


Uribo2          ----------------------------------------------------LSYYK 496
PiggyBat        ----------------------------------------FKMFLKQTAIHWLTDDIP 478
PiggyBac        ------------------------------------------------------EKVQ 497
PiggyMac        CSECGNETKVFCQECTILKAEVVGLCHEKDTIKCQRFHEFMDFELDKNKEVIDKRKGKDP 771
PGBD5           ----------------------------------------------------YSRAQ 436


Uribo2          YQLQILPALLFGGVEEQTVPEMPPSDNVARL---IGKHFIDTLPPTPG---KQRPQKGCK 550
PiggyBat        EDMDIVPDLQPVPSTSGMRAKPPTSDPPCRLSMDMRKHTLQAIVGSGK---KKNILRRCR 535
PiggyBac        SRKKFMRNLYMSLTSSFMRKRLEAPTLKRYLRDNISNILPNEVPGTSDDSTEEPVMKKRT 557
PiggyMac        YKPNFLEKLNQRTNAKGNQSSQKKESPLVNLLNKINDQIKQEARVKQEVKREDNTNKQTT 831
PGBD5           FGERLVRELLGLEDASPTH----------------------------------------- 455
                    ::  *      .

Uribo2          VCRKR-----GIRRDTRYYCPKCPRNPGLCFKPCFEIYHTQLHY---------------- 589
PiggyBat        VCSVH-----KLRSETRYMCKFCN--IPLHKGACFEKYHTLKNY---------------- 572
PiggyBac        YCTYCP---SKIRRKANASCKKCK--KVICREHNIDMCQSCF------------------ 594
PiggyMac        YIIPEVKSEDESSTDSDIYIQRTTNQRLIEIHQKIEQMSMCSEEFLNGSVANSQENFAER 891
PGBD5           ---------------------------------------------------------


Uribo2          ---------------------------------------------------------
PiggyBat        ---------------------------------------------------------
PiggyBac        ---------------------------------------------------------
PiggyMac        DENDQFNDNNGNDDNQFQFPQQRAQQIDDDDEQRNSKNEEQQKDFIKKMMEFADDEGSEN 951
PGBD5           ---------------------------------------------------------


Uribo2          ---------------------------------------------------------
PiggyBat        ---------------------------------------------------------
PiggyBac        ---------------------------------------------------------
PiggyMac        EEIQYPEDEADHFYQQLLQQEEQAIKYQQKKQLQQQLEEESERSNISHKSKKQQKLEQKF 1011
PGBD5           ---------------------------------------------------------


Uribo2          -------------------------------------------------------
PiggyBat        -------------------------------------------------------
PiggyBac        -------------------------------------------------------
PiggyMac        IETSMRGIKQSQIQSNSEIGQDLQKIISASQDLNQISKQITESKGDNQNSQSDQ 1065
PGBD5           -------------------------------------------------------
```

Fig. S1. PGBD5 is a distinct transposase in the human genome with no apparent similarity in DDD motif with other piggyBac proteins. Clustal W2 alignment of Piggybac like transposases, Uribo 2, PiggyBat, Piggybac, and PGBD5. Canonical conserved piggyBac catalytic residues are highlighted in red.

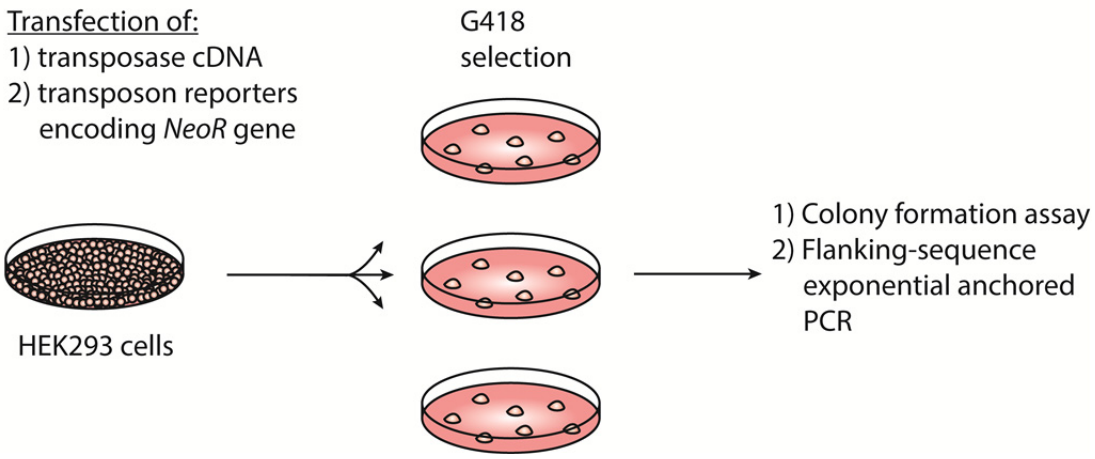

Fig. S2. Assay for genomic integration of transposon reporters. Schematic showing the procedure to assay for genomic integration of transposon reporters using G418 selection to clone genomically-integrated neomycin resistant cells.

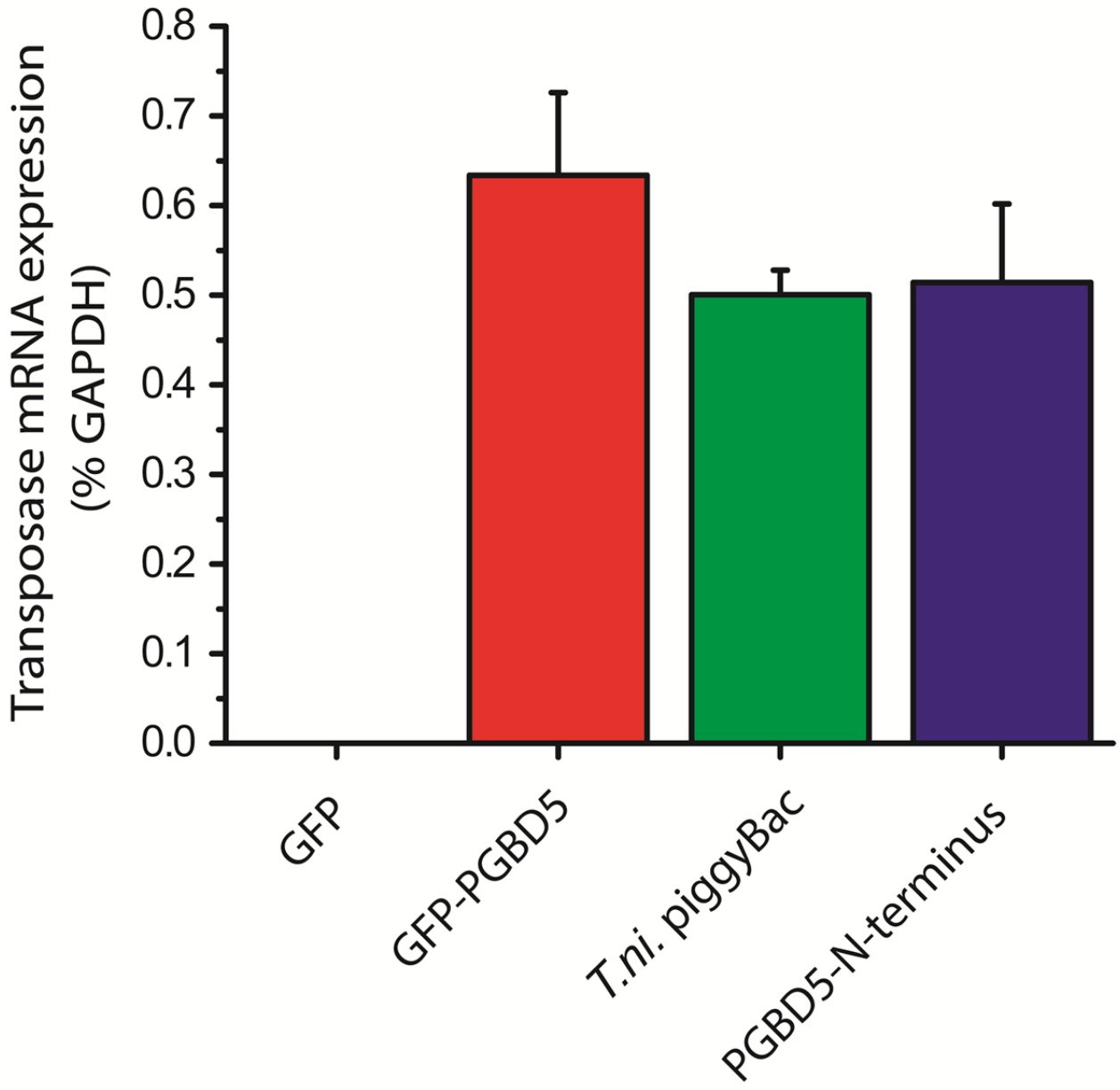

Fig. S3. GFP-PGBD5, PGBD5 N-terminus and T.ni. piggyBac are equally expressed upon transfection in HEK293 cells. Quantitative RT-PCR specific to GFP-PGBD5, PGBD5 N-terminus and T.ni. piggyBac show equal mRNA expression of all 3 transposases (PGBD5 N-terminus $p = 0.17$, T.ni. piggyBac $p = 0.092$).

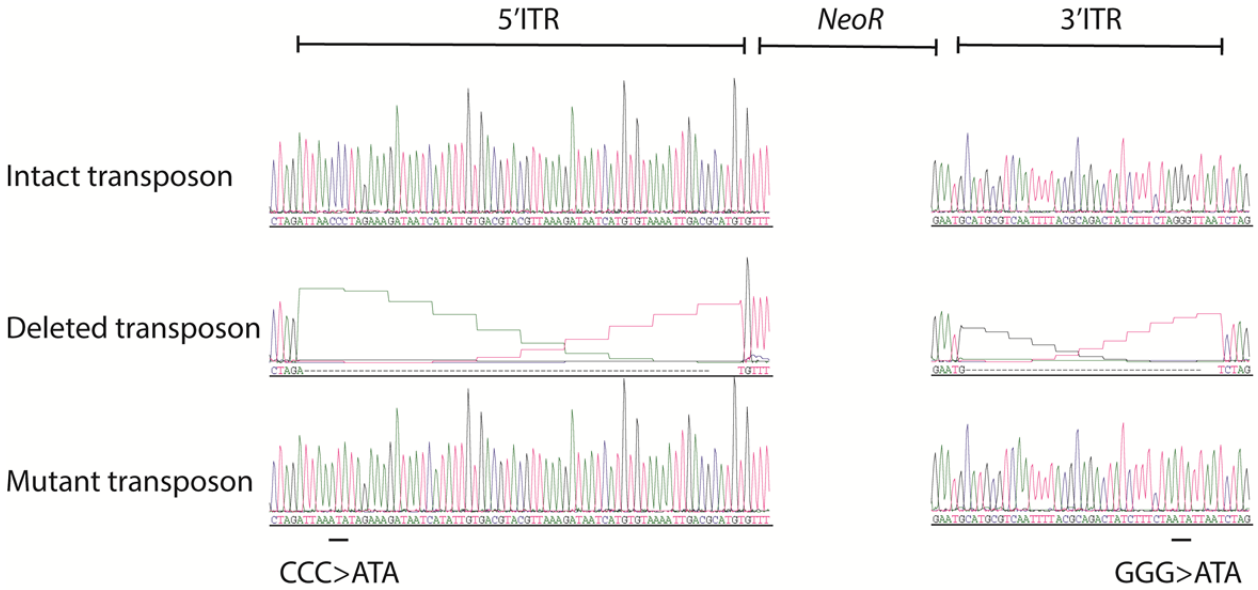

Fig. S4. Sanger sequencing traces of the inverted terminal repeats of the synthetic transposon reporter plasmids. Top to bottom: Intact transposon, deleted transposon and mutant transposon. Black line indicates location of GGG to ATA mutations in mutant transposon ITR.

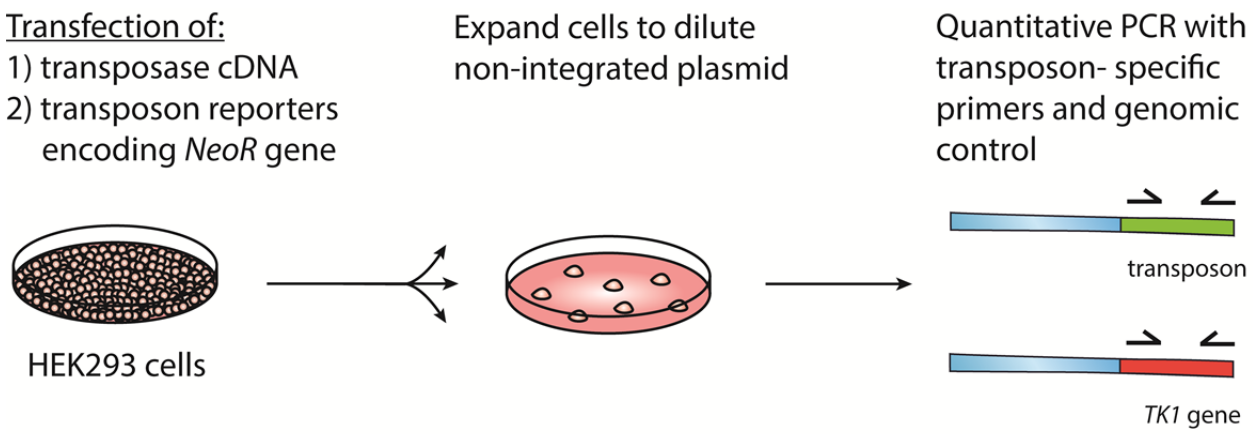

Fig. S5. Quantitative assay of genomic integration of transposon reporters. Cells were transfected as described in Materials and Methods. Next, cells were expanded and genomic DNA was isolated. Quantitative real-time PCR was performed with primers specific to the transposon sequence as well as to the *TK1* reference gene.

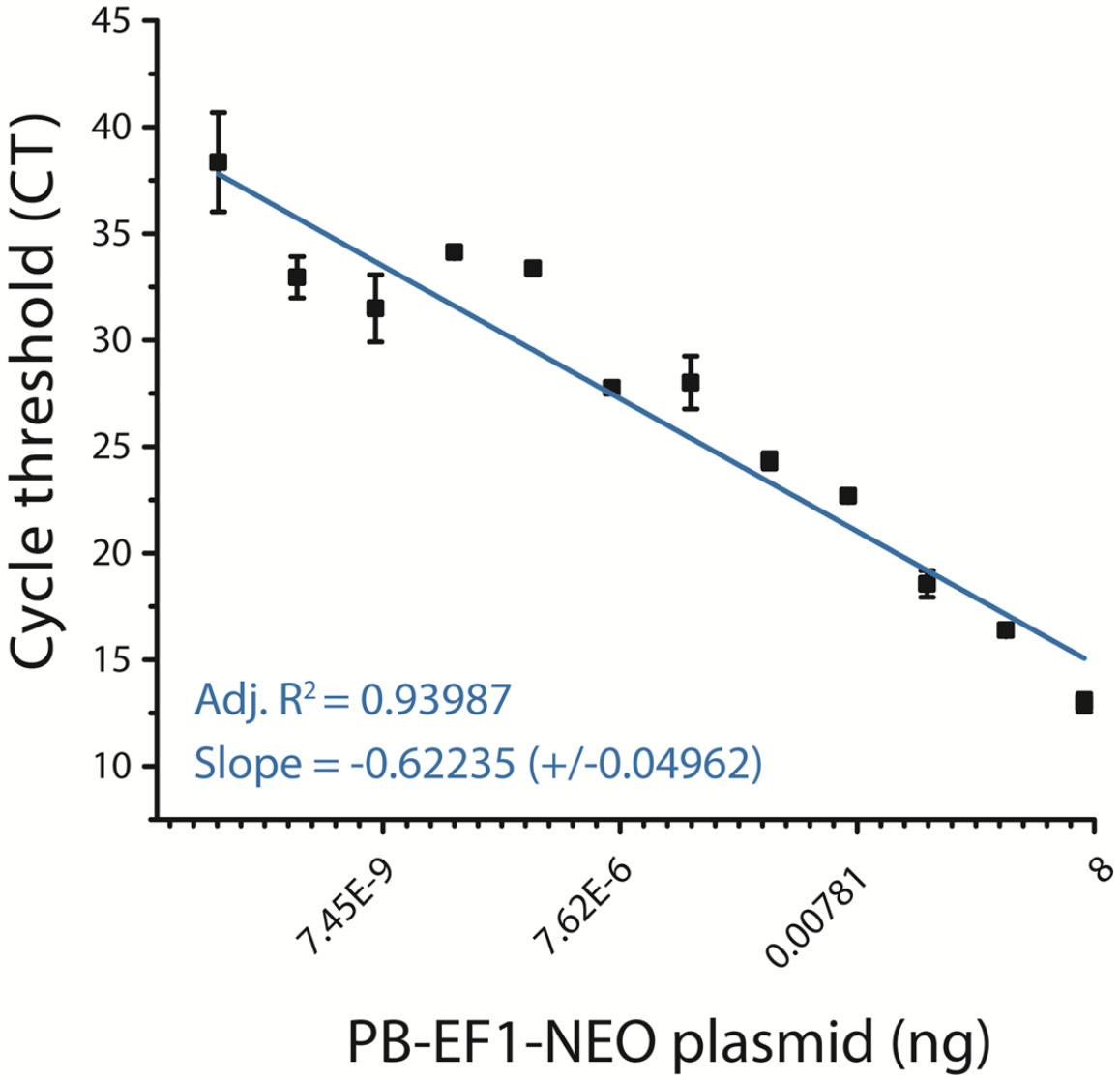

Fig. S6. Quantitative genomic PCR standard curve for transposon specific primers.

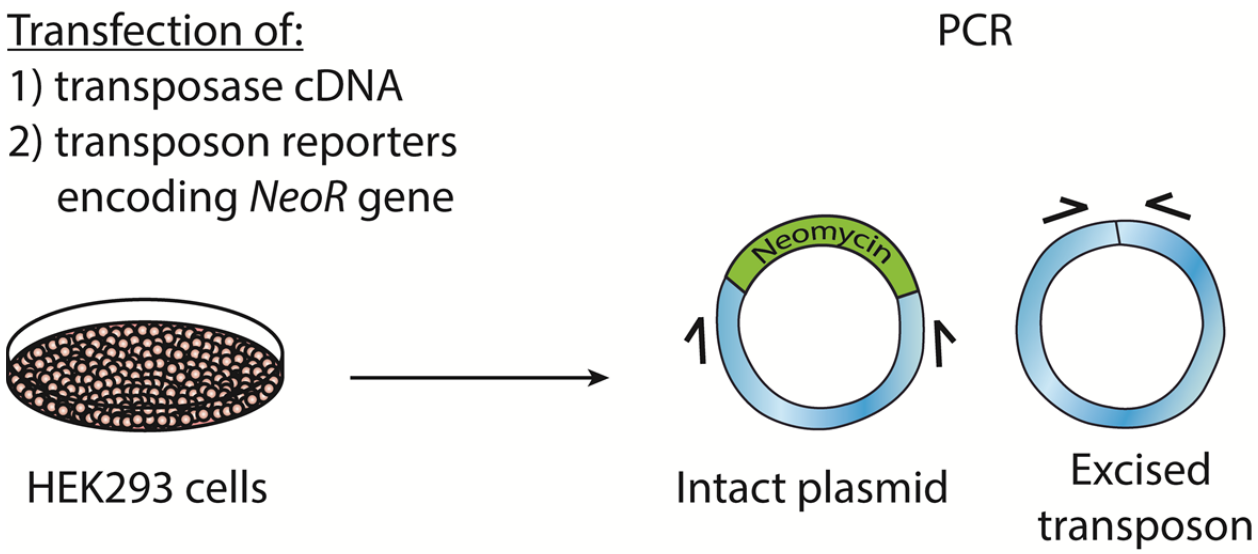

Fig. S7. Schematic of transposon excision assay. Cells were transfected as described in Materials and Methods. Next DNA was isolated. PCR was performed with primers specific to sequences flanking the transposon.

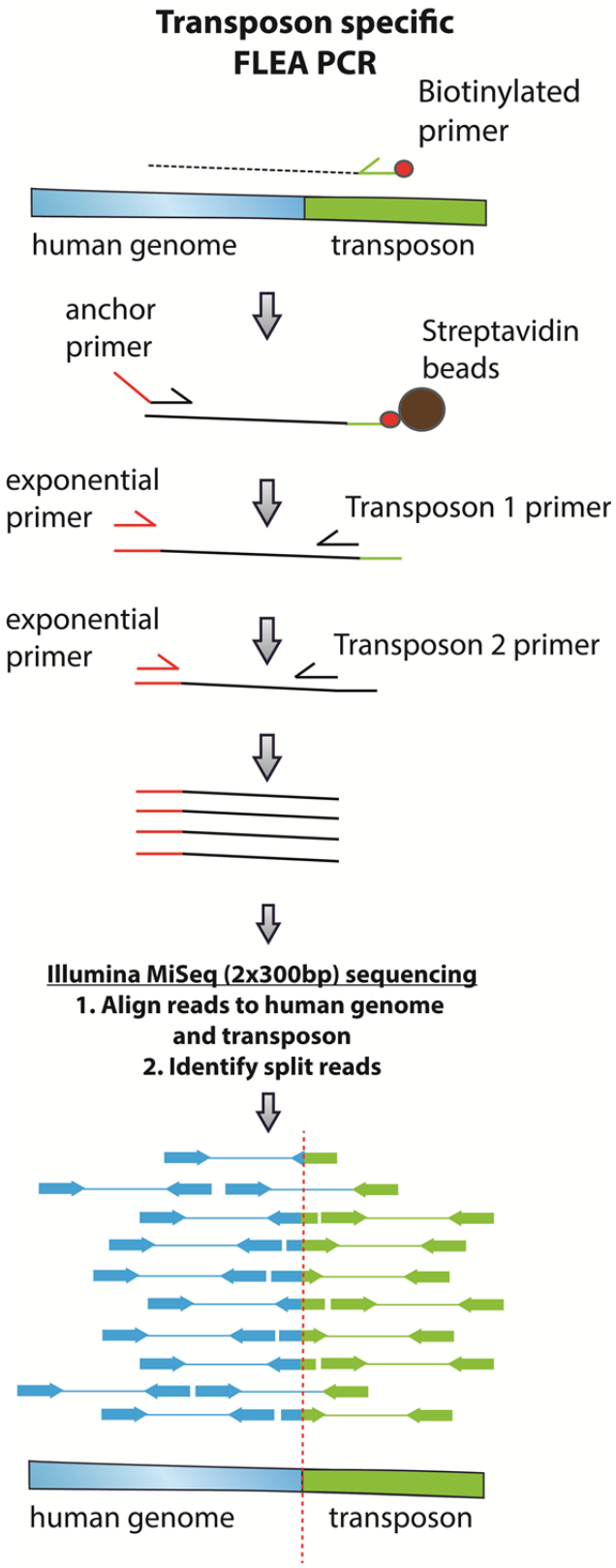

Fig. S8. Schematic of transposon specific flanking-sequence exponential anchored–polymerase chain reaction amplification (FLEA-PCR) and massively parallel single molecule sequencing assay for mapping and sequencing transposon insertions.

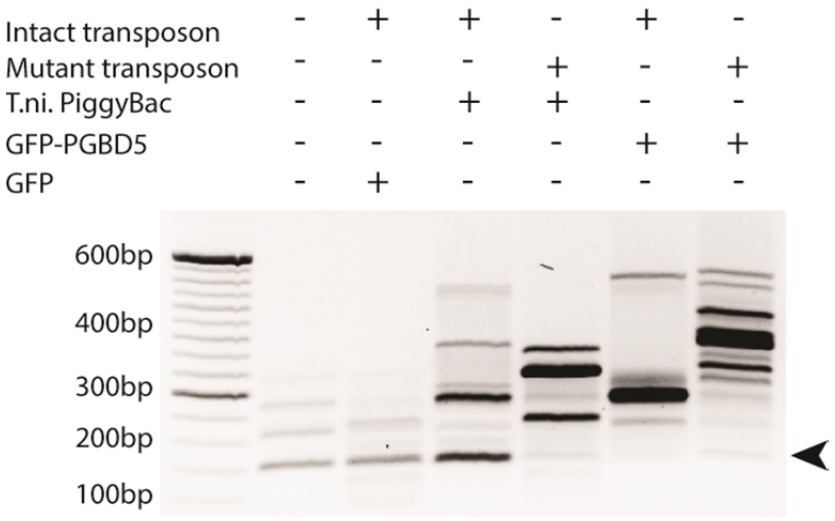

Fig. S9. Representative agarose gel image of amplicons from flanking-sequence exponential anchored–polymerase chain reaction amplification (FLEA-PCR). Arrow indicated expected size of degenerate anchor primer amplicons.

Fig. S10. Sanger sequencing trances of pRecLV103-GFP-PGBD5 D>A and E>A mutants.

Fig. S10. Sanger sequencing trances of pRecLV103-GFP-PGBD5 D>A and E>A mutants.

Fig. S10. Sanger sequencing trances of pRecLV103-GFP-PGBD5 D>A and E>A mutants.

Table S1: Sequences of PCR primers

Name	Description	Sequence (5'-3')
PGBD5_qpcr_F	qPCR of PGBD5, forward primer	GCTTATTCTTCAGCGCATCC
PGBD5_qpcr_R	qPCR of PGBD5, reverse primer	CAGCCTCTGGGTCAGACAAT
NTerm_qPCR_F	qPCR of PGBD5 N-terminus, forward primer	AGAACATGGTGGTGCAGACA
NTerm_qPCR_R	qPCR of PGBD5 N-terminus, reverse primer	GGAGATCATGTAGCCCAGGA
TniPB_qPCR_F	qPCR of T.ni. piggyBac, forward primer	TGAGCATGGTGTACGTGTCC
TniPB_qPCR_R	qPCR of T.ni. piggyBac, reverse primer	CAGGAACATCACCTGCGACA
PB_ExPCR_F	Excision PCR assay, forward primer	GGGTTCCGCGCACATTTC
PB_ExPCR_R	Excision PCR assay, reverse primer	CAGTCATCCTCGGCAAACTCTTT
PB_qPCR_F	qPCR of transposon reporter, forward primer	GATGTCGTGTACTGGCTCCG
PB_qPCR_R	qPCR of transposon reporter	CGCGTGAAGGAGAGATGCGAG
TK1_qPCR_F	qPCR of TK1, forward primer	ATGCTGATGTCTGGGTAGGGTG
TK1_qPCR_R	qPCR of TK1, reverse primer	TGAGTCAGGAGCCAGCGTATG
Bio-linear	FLEA-PCR	[BioTEG]CATTTTGACTCACGCGGTCGT
Anchor	FLEA-PCR	GTGGCACGGACTGATCNNNNN(N-Q)
Exponential	FLEA-PCR	GTGGCACGGACTGCA
Transposon1	FLEA-PCR	ATTGACAAGCACGCCTCACG
Transposon2	FLEA-PCR	ATGCACAGCGACGGATTCG

Table S2: List of supplemental data files

File name	Description
List_insertion_sites.xlsx	FLEA-PCR transposon insertions genomic locations, and integration classes
PRJNA291089	FLEA-PCR Illumina sequencing (fastq and bam files, uploaded to SRA)