Genome-wide Profiling Reveals Remarkable Parallels Between Insertion Site Selection Properties of the MLV Retrovirus and the piggyBac Transposon in Primary Human CD4⁺ T Cells

Andreas Gogol-Döring¹,², Ismahen Ammar³, Saumyashree Gupta⁴, Mario Bunse³, Csaba Miskey⁵, Wei Chen¹, Wolfgang Uckert³, Thomas F Schulz⁴, Zsuzsanna Izsvák³ and Zoltán Ivics⁵

¹Institute of Computer Science, Martin Luther University Halle-Wittenberg, Wittenberg, Germany; ²German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Leipzig, Germany; ³Max Delbrück Center for Molecular Medicine, Berlin, Germany; ⁴Institute of Virology, Hannover Medical School, Hannover, Germany; ⁵Division of Medical Biotechnology, Paul Ehrlich Institute, Langen, Germany

The inherent risks associated with vector insertion in gene therapy need to be carefully assessed. We analyzed the genome-wide distributions of Sleeping Beauty (SB) and piggyBac (PB) transposon insertions as well as MLV retrovirus and HIV lentivirus insertions in human CD4⁺ T cells with respect to a panel of 40 chromatin states. The distribution of SB transposon insertions displayed the least deviation from random, while the PB transposon and the MLV retrovirus showed unexpected parallels across all chromatin states. Both MLV and PB insertions are enriched at transcriptional start sites (TSSs) and co-localize with BRD4-associated sites. We demonstrate physical interaction between the PB transposase and bromodomain and extraterminal domain proteins (including BRD4), suggesting convergent evolution of a tethering mechanism that directs integrating genetic elements into TSSs. We detect unequal biases across the four systems with respect to targeting genes whose deregulation has been previously linked to serious adverse events in gene therapy clinical trials. The SB transposon has the highest theoretical chance of targeting a safe harbor locus in the human genome. The data underscore the significance of vector choice to reduce the mutagenic load on cells in clinical applications.

Received 28 July 2015; accepted 6 January 2016; advance online publication 9 February 2016. doi:10.1038/mt.2016.11

INTRODUCTION

The ability to efficiently deliver foreign genes into cells offers opportunities to use gene therapy to correct genetic diseases and to augment cellular processes to achieve a therapeutic effect (reviewed in refs. 1,2). Hematopoietic stem cell (HSC)-based gene therapy has clearly provided therapeutic benefit in primary immunodeficiencies (including SCID-X1, ADA-SCID), thalassemia, and leukodystrophies.³–⁶ However, uncontrolled integration of contemporary retroviral gene therapy vectors may result in insertional mutagenesis by activating oncogenes,⁷,⁸ as observed in clinical trials for SCID-X1⁹,¹⁰, X-CGD,¹¹ and WAS.¹² In contrast to HSC-based gene therapy, leukemia was never observed in preclinical animal models or clinical trials involving gene transfer into peripheral blood-derived T lymphocytes.¹³,¹⁴ Thus, mature T cells seem to be less susceptible to transformation by genotoxic events than are HSCs, and retroviral gene therapy in T cells therefore has not been thought to involve a major risk of insertional mutagenesis and development of cancer. Importantly, however, recent studies indicate that some HIV integrations into genes associated with cancer or cell cycle regulation may confer a survival advantage of HIV-infected cells and thus a clonal imbalance of HIV integrations in AIDS patients.¹⁷,¹⁸

The risk of insertional oncogenesis in gene therapy is inherently linked to a fundamental step of the life cycle of mobile genetic elements (retroviruses and transposons): genomic insertion. Vector architecture, the enhancer/promoter elements used to drive transgene transcription, copy numbers, the underlying disease, and insertion site selection properties of the vectors can strongly influence the actual risk of insertional oncogenesis. There is a wide spectrum of specificity in target site selection by mobile genetic elements. For example, retroviral/lentiviral integration displays little specificity on the primary DNA sequence level but biased patterns of distribution on the genome level, which is likely due to interaction of the viral components with certain host proteins or recognition of different chromatin states of the chromosomes during integration.¹⁹ For example, the bias of HIV toward integration into active cellular transcription units²⁰ was proposed to be due to tethering interactions with cellular proteins rather than to chromatin accessibility. In particular, the cellular lens epithelium-derived growth factor (LEDGF)/p75 was shown to influence HIV target site selection.²¹

Similar studies showed that MLV has a strong preference for integrating into regions surrounding transcriptional start sites (TSSs).²² However, a recently generated, high-resolution insertion site map based on >3 million unique integration events in two ENCODE-characterized human cell lines revealed that a subset of strong enhancers and active promoters characterized by high enrichment of multiple marks of active chromatin (including H3K4me1, H3K4me2, H3K4me3, H3K27ac, and H3K9ac
are preferentially targeted, and thus, these regions are better predictors of MLV integration than TSSs.\textsuperscript{23} Finally, it was recently reported that the cellular bromodomain and extraterminal (BET) domain proteins (BRD2, BRD3, and BRD4) physically interact with the MLV IN\textsuperscript{24–26}. The N-terminal bromodomains of BET proteins bind to acetylated H3 and H4 tails,\textsuperscript{27} which are associated with TSSs. Thus, MLV integration site distribution parallels the chromatin-binding profile of BET proteins. Furthermore,
Comparative Genome-wide Profiling of Transposon and Retrovirus Integration Sites in Human T Cells

Enrichments of integration sites versus control sites of MLV compared to PB at 40 states, 70 datasets, and 16 genomic features, as described in *f*i.e.*, Fisher's exact test not significant.*

**Sleeping Beauty** (SB) is the most thoroughly studied vertebrate transposon to date, and it has shown highly efficient transposition in different somatic tissues of a wide range of vertebrate species including humans (reviewed in refs. 29,30). SB has been shown to provide long-term transgene expression in preclinical animal models (see refs. 31–33 for recent reviews) and is currently under clinical evaluation as an integrating, nonviral vector system for gene therapy.44,45 The SB transposon preferentially inserts into TA dinucleotides and shows additional target site preferences based on physical properties of the DNA.46,47 On the genomic scale, SB transposons exhibit a close-to-random integration profile with a slight bias toward integration into genes and their upstream regulatory sequences in cultured mammalian cell lines;48–49 this tendency, however, is not as pronounced as seen for viral vectors. The piggyBac (PB) element, a DNA transposon isolated from the cabbage looper moth, has shown transpositional activity in mouse and human cells and thus also has a potential as a vector in gene therapy.44 PB preferentially integrates into TTAA sequences44 with a significant bias toward transcriptionally active regions including genes, TSSs, and DNaseI hypersensitive sites in mammalian cells43,44,46–48 and in *Drosophila.*49 Target site selection properties of the SB and PB transposons together with the Mouse Mammary Tumor Virus (MMTV) have been comparatively analyzed in great detail in mouse embryonic stem (ES) cells;50 however, a similar analysis of these transposons in therapeutically relevant human cell types and against retroviral vector systems that are currently used in several gene therapy clinical trials (Journal of Gene Medicine Clinical Trial Database, 2015) has been lacking.

Although characterization of the target site selection properties of different vector systems still falls short of predicting the actual risk of insertional oncogenesis in a clinical trial, it is highly useful for ranking the different vector types and designs according to their genotoxic potential.2 Thus, we have undertaken a comparative study addressing target site selection properties of the SB and PB transposons as well as the MLV and HIV viral systems in primary human CD4+ T cells. We have chosen this cell type due to the availability of the rich genome-wide mapping data for chromatin marks as well as other genomic features and because currently running phase 1 gene therapy clinical trials with SB use this cell type as target.51 We find that, in contrast to PB, MLV and HIV that all show biased insertion patterns into expressed genes, the SB transposon displays a close-to-random insertion profile, thereby supporting relative safety of SB in human applications. The PB transposon shows an intriguing, MLV-like profile with pronounced preference for integrating into the 3′-transcriptional regulatory regions of genes, and we show that this is largely shaped by physical interaction between the transposase and BET proteins. We describe an additional tethering mechanism that involves chromatin-associated transposase molecules in SB transposition. In sum, our findings have important implications for the safety of these integration systems for genome engineering, including human gene therapy.

**RESULTS**

**Genomic states define chromosomal regions preferred for integration**

In order to generate datasets representing *de novo* transposon integration sites, primary human CD4+ T cells were electroporated with pairs of transposase and transposon plasmids of the SB and the PB systems (Supplementary Figure S1a). Linear amplification-mediated PCR was used for the recovery of genomic transposon integrations, and the PCR libraries were sequenced using the Illumina/Solexa HiSeq Platform (Supplementary Figure S1b). The retrieved transposon integration sites (Supplementary Figure S1c) for the SB (8,290 sites) and PB (8,954 sites) systems were used together with datasets generated in CD4+ human T cells with the MLV retrovirus (66,764 sites)52 and with the HIV lentivirus (7,765 sites)52 in a comparative manner for all downstream analyses (schematic maps of all four vectors are shown in Supplementary Figure S1d). Consensus sequences at transposon integration sites revealed that the highly preferred TA target site dinucleotides for SB and the TTAA tetranucleotide motif for PB are embedded in AT-rich DNA, as noted previously44,53 (Supplementary Figure S2a). In contrast, the viral systems display very weak preferences at the primary DNA sequence level for integration (Supplementary Figure S2a,b).

Mapping of SB and PB insertions on the human chromosome set revealed an overall wide distribution across the human genome (Supplementary Figure S3). In the context of describing insertional preferences on a genome-wide scale, it has been recently shown that chromatin “states” defined by co-occurrence of multiple chromatin marks are far better predictors for integration than any of those marks separately.51 Keeping this in mind, we used an unsupervised machine-learning approach (ChromHMM, which applies a Hidden Markov Model)54 in order to define 40 different genomic states, each characterized

**Figure 1** Analysis of chromatin states. The ChromHMM software (which uses a Hidden Markov Model (HMM)) was applied to derive 40 different genomic states from 70 empirical datasets including histone marks, DNA methylation, chromatin accessibility, or the binding sites of numerous transcription factors, polymerases, histone acetylases, and deacetylases. (a) The output probabilities of the trained HMM. Black: the state and the dataset are disjoint. (b) Enrichments of each state for a set of genomic regions, computed using the OverlapEnrichment function of ChromHMM. (c) Genome coverage of states. More frequent states correspond to darker colors. (d–f) Enrichment of integration sites versus control sites at the different (d) states, (e) dataset regions, or (f) genomic features. Blue: depletion of integration sites at the given regions; red: enrichment of integration sites. If the depletion/enrichment of integration sites at a given region is not significant (i.e., Fisher’s exact test $P$ value $>0.05$), then the field is marked gray. (g) Comparison of PB and MLV. The scatter plots show the fold enrichments of integration sites versus control sites of MLV compared to PB at 40 states, 70 datasets, and 16 genomic features, as described in a–f. Gray dots correspond to region sets in which PB or MLV show no significant enrichment/depletion of integrations.
by a coherent pattern of presence or absence of genomic characteristics specified in 70 different datasets (Figure 1a). Although it is not obvious how to assign a clear definition to each of these automatically generated states, one can get hints about possible functions of the different states from Figure 1b, which shows the level of enrichment or depletion of states within certain genomic features. For example, state 1 represents the genomic (intergenic) background without a marked association with any specific measured characteristics. States 2–7 correspond to heterochromatic regions, where state 6 specifically covers methylated DNA. States 11–13 are weak enhancer regions, mainly inside gene bodies, while states 27, 33, and 34 are strong enhancers located in intergenic regions. States 34 and 37–40 are promoter regions, whereas state 39 is associated with a strong histone acetylation signal. State 34 marks active promoters, whereas promoters in state 37 feature a less accessible chromatin and tend to be associated with low or unexpressed genes. States 16, 17, 31, and 32 correspond to transcriptional ends of active genes. The tool then assigned to each genomic region one of these states. The states cover between 25% (state 1) and 0.01% (state 33) of the genome (Figure 1c). We then determined the fold enrichments of integration sites versus control sites at the different states (Figure 1d), dataset regions (Figure 1e), and genomic features (Figure 1f). Unexpectedly, the enrichments for PB and MLV are very tightly correlated in all three cases (coefficients of determination $R^2$ between 0.90 and 0.95; Figure 1g). In contrast, SB and PB display much weaker correlation ($R^2$ between 0.18 and 0.43, data not shown).

![Figure 2 Analysis of integration frequencies relative to annotated genes.](image)

(a) Distribution of distance between integration sites outside of genes and the nearest gene. The figures were created using the R density function with Gaussian kernel and bandwidth 0.2 on the log10 distances. The dashed lines show the median distances for SB (48.1 kb), PB (24.3 kb), HIV (13.2 kb), MLV (12.4 kb), and the random controls (85.6 kb). (b) Fold enrichments (i.e., the proportion of experimentally determined integration sites falling into a given set of regions divided by the proportion of control sites falling into the same set of regions) of integrations sites within RefSeq genes and of integration sites found upstream (−50 kb to −1 kb) or downstream (1 kb to 50 kb) of RefSeq genes. The horizontal line at 1 represents the value expected for random sites. With exception of the integration frequency of SB and HIV into 1 and 2 kb around RefSeq genes, the comparison between the experimentally determined integration sites and the random integration sites achieved in all panels $P \leq 0.001$ (Fisher’s exact test). (c) The fold enrichments of SB, PB, HIV, and MLV insertion sites are plotted over the gene body from the 5′ to the 3′ end. Each gene was split into 40 equal bits, and fold enrichment of integration was calculated for each bit. The figure shows fold enrichments of insertion per bit averaged over all RefSeq genes. The dashed line marks fold enrichment of 1. (d) Insertions around transcriptional start sites (TSSs). The figure shows fold enrichment of insertions around RefSeq TSSs. The numbers of sites at each distance was averaged over all RefSeq genes and smoothed by computing the mean in a 500-bp rolling window using the R zoo package. The fold changes were computed after smoothing. The transcription starts at position 0 and proceeds to the right, i.e., positive $x$ values represent the transcribed region, while negative $x$ values represent regions upstream of the TSSs.
Distribution of vector integration sites with respect to genes and TSSs

We next determined the frequencies of integration into genomic features including genes, exons, introns, TSSs (TSS ± 1 kb), CpG islands and DNaseI hypersensitive sites relative to computer-generated random data sets matched for each of the four vector systems (Figure 2 and Supplementary Figure S4). Distribution of distance between integration sites outside of genes and the nearest gene revealed that intergenic insertions tend to be closer to genes than the control datasets (Figure 2a). For example, the median distance for HIV integration sites is around 13 kb, whereas for the control dataset, it is 85 kb. Out of the four systems, the SB transposon insertions are the farthest away from genes (median distance 48 kb), suggesting a very low preference for integrating near gene sequences. For PB and MLV, Figure 2a shows a bimodal shape reflecting the difference between integration sites close to TSS (left peak) and sites more distal to genes (right peak). Insertions by all four systems were enriched in genes (Figure 2b) (both in exons and introns; Supplementary Figure S4). The preference of HIV to target genes is well described.20 Remarkably, the PB transposon displayed an MLV retrovirus-like insertion profile with respect to preferentially targeting the upstream regions of genes, whereas the SB transposon displayed the weakest preference toward genes (Figure 2b). In the case of MLV, we found a significant orientation bias (6% more MLV integrations into genes in sense than in antisense direction). Plotting insertion sites over the gene bodies revealed high enrichment of both MLV and PB insertions at the 5′-ends of genes, whereas HIV insertions tend to be enriched in the gene bodies and relatively disfavored at the 5′- and 3′-ends of genes (Figure 2c). Finally, both PB and MLV are enriched (with no significant difference between the frequencies of insertions with sense and antisense orientation) and HIV depleted around TSSs (Figure 2d and Supplementary Figure S4).

Interestingly, the MLV insertions concentrate at two peaks up- and downstream of the actual TSSs leaving a small drop of insertions at the TSSs, whereas PB insertions map directly at TSSs (Figure 2d). The bimodal MLV integration pattern can be explained by (i) a tendency of retroviruses to direct integration into outward-facing major grooves on nucleosome-wrapped DNA51,55 and (ii) TSSs of expressed genes are nucleosome depleted.56 In contrast, high-density integration profiling of the Hermes transposon in Saccharomyces cerevisiae and in Schizosaccharomyces pombe revealed a strong association of Hermes integration sites with nucleosome-free chromatin.57,58 Target site choice by the SB and PB transposons with respect to nucleosomal versus nucleosome-free DNA has not been investigated.

We determined genomic positions of nucleosomes from micrococcal nuclease sequencing (MNase-Seq) data19 using the TemplateFilter program99 with template “Templates_7.1.txt” and default parameters. We assumed that the 40-bp regions directly neighboring each nucleosome belong to linkers. We then counted all insertion sites and control sites falling into the nucleosomes or linker regions. All sites outside of these regions were excluded from the subsequent analysis, since the chromatin states of these regions were unknown. In order to avoid wrong assignments, we only counted sites which were more than 5 kb away from the nucleosome borders. The figure shows the fractions of sites inside of the nucleosomes; the rest of the sites fall into linker regions. SB and PB prefer to insert into linker regions (P value < 10−14 for SB and P value <10−5 for PB, Fisher’s exact test), while MLV prefers nucleosomes (P value <0.01, Fisher’s exact test).

Figure 3 Insertions into nucleosomes versus linkers. We determined genomic positions of nucleosomes from MNase-Seq data19 using the TemplateFilter program99 with template “Templates_7.1.txt” and default parameters. We assumed that the 40-bp regions directly neighboring each nucleosome belong to linkers. We then counted all insertion sites and control sites falling into the nucleosome or linker regions. All sites outside of these regions were excluded from the subsequent analysis, since the chromatin states of these regions were unknown. In order to avoid wrong assignments, we only counted sites which were more than 5 kb away from the nucleosome borders. The figure shows the fractions of sites inside of the nucleosomes; the rest of the sites fall into linker regions. SB and PB prefer to insert into linker regions (P value < 10−14 for SB and P value <10−5 for PB, Fisher’s exact test), while MLV prefers nucleosomes (P value <0.01, Fisher’s exact test).

Vector integration positively correlates with gene expression levels and gene densities

We next investigated a potential association between the expression status of genes and the relative frequencies of integrations. Transcriptional regulatory elements often coincide with CpG islands.61 Indeed, the patterns of integrations at TSSs are mirrored for CpG islands in that HIV is depleted, whereas PB and MLV are enriched at CpG islands (Supplementary Figure S5). SB showed no enrichment at CpG islands. We also investigated if the actual methylation status of CpG sites influences integration. Interestingly, even though both MLV and PB insertions are enriched in CpG islands, both elements avoid methylated and prefer nonmethylated CpG sites (Supplementary Figure S5). Because actively transcribed genes tend to be associated with nonmethylated CpG islands, this finding suggests that both MLV and PB target transcriptionally active regions, including promoters. Indeed, both MLV and PB integration sites were also found to highly correlate to DNaseI hypersensitive sites (Supplementary Figure S4).

We next systematically analyzed frequencies of insertions and the transcriptional status of genes and their promoters. In general, a positive correlation between gene expression and integration frequencies can be observed; i.e., integrations tend to be depleted in nonexpressed genes but enriched in expressed...
Figure 4 Correlation between gene expression and integration into genes and transcriptional start sites. (a) Fold enrichment (i.e., the proportion of experimentally determined integration sites falling into a given set of regions divided by the proportion of control sites falling into the same set of regions) of integrations into not expressed (absent = A; 5,603 genes), marginal (M; 5,253 genes), and clearly expressed (present = P; 6,947 genes) genes. The gene expression levels were determined by using Affymetrix Human Genome U133 Plus 2.0 Arrays. The classification into A, M, and P genes was done by Schones et al using the Affymetrix GCOS software package. (b) Fold enrichment of integrations for genes of different expression level. We split the set of 12,200 marginal or clearly expressed genes into 10 groups (each of 1,220 genes) according to their expression level from the lowest (left) to the highest (right) genes. The enrichment is higher for higher expression levels. (c) Fold enrichment (i.e., the proportion of experimentally determined integration sites from −1 kb to +1 kb around the TSSs falling into a given set of regions divided by the proportion of control sites falling into the same set of regions) of integrations into genes categorized as above. (d) Fold enrichment of integrations for genes of different expression levels. HIV was omitted from this plot since the total number of HIV insertions close to TSSs was too low. (e) Chromosomal densities of insertions. The relative insertion densities, i.e., the fraction of insertions falling in a certain chromosome divided by the fraction of possible insertions falling in this chromosome. In the case of SB, all occurrences of TA dinucleotides are possible insertion sites. In the case of PB, all occurrences of TTAA are possible insertion sites. In the case of HIV and MLV, they could integrate at any genomic position. The relative insertion density compares the density of insertion to the density of all possible insertion sites. For completely random insertions, the relative insertion density would be 1; values above 1 mean an enrichment of insertions, values below 1 mean a depletion of insertions. (f) Gene densities, i.e., number of RefSeq genes/isoforms per chromosome divided by the chromosomal length (in million bp).
genes (Figure 4a). Moreover, the enrichment is higher in highly expressed than in marginally expressed genes. In other words, the stronger a gene is expressed, the more likely it would be targeted by any of the four integration systems analyzed here, by HIV displaying the most significant correlation (Figure 4a and Supplementary Figure S6). Breaking down gene expression levels into a larger number of categories indicates that the enrichment in insertion frequencies gradually grows with increasing expression level (Figure 4b). This is interesting for MLV and PB, since they are enriched at TSSs and not in the gene bodies. Thus, we also looked at association between the expression status of genes and frequencies of integrations into TSSs. Figure 4c shows that HIV is always depleted and PB and MLV are always enriched around TSSs, but in all three cases, the number of insertions increases with the expression levels. However, the level of enrichment for PB and MLV reaches a plateau at a certain level of gene expression (Figure 4d). This is a hint that for PB and MLV the gene expression level itself is less important than the open chromatin state at promoter regions. Interestingly, frequencies of SB insertions into TSSs do not seem to correlate with the expression levels of the targeted genes (Figure 4c,d).

Having seen an overall correlation of integration frequencies and genes, we analyzed gene densities across chromosomes and found that insertion densities correlate very well with the gene densities (Pearson correlation coefficients are between 0.87 and 0.94; Figure 4e). That is, gene-rich chromosomes including chr17, chr19, and chr22 are more frequently hit by insertions, with HIV displaying the most significant correlation (0.94). The bioinformatic analysis revealed an almost random integration profile for SB with only a slight bias for euchromatin marks and no bias with respect to heterochromatin marks (Figure 5a). In contrast, the PB transposon as well as the MLV and HIV viruses avoid heterochromatin marks and are enriched in chromosomal regions characterized by open chromatin (Figure 5a). Both PB and MLV are highly enriched in chromatin marks associated with TSSs (H3K4me3 and enhancer regions (H3K4me4)) and enhancer regions (H3K4me1). Differences between PB and MLV become visible when inspecting the density of histone marks as a function of the distance to insertion sites (Figure 5b). In case of marks prominent at TSSs, i.e., H3K4me3, Pol II, and CTCF, MLV integration sites have the characteristic “shifted” peak at a distance of about 500–1,000 bp (Figure 5b). A similar distribution of MLV insertions was seen with respect to DNaseI cleavage sites, consistent with DNaseI preferentially cleaving nucleosome-free DNA, while MLV preferentially inserts into nucleosome-bound DNA. Finally, consistent with preferential integration into gene bodies, HIV integration sites are highly associated with open chromatin characterized by H3K36me3.

**Tethering mechanisms affecting SB and PB integrations**

Chromosomal tethering by interaction of the integration machinery of transposable elements and viruses with host cell-encoded, DNA- or chromatin-binding factors may introduce an insertional bias in target site selection. Such targeting mechanism, based on interactions between LEDGF/p75 and HIV IN, and between BET domain proteins and MLV IN is known to play a role in target site distribution of HIV and MLV, respectively. Although SB is a fish transposon, and the human genome does not contain SB-like transposons, DNA sequences resembling the transposase-binding sites at the ends of the transposon may occur in human DNA by chance, and such sites might be bound by the SB transposase. Can
it then be that, in an analogous fashion, the transpositional complex is tethered to chromosomal regions bound by excess transposase molecules during transposition?

We addressed this possibility by mapping our insertion datasets with respect to the 5′ GTTTACATACAC 3′ sequence motif representing the SB transposase core binding site, allowing one mismatch in the motif. In total, we found 20,188 occurrences of this motif in the human genome. We detected a highly significant enrichment of SB insertions within 100 bp of the motifs (Figure 6), consistent with a tethering mechanism, in which the transpositional complex is anchored to certain chromosomal sites bound by excess transposase molecules (Supplementary Figure S7). Neither PB nor the MLV and HIV insertions displayed enrichment close to SB transposase-binding sites (Figure 6). Highly significant enrichments of SB insertions could also be found for different window sizes and for alternative definitions of SB transposase-binding sites (e.g., all occurrences of 5′ GTTTACATACAC 3′ with up to

![Figure 6 Enrichment of integrations near chromosomal SB transposase-binding sites in the human genome.](image6)

(a) Fold enrichments of insertion sites compared to random controls in up to 100-bp distance of SB transposase-binding sites, i.e., genomic occurrences of 5′ GTTTACATACAC 3′ with up to one mismatch. SB integrations are significantly enriched (P value < 1 × 10^15, Fisher’s exact test), while the integration frequencies of PB, HIV, and MLV compared to their respective backgrounds are not significantly changed. (b) Fold enrichments of insertions sites compared to random controls around SB transposase-binding sites, i.e., genomic occurrences of 5′ GTTTACATACAC 3′ with up to three mismatches. The numbers of sites at each distance was averaged over all RefSeq genes and smoothed by computing the mean in a 100-bp rolling window using the R zoo package. The displayed fold changes were computed after smoothing. There are 3.2 million genomic occurrences of 5′ GTTTACATACAC 3′ with up to three mismatches. Four hundred and eighteen out of 8,290 (5%) SB insertions fell directly inside of one of these SB transposase-binding sites, which is a highly significant enrichment (P value < 1 × 10^15 Fisher’s exact test, fold change 2.02) compared to random controls. The human reference genome (hg18) was screened for SB transposase-binding sites using Bowtie.65

![Figure 7 Interaction of the transposase with BET proteins targets piggyBac insertions to BRD4-binding sites.](image7)

(a) Enrichment of PB and MLV insertions near BRD4-associated sites. The fold enrichment of insertions around 33,544 BRD4-binding sites detected by Zhang et al.65 The figure was created with the same method as in Figure 2d for TSSs. The numbers of sites at each distance were averaged over all RefSeq genes and smoothed by computing the mean in a 500-bp rolling window using the R zoo package. The fold changes were computed after smoothing. The profiles for PB and MLV resemble the pattern observed for TSSs. Fifty-three percent of the BRD4-binding sites are inside of RefSeq genes, this explains the relative high number of HIV integrations nearby. (b) The PB transposase interacts with the BET proteins BRD2 and BRD4. Extracts of HEK293T cells expressing PB-HA along with GFP-tagged BRD2 or BRD4 were immunoprecipitated (IP) with a GFP affinity matrix, and immunoprecipitates were analyzed by Western blotting (WB) using rat anti-HA or mouse anti-GFP antibody as indicated.
2 or 3 mismatches, or all occurrences of 6-mers or 8-mers of 5’ GTTTACATACAC 3’; data not shown).

Given the highly similar patterns of PB and MLV insertions in our datasets with respect to proximity to TSSs (Figure 2d), we wondered if PB insertions are also enriched at sites associated with BET proteins. Thus, we next evaluated whether there is a correlation between PB integration sites and the chromatin-binding sites of BET proteins mapped using ChIP-Seq data obtained in CD4+ T cells.46 Figure 7a shows a significant co-localization of BRD4-binding sites, MLV integration sites, and PB integration sites. The data also revealed a very strong enrichment and positioning of PB insertion sites directly at the BRD4-binding sites, whereas MLV insertions are positioned at the flanks of the BRD4-binding sites. Thus, the correlation between PB and MLV insertions with respect to BRD4-binding sites is very similar to the distribution of insertions with respect to TSSs as shown in Figure 2d. These results suggest that PB, in a fashion analogous to MLV, is possibly targeted to TSSs through a tethering mechanism dictated by chromatin-bound BRD4.

To investigate a possible physical interaction between the PB transposase and BET proteins, a co-immunoprecipitation experiment was performed (Figure 7b). Cell extracts were prepared from HEK293T cells expressing HA-tagged PB transposase and GFP-tagged BET proteins, BRD2 and BRD4. An antibody against GFP was used for immunoprecipitation. Precipitated proteins were subsequently detected with an antibody against the HA tag of the PB transposase. PB transposase was co-precipitated with BRD2 and BRD4, but not with control (GFP) (Figure 7b, lanes 1, 2, and 3, respectively). Thus, we conclude that PB transposase interacts with BET proteins. We further found that the C-termini of BET proteins spanning their highly conserved ET domains that are required for interaction with the MLV IN (aa640-801 for BRD2, aa6539-726 for BRD3, and aa607-722 for BRD4)24–26 was sufficient to interact with the PB transposase (Supplementary Figure S8a). Finally, we tested if the residues in the BRD2 ET domain that contribute to binding to MLV IN and FeLV IN26 are also involved in binding to the PB transposase. We found that none of the residues involved in binding to INs are important for interaction with the PB transposase (Supplementary Figure S8b). Thus, although PB transposase interacts with the ET domain, it does so in a manner different from MLV IN and FeLV IN.

**Potential deregulation of gene expression upon integration and genomic safe harbors**

Integration of therapeutic gene constructs into safe sites in the human genome would prevent insertional mutagenesis and associated risks of oncogenesis in gene therapy. Genomic “safe harbors” (GSHs) are regions of the human genome that are able to accommodate the predictable expression of newly integrated DNA without adverse effects on the host cell or organism. It was previously proposed that GSHs should meet the following five criteria: (i) distance of at least 50 kb from the 5’-end of any gene, (ii) distance of at least 300 kb from any cancer-related gene, (iii) distance of at least 300 kb from any microRNA (miRNA), (iv) location outside a transcription unit, and (v) location outside ultraconserved regions of the human genome.66-67 We compiled our datasets to investigate the relative frequencies of integration into a GSH by any of the four integration systems. Figure 8a shows that the viral systems have very reduced chance (as low as ~3% for HIV) of integrating into a GSH. The PB transposon has a ~12% chance of integrating into a GSH, thereby it is expected to be safer than the two viruses. However, based on these criteria, the SB transposon is predicted to be the safest in a therapeutic context, with an overall chance of ~20% of integration into GSHs.

Some of the adverse events observed in HSC-based clinical trials revealed a clonal imbalance in reconstituted hematopoiesis in patients associated with gammaretroviral insertions into the LMO2 (refs. 9–12,14), MN1 (14), CCND2 (9), BMI1 (ref. 9), MECOM/MDS1/EVI1 (refs. 13,14), PRDM16 (68), and SETBP1 (68) genes, and some of these insertions have been shown to be causally linked to oncogenesis. Because our insertional datasets allow us to assess relative enrichment of integrations in genes in the absence of biological selection, we wondered if any of the genes recovered in the clinical trials are actually favored targets by the four vector systems. Figure 8b reveals that MLV indeed favors integration into some of these genes with SETBP1 favored more than twofold over random chance to be hit. Thus, vector choice greatly influences the relative chance of insertional oncogenesis in gene therapy clinical trials.

Although lentiviral vectors have long been considered to be safe for genetic engineering in differentiated T cells, HIV integrations have been recently associated with clonal cell expansion in AIDS patients.17,18 Some of these HIV integrations occurred in genes playing roles in cell growth, development, and cancer, suggesting that proviral integrations into some of these genes can drive biological selection on the level of cell survival and selective proliferation. We selected a total of 29 genes from these two studies (these genes were recovered in two out of three patients in the Wagner et al. study or in two out of five patients in the Maldarelli et al. study) and analyzed if any of them is a favored target by the four vector systems. Figure 8b reveals that 55% (16/29) of these genes are favored by MLV, 41% (12/29) by HIV, and 24% (7/29) by PB, with significant overlap between these gene lists. For example, five genes (CYTH1, IKZF3, NFATC3, RPTOR, and TNRC6B) are mutually favored by MLV, HIV, and PB. Not a single gene from this list appears to be preferentially targeted by SB. Thus, the data reveal that vector choice can greatly contribute to a reduced likelihood of insertional mutagenesis of genes implicated in driving clonal dominance in T cells.

**DISCUSSION**

DNA-based, cut-and-paste transposons display a wide spectrum of selectivity with respect to chromosomal integration. In this work, we mapped ~9,000 de novo SB and PB insertions in primary human CD4+ T cells and compared their insertion profiles with those of the MLV retrovirus and the HIV lentivirus. Our bioinformatic analyses included mapping against the T cell genome with respect to proximity to genes, TSSs, CpG islands, DNaseI hypersensitive sites, chromatin marks, and transcriptional status of genes (Supplementary Table S1). The SB transposon displayed the least deviation from random with respect to genome-wide distribution: no apparent bias was seen for either heterochromatin marks or euchromatin marks, and only a weak correlation with transcriptional status of targeted genes was detected (Figures 2,
Figure 8 Insertions into genomic safe harbors and selected genes. (a) Instead of “ultra-conserved” regions, we used a database of regions conserved in primates created by the phyloP program from the PHAST package (phyloP46waysPrimates downloaded from UCSC table browser). A list of 3,165 cancer-related genes was downloaded from http://cbio.mskcc.org/CancerGenes. “Safe harbor” is defined as the intersection of the five conditions: not conserved, >300 kb from cancer genes and miRNA genes, >50 bp from TSS, and outside of genes. The error bar for “control” results from the divergences of the four different control sets. (b) Relative enrichment of Sleeping Beauty and piggyBac transposon, MLV retrovirus, and HIV lentivirus integration sites in selected genes. Number of integration sites into genes and their relative enrichment compared to control sites is shown. The list contains selected genes that were either insertional deregulated by MLV-based vectors in gene therapy clinical trials leading to oncogenesis or that were found to be associated with HIV insertions and a clonal imbalance in the T cell repertoires of AIDS patients. The table contains the number of insertion sites in the given genes and the relative enrichment of integration sites compared to the respective set of random control sites. If no integration site was recovered in a given gene, then this enrichment is set to “0x.” P values were calculated using Fisher’s exact test; multiple test correction was done using the Benjamini–Hochberg method. Significant enrichments (corrected P values ≤0.05) are marked orange.
4, and 5). This is in marked contrast to target site distributions of several other transposons including Tol2 (refs. 39, 41), TcBuster,\(^{43}\) SPIN,\(^{43}\) and PB\(^{40,41,43,44}\) that all show significant difference from random insertion with respect to favored integration into genes and near chromatin marks characteristic of active transcription units (e.g., H3K27 acetylation and H3K4 monomethylation) and disfavored integration near marks characteristic of inactive chromatin (e.g., H3K27 trimethylation). The PB transposon, in particular, has been shown to favor open chromatin, expressed genes, and TSSs (±5kb) associated with DNase1 hypersensitive sites, H3K4me3 marks, and Pol II-bound regions in mouse and human cells.\(^{54,46-48,50,53}\)

We have identified remarkable parallels between integration site distributions of the PB transposon and the MLV retrovirus across 40 different chromatin states defined by combinations of genomic features specified in 70 datasets (Figure 1). Both PB and MLV were highly enriched in chromatin marks associated with TSSs (H3K4me3), in regions characterized by Pol II and CTCF binding, in proximity to expressed genes and in genes with higher expression levels (Figures 2, 4, and 5), suggesting that a major determining factor of insertion site distribution is physical accessibility of chromatin. Our studies, however, also highlight the potential involvement of an active mechanism of shaping the characteristic, MLV-like insertion profile of the PB transposon. Namely, PB insertions co-localize with BRD4-associated sites at TSSs, and the PB transposase interacts with BET proteins, including BRD4 (Figure 7). Recent studies revealed the role of an interaction of the MLV IN with BET domain proteins in tethering the viral pre-integration complex to TSSs,\(^{24-26}\) and our results suggest a similar mechanism influencing PB integration. Although both the PB transposase and the MLV IN are DDE recombinases, these two proteins are only distantly related; therefore, a BRD/BET protein-dependent tethering mechanism in their chromosomal integration process is likely a result of convergent evolution. Indeed, in addition to MLV, some other viruses exploit cellular BET proteins for different aspects of their life cycle (reviewed in ref. 69). One prominent mechanism is anchoring of the episomal genomes of papillomaviruses,\(^{70,71}\) Kaposis sarcoma-associated herpesvirus,\(^{72-75}\) and Epstein–Barr virus\(^{76}\) to either interphase chromatin or mitotic chromosomes by interactions of viral proteins with BET domain proteins. Thus, tethering mechanisms relying on interactions between virally encoded factors and BET proteins appear to have independently arisen during viral evolution.

Despite the similarities, a close inspection of the integration sites of PB and MLV at TSSs revealed a characteristic difference: the MLV insertions map at two peaks just up- and downstream of the TSSs, whereas PB insertions map directly at TSSs (Figure 2d). The bimodal MLV integration pattern is likely the result of a preference of MLV to integrate into nucleosome-wrapped DNA\(^{52,55}\) and that TSSs of expressed genes tend to be nucleosome depleted.\(^{56}\) We have shown that, in contrast to MLV, PB insertions favor nucleosome-free DNA (Figure 3), thereby providing a likely explanation for the characteristic difference between PB and MLV insertion patterns at TSSs.

We provide evidence for enriched insertion of the SB transposon near chromosomal sites that resemble binding sites of the SB transposase (Figure 6). These data are consistent with a tethering mechanism that involves interaction of the transpositional nucleoprotein complex with chromatin-bound excess transposase molecules (Supplementary Figure S7). It has been proposed that SB transposition involves a transposase tetramer associated with the transposon ends.\(^{77}\) Can it be that this tetrameric complex can establish contacts with additional transposase molecules bound elsewhere in the genome? Although the relative contributions of the four transposase monomers to the catalytic steps of transposition have not been elucidated, it is possible that not all monomers are equally engaged in the reaction. Indeed, the bacterial Mu transposase forms a stable tetramer with the Mu DNA ends but only two of the active sites within the tetramer are involved in catalysis.\(^{78-80}\) Although the other two subunits of the tetramer do not supply DDE residues to the active sites, they are likely to play other important roles, including maintaining the structural integrity of the transpososome.\(^{81,82}\) Similarly, the foamy virus retroviral intasome structure revealed a tetramer of IN, but the catalytic DDE residues are contributed by only two IN subunits.\(^{83}\) Thus, other mobile elements (including SB) may also require “surplus” recombinase subunits for temporarily stabilizing pairing of the transposon ends and for the formation of a catalytically primed synaptic complex.

Finally, our data allow us to estimate the relative safety of the four integrating genetic elements in the context of human applications (Figure 8). Our compiled datasets allow us to rank these vector systems with respect to their projected relative “safety” based on the frequencies of integration into GSHs (Figure 8a) as well as into selected genes that were targeted by retroviral insertions in gene therapy clinical trials leading to serious adverse events or by HIV insertions on AIDS patients leading to a clonal imbalance in their T cell repertoire (Figure 8b). Our analyses collectively establish a favorable integration profile of the SB transposon. It has to be noted that those insertions that are not in GSHs are not necessarily equally genotoxic. Indeed, it has been demonstrated that MLV-based gammaretroviral insertions, although they target GSHs >2-fold more frequently than HIV-based lentiviral vectors (Figure 8a), were approximately threefold more likely to trigger transformation of primary HSCs in a cell-based immortalization assay.\(^{84}\) This suggests that an MLV insertion next to a TSSs tends to be more genotoxic than an HIV insertion in a gene body. Furthermore, the mutagenic potential of any integrating gene vector will ultimately be defined not only by its insertional pattern, but also by its cargo (including the transcriptional regulatory elements that drive transgene expression) and by vector copy number per genome. Importantly, vector copy number can be experimentally adjusted by titrating the components of the SB transposon system in the electroporation reactions to yield primarily one or two insertions per cell,\(^{29}\) very much like multiplicity of infection largely determines vector copy number in viral vector transductions. Finally, important steps have been made toward introducing an experimental bias into the natural target site selection properties of integrating gene vector systems. First, for both the SB\(^{42}\) and the PB\(^{45}\) transposon systems, it has been shown that engineered DNA-binding domains can drive at least a fraction of integration events into a chromosomal locus or region defined by sequence-specific DNA–protein interactions, suggesting a possibility to target vector integrations into validated GSHs in the
future. Second, disrupting the interaction of BET proteins with the MLV preintegration complex by targeted mutagenesis of IN has been shown to result in detargeting of TSSs, thereby yielding a potentially safer genome-wide insertion profile. Thus, continuing efforts of vector engineering will likely have a considerable impact on the safety of future vector designs.

MATERIALS AND METHODS

Vector construction. Transposase expression vectors pCaggs-SB100 and pCaggs-pB were kindly provided by Grabundzija. The transposon vectors pUC19SBcaggsGFP and pUC19pBCaggsGFP were generated by replacing the SV40neo cassette in pUC19SBneo and pUC19pBneo by a Caggs promoter-driven GFP expression cassette.

Isolation of human peripheral blood mononuclear cells and enrichment of CD4+ T cells. After obtaining informed consent, venous blood from healthy volunteers was drawn into ethylenediaminetetraacetic acid-containing tubes (S-Monovette, Sarstedt, Nümbrecht, Germany) and diluted 1:2 with buffer (phosphate-buffered saline (PBS), 2% heat-inactivated fetal calf serum). The diluted blood was layered onto Biocoll separating solution (Biochrom AG, Berlin, Germany) at a volume ratio of 2:1, and after centrifugation for 20 minutes at 648 g without brake and low acceleration, the layer of mononuclear cells was aspirated. Peripheral blood mononuclear cells were washed twice with buffer (first cycle, 10 minutes at 300 xg; second cycle, 10 minutes at 200 xg), and the cell density was adjusted to 5 x 10^6/ml.

CD4+ T cells were isolated from peripheral blood mononuclear cells by negative immunomagnetic selection using the EasySep Human CD4+ T Cell Enrichment Kit (Stemcell Technologies, Grenoble, France) according to the manufacturer’s instructions. Hundred microliter antibody cocktail were added to 1 x 10^7 peripheral blood mononuclear cells in 2 ml buffer and incubated for 10 minutes at room temperature. Hundred microliter magnetic particles were added, and after 5 minutes of incubation at room temperature, the cell suspension was adjusted to 2.5 ml by adding buffer. The tube was placed into the EasySep magnet (Stemcell Technologies, Grenoble, France), and after 5 minutes, the unlabeled CD4+ T-cell fraction was poured off into a new tube.

Electroporation, activation, and expansion of T cells. T cells were electroporated using the Nucleofector I device and the Human T cell Nucleofector Kit (Lonza, Cologne, Germany) according to the manufacturer’s instructions. The time between the electroporation of the transposon to transposase ratio of 1:2 in 100 µl DNA-nucleofector solution and electroporation was performed using the program U-14. RPMI 1640 Glutamax (Life Technologies, Darmstadt, Germany) supplemented with 10% heat-inactivated fetal calf serum (Biochrom, Berlin, Germany) and 10 mM HEPES was used as T cell medium (TCM). Immediately after the electroporation, 0.5 ml TCM was added to the cuvette, the cell suspension was transferred into a 24-well tissue culture plate containing 1.5 ml pre-warmed TCM, and cultured overnight in a humidified incubator at 37 °C and 5% CO2.

T cells were activated 24 hours after electroporation by transfer into a new plate precoated with 5 µg/ml anti-CD3 and 1 µg/ml anti-CD28 antibodies (BD Pharmingen, Heidelberg, Germany) and addition of 100 U/ml IL-2 (Proleukin, Novatis, Basel, Switzerland). After 3–4 days, the activated T cells were transferred into a new cell culture flask, and 2–3 ml TCM supplemented with 100 U/ml IL-2 was added daily.

Analysis of T cells by flow cytometry. The transposon vectors were tagged with a GFP expression cassette that allowed an estimation of the respective transfection and transpositional efficiencies at day 1 and day 10 post electroporation. Sustained GFP expression, as judged by fluorescence-activated cell sorting (FACS) analysis, in the presence of the respective transposasses is indicative of stable, transposon-mediated genetic modification of human CD4+ T cells (Supplementary Figure S1a). Both transposon systems were about equally efficient in stable gene transfer and resulted in ~40% GFP+ cells. For FACS analysis, cells were washed and incubated with APC-labeled anti-human CD4 antibody (BD Pharmingen, Heidelberg, Germany) in FACS buffer (PBS, 2% fetal calf serum, 2 mM ethylene-diaminetetraacetic acid, 0.05% NaN3) for 30 minutes at 4 °C and washed twice afterwards. Before measurement, SYTOX blue (Life Technologies, Darmstadt, Germany) was added to stain dead cells. Flow cytometry data was acquired using a FACS canto II (BD Bioscience, Heidelberg, Germany), and data were analyzed with FlowJo software (TreeStar, Ashland, OR).

Recovery of integration sites. Unsorted cell populations at day 10 post-electroporation were harvested for genomic DNA preparation. Two variations of the linear amplification-mediated PCR were performed to amplify the vector–genomic DNA junctions. For both approaches, T cells were harvested 10 days after electroporation, and genomic DNA was extracted using the DNeasy kit (Qiagen, Hilden, Germany). Five microworm of genomic DNA was either pre-digested with Dpn1 and BamHI (SB) or Dpn1 and Kpn1 (PB) or sonicated to small pieces ranging from 100 to 500 bp, with an average size of 250 bp using the Covaris S2 sonication device. Thereafter, the sonicated DNA was ethanol precipitated. The digested DNA was subjected to gel electrophoresis, and the genomic DNA was isolated from 0.7% agarose gel and purified using GenElute Gel extraction kit (Sigma-Aldrich, St Louis, MO). For both, 500 ng DNA was used for linear amplification-mediated PCR. Biotinylated SB- and PB- transposon inverted terminal repeat-specific primers (see Supplementary Materials for primer sequences) were used in 50 rounds of linear amplification to enrich DNA species containing transposon–chromosomal DNA junctions. The single-stranded products were immobilized on streptavidin-coated magnetic beads using the Dynabeads kilobase BINDER kit (Invitrogen, Carlsbad, CA). All subsequent steps were performed on the magnetic bead-bound DNA. Repeated washing steps with water followed each reaction. Second strand synthesis was performed with random hexamer oligos (Roche, Basel, Switzerland) using Klenow (New England Biolabs, Ipswich, MA). The free ends of the double-stranded sonicated DNA were blunt ended and phosphorylated using the End-IT DNA End-Repair kit (Epicentre, Madison, WI). Klenow fragment exo- (New England Biolabs) and dATP were used to add a single “A” nucleotide to the 3’ ends. The double-stranded DNA of the second approach was subjected to restriction digests with MboI, HpaII, or CviQI. The DNA fragments with an “A” overhang were ligated to linkers with a “T” overhang, whereas the digested DNA was ligated with linkers having the equivalent overhang created by the respective restriction enzyme. Next, the bead-bound DNA was subjected to a PCR using primers specific for the inverted terminal repeat sequences and the linkers. During the amplification, we used barcode primers so that we could pool different libraries. Finally, primers corresponding to Illumina adapter sequences were used to yield a directional library, in which sequences complementary to the Illumina genomic DNA sequence primers were located upstream to the transposon inverted terminal repeats. Thus, the resulting libraries could be pooled and sequenced on a single flow cell lane on the Illumina HiSeq platform with single end run settings.

Analysis of sequencing data. We selected all sequencing reads which passed the quality filter of the Illumina real-time analysis program and which started with the barcode (exact sequence) and the transposon inverted terminal repeat-specific primer sequences (with up to one mismatches). The rest of the reads (27 bp for SB starting with TA dinucleotide, and 30 bp for PB starting with TTAA) were mapped to the human genome (hg18, downloaded from genome.ucsc.edu) using the following procedure to avoid spurious insertion sites. First, we determined all sequencing reads mapping exactly to one or more positions within the reference genome using Bowtie. The resulting reads were mapped against each other and then clustered such that any two reads with up to two mismatches belong to the same cluster. We kept only reads which occurred at least twice and
contributed at least one-fifths of the total number of reads within their cluster. From the resulting reads, we kept only those which mapped exactly to a unique position in the reference genome. Reads mapping to the same TA (in the case of SB) or TTAA (in the case of PB) sequence in the genome were then merged together. Sequence data is accessible in the GEO database at http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?token=wtssfakcnv yfroz&acc=GSE58744.

**Generation of control sites.** We created sets of control sites for SB and PB separately; 12 control sites per insertion site. The four protocol variants, i.e., using sonication or digestion with Mbol, HpaII, or CviQI, require slightly different methods of control site selection. For insertion sites found by sonication, we randomly selected the control sites from all occurrences of TA (in the case of SB) or TTAA (in the case of PB) in the genome. For insertion sites found by enzyme digestion, we selected occurrences of TA/TTAA having the same distance to the closest enzyme restriction site as the original insertion site. Since each insertion site could be found by any of these four protocol variants, we adjusted the number of control sites accordingly: if, for example, one insertion site was found both by sonica-
tion and Mbol, then we created six control sites using the sonication method and six control sites using the Mbol method. We then retrieved the genomic sequences at the control sites, 27 bp in the case of SB and 30 bp in the case of PB, and then processed them in the same way as the sequenc-
ing reads (see above) keeping at the end only the uniquely mapped control sites. For HIV and MLV, we used the random control sites described by Roth et al.21

**ChromHMM analysis.** The following ChIP-Seq data sets were retrieved from various public repositories: BRG1 (ref. 89), CTCF, H2BKT5me1, H3K27me1, H3K27me2, H3K27me3, H3K36me3, H3K4me1, H3K4me3, H3K79me1, H3K79me2, H3K79me3, H3K9me1, H3K9me2, H3K9me3, H3R2me1, H3R2me2, H4K20me1, H4K20me3, H4R3me2 (ref. 62), PolIII.90 HMGN1, Y1 (ref. 91), STAT1, STAT4, STAT5A, STAT5B, 92 gH2AX, H2AX, H2AZ, INO80, SRCAP, 93 H2AK5ac, H2AK9ac, H2BK120ac, H2BK12ac, H2BK20ac, H2BK5ac, H3K14ac, H3K18ac, H3K23ac, H3K27ac, H3K36ac, H3K4ac, H3K9ac, H4K12ac, H4K16ac, H4K5ac, H4K8ac, H4K91ac.44 CBP, MOE, p300, PCAF, Tip60, HDAC1, HDAC2, HDAC3, HDAC6 (ref. 95), BRD4, PolIII, PolIIIS2P, and PolIIIS2P. Moreover, we used DNase-Seq data,94 total RNA-Seq and polyαRNA-Seq data,95 and MRE and MeDIP assays measuring DNA methylation.96

For creating sets of genomic features, we downloaded genome annotation files from the UCSC Genome Bioinformatics Site (http://
genome.ucsc.edu). The splitting of genes into the groups “absent,” “marginal,” and “present” according to their expression levels was done as described in ref. 59. DNase hypersensitive sites were taken from ref. 63. “Absent,” “Marginal,” “Present,” and “Strong” according to their expression levels was done as described in ref. 98. Conserved regions and safe harbors were defined as described in refs. 66,67.

**Materials and Methods**

**ACKNOWLEDGMENTS**

We thank N. Malani and F. Bushman for kindly providing raw inser-
tional datasets for MLV and HIV obtained in T cells. This work was sup-
ported by EU FP7 (PERSIST, grant number 222878) and grants from the Deutscher Forschungsgemeinschaft “Mechanisms of gene vector entry and persistence” (SPP1230, grant number 17/4-2), Collaborative Research Centre “Chronic Infections: Microbial Persistence and its Control” (SFB900 TP C1), and from the Bundesministerium für Bildung und Forschung (NGFNPlus ENGINE, grant number 01GS08199).

The authors declare no conflict of interest.

**REFERENCES**

1. Mavilio, F and Ferrari, G (2008). Genetic modification of somatic stem cells. The progress, problems and prospects of a new therapeutic technology. EMBO Rep 9 (suppl. 1): 564–569.
2. Naldini, L (2011). Ex vivo gene transfer and correction for cell-based therapies. Nat Rev Genet 12: 301–315.
3. Aji, A, Biasco, L, Scaracuzza, S, Ferrua, F, Cicalese, MP, Baricordi, C et al. (2013). Lentiviral hematopoietic stem cell gene therapy in patients with Wiskott-Aldrich syndrome. Science 341: 1233151.
4. Biffi, A, Montroni, E, Loriodi, L, Cesani, M, Fumagalli, F, Plati, E et al. (2013). Lentiviral hematopoietic stem cell gene therapy benefits metachromatoc leukodystrophy. Science 341: 1233158.
5. Cavazzana-Calvo, M, Payen, E, Negre, O, Wang, G, Hehir, K, Fussl, F et al. (2010). Transduction independence and HMG2 activation after gene therapy of human β-thalassaemia. Nature 467: 318–322.
6. Cartier, N, Hacen-Bey-Aaina, B, Bartholomae, CC, Veres, G, Schmidt, M, Kutschera, I et al. (2009). Hematopoietic stem cell gene therapy with a lentiviral vector in x-linked adrenoleukodystrophy. Science 326: 818–823.
7. Baum, C, von Kalle, C, Staal, FJ, Li, Z, Fehse, B, Schmidt, M et al. (2004). Chance or necessity? Insertional mutagenesis in gene therapy and its consequences. Mol Ther 9: S1–S11.
8. Deichmann, A, Hacen-Bey-Aaina, B, Schmidt, M, Garrigue, A, Brugman, MH, Hu, J et al. (2007). Vector integration is nonrandom and clustered and influences the fate of lymphopoiensis in SCID-X1 gene therapy. J Clin Invest 117: 2225–2232.
9. Hacen-Bey-Aaina, B, Garrigue, A, Wang, GP, Soulier, J, Lim, A, Morillon, E et al. (2008). Insertional oncogenesis in 4 patients after retrovirus-mediated gene therapy of SCID-X1. J Clin Invest 118: 3132–3142.
10. Hove, S, Mansour, MR, Schwartzwoelder, K, Bartholomae, C, Hubank, M, Kempski, H et al. (2008). Insertional mutagenesis combined with acquired somatic mutations causes leukemogenesis following gene therapy of SCID-X1 patients. J Clin Invest 118: 3114–3150.
11. Hacen-Bey-Aaina, B, Von Kalle, C, Schmidt, M, McCormack, MP, Wulfraat, N, Lebouche, P et al. (2003). LMO2-associated clonal T cell proliferation in two patients after gene therapy for SCID-X1. Science 302: 415–419.

**SUPPLEMENTARY MATERIAL**

**Figure S1.** Sustained GFP expression in human CD4+ T cells after genetic modification using the Sleeping Beauty and piggyBac transposon systems.

**Figure S2.** Local sequence information content at vector integration sites.

**Figure S3.** Genome-wide mapping of Sleeping Beauty and piggyBac integrations in primary human T cells.

**Figure S4.** Insertions into genomic features.

**Figure S5.** Integration and DNA-methylation.

**Figure S6.** Correlation between gene expression and integration into genes.

**Figure S7.** A model for a tethering mechanism in Sleeping Beauty transposon integration that involves interaction between the trans-
 positional nucleoprotein complex and chromatin-bound excess trans-
 posase molecules.

**Figure S8.** The PB transposase interacts with the BET proteins BRD2 and BRD4 via their C-termini.

**Table S1.** Summary of genomic features around insertion sites.

www.moleculartherapy.org vol. 24 no. 3 mar. 2016
Comparative Genome-wide Profiling of Transposon and Retrovirus Integration Sites in Human T Cells

Molecular Therapy vol. 24 no. 3 mar. 2016

12. Thressor, A., Gaspar, H., Baum, C., Modlich, U., Schambach, A., Candotti, F. et al. (2006). Gene therapy: X-SCID transgene leukemogenesis. Nature 443: 43–46; discussion 46.

13. Stein, S., Ott, M., Schütte-Strasser, S., Jauch, A., Burwinkel, B., Kinner, A. et al. (2010). Genomic instability and myelodysplasia with monosomy 7 consequent to ETV6-NTRK3 mutation after gene therapy for chronic granulomatous disease. Nat Med 16: 198–204.

14. Braun, C., Boztug, K., Parviznina, A., Witzel, M., Schwarzwar, A., Rothe, M. et al. (2014). Gene therapy for Wiskott-Aldrich syndrome long-term efficacy and genotoxicity. Sci Transl Med 6: 227ra33.

15. Recchia, A., Bonini, C., Magnani, Z., Urbinati, F., Sartori, D., Muraro, S. et al. (2006). Retroviral vector integration deregulates gene expression but has no consequence on the biology and function of transplanted T cells. Proc Natl Acad Sci USA 103: 1457–1462.

16. Neuvrelz, S., Corinillos, K., Li, Z., Baum, C., Brugman, M., Hirtmann, M. et al. (2008). Resistance of mature T cells to oncogene transformation. Blood 112: 2278–2286.

17. Maldarelli, F., Wu, X., Xu, L., Simonetti, FR., Shao, W., Hill, S. et al. (2014). HIV latency. Specific HIV integration sites are linked to clonal expansion and persistence of infected cells. Science 345: 179–183.

18. Wagner, TA., McLaughlin, S., Garg, K., Cheung, CY., Larsen, BB., Stychak, S. et al. (2014). HIV latency. Proliferation of cells with HIV integrated into cancer genes contributes to persistent infection. Science 345: 570–573.

19. Mitchell, RS., Beitzel, BF., Schroder, AR., Shin, P., Chen, H., Berry, CC. et al. (2004). Retroviral DNA integration: ASLV, HIV, and MLV show distinct target site preferences. PLoS Biol 2: e234.

20. Schroder, AR., Shin, P., Chen, H., Berry, C., Ecker, JR. and Bushman, F. (2002). HIV-1 integration in the human genome favors active genes and local hotspots. Cell 110: 215–226.

21. Ciuffi, A., Llano, M., Poeschla, E., Hoffmann, C., Leipzig, J., Shinn, P. et al. (2010). Role of LIGase/PC in targeting HIV DNA integration. Nat Med 16: 1287–1289.

22. Wu, X., Li, Y., Cripe, B. and Burgess, SM. (2003). Transcription-dependent initiation of novel target sites for HIV-1 integration. Nature 420: 1749–1751.

23. LaFeve, MC., Varshney, GH., Gildea, DE., Wolfsberg, TG., Baxevanis, AD. and Burgess, SM. (2014). MLV integration site selection is driven by strong enhancers and active promoters. Nature 511: 4257–4269.

24. Gupta, SS., Maetzigt, T., Maertens, GN., Sharif, A., Rothe, M. and Weidner-Glunde, M. et al. (2013). Bromo- and extraterminal domain chromatin regulators serve as cofactors for the leukemia virus integration machinery. Nat Struct Mol Biol 20: 22–28.

25. Sharma, A., Larue, RC., Plumb, MR., Malani, N., Male, F., Sugrue, A. et al. (2013). BET proteins promote efficient murine leukemia virus integration at transcription start sites. Proc Natl Acad Sci USA 110: 12360–12364.

26. Filipakesopoulou, P., Qi, J., Picard, S., Shens, Y., Smith, WB., Fedorov, O. et al. (2010). Selective inhibition of BET bromodomains. Nature 468: 1067–1073.

27. Ayer, S., Swapova, CV., Malani, N., Aramini, JM., Schneider, WM., Plumb, MR. et al. (2012). Altering murine leukemia virus integration through disruption of the BET and TAF protein family interaction. Cell Rep 3: 88–96.

28. LaFeve, MC., Varshney, GH., Gildea, DE., Wolfsberg, TG., Baxevanis, AD. and Burgess, SM. (2014). MLV integration site selection is driven by strong enhancers and active promoters. Nature 511: 4257–4269.

29. Shulha, HP., Meltzer, P., Margulies, EH., Weng, Z., Chen, X., Parkinson, HD. et al. (2007). Comparative analysis of transposable element vector systems in human cells. Mol Ther 12: 227–237.

30. Li, X., Ewis, H., Hice, RH., Malani, N., Parker, N., Zhou, L. et al. (2013). A resurrected mammalian hAT transposable element and a closely related isolated element are highly active in human cell culture. J Biol Chem 288: 12992–12999.

31. Hackett, PB., Largaespada, DA. and Cooper, LJ. (2010). A transposon and transposase system for targeted genome modification in vertebrates. EMBO J 29: 401–409.

32. Atwell, S., Watanabe, K., Pacifici, R., Comer, J., Auerbach, B., Aldrich, D. et al. (2009). Genome editing of Kaposi sarcoma-associated herpesvirus and murine gammaherpesvirus 68 using the piggyBac transposon system. PNAS 106: 2677–2682.

33. Li, MA., Pettitt, SJ., Eckert, S., Ning, Z., Rice, S., Cadiñanos, J. et al. (2006). BET bromo- and extraterminal domain chromatin regulators serve as cofactors for the leukemia virus integration machinery. Nat Struct Mol Biol 13: 926–932.

34. Liang, Q., Kong, J., Stalker, J. and Bradley, A. (2009). Bromodomain and extraterminal domain chromatin remodeling complexes are essential for transcriptional repression of murine leukemia virus. EMBO J 28: 731–743.

35. Filippakopoulos, P., Qi, J., Picard, S., Shens, Y., Smith, WB., Fedorov, O. et al. (2010). Selective inhibition of BET bromodomains. Nature 468: 1067–1073.

36. Ayer, S., Swapova, CV., Malani, N., Aramini, JM., Schneider, WM., Plumb, MR. et al. (2012). Altering murine leukemia virus integration through disruption of the BET and TAF protein family interaction. Cell Rep 3: 88–96.

37. Liu, G., Geurts, AM., Yae, K., Srinivasan, AR., Fahrenkrug, SC. and Largaespada, DA. et al. (2005). Target-site preferences of Sleeping Beauty transposons. Mol Cell 19: 161–173.

38. Yant, SR., Wu, X., Huang, Y., Garrison, B., Burgess, SM. and Ray, MA. (2005). High-resolution genome-wide mapping of transposition in mammals. Mol Cell Biol 25: 2085–2094.

39. Grabundzija, I., Irgang, M., Malet, L., Belay, E., Matra, J., Gogol-Döring, A. et al. (2010). Comparative analysis of transposable element vector systems in human cells. Mol Ther 18: 1200–1209.

40. Moldt, B., Miskey, C., Staunstrup, NH., Gogol-Döring, A., Bak, RO., Sharma, N. et al. (2011). Comparative genomic integration profiling of Sleeping Beauty transposons by high-sensitivity assays from integrated-defective vectomice in primary human cells. Mol Ther 19: 1499–1507.

41. Ammar, I., Gogol-Döring, A., Miskey, C., Chen, W., Cathermon, T., Izsak, Z. et al. (2012). Retargeting sleeping beauty transposon insertions by engineered zinc finger DNA binding domains. Mol Ther 20: 1852–1862.

42. Li, X., Evis, H., Rice, RH., Malani, N., Parker, N., Zhou, L. et al. (2013). A resurrected mammalian hAT transposable element and a closely related isolated element are highly active in human cell culture. J Biol Chem 288: 12992–12999.
