


RESEARCH

Open Access



# Sleeping Beauty transposon integrates into non-TA dinucleotides

Yabin Guo<sup>\*</sup> , Yin Zhang and Kaishun Hu

## Abstract

**Background:** Sleeping Beauty transposon (SB) has become an increasingly important genetic tool for generating mutations in vertebrate cells. It is widely thought that SB exclusively integrates into TA dinucleotides. However, this strict TA-preference has not been rigorously tested in large numbers of insertion sites that now can be detected with next generation sequencing. Li et al. found 71 SB insertions in non-TA dinucleotides in 2013, suggesting that TA dinucleotides are not the only sites of SB integration, yet further studies on this topic have not been carried out.

**Results:** In this study, we re-analyzed 600 million pairs of Illumina sequence reads from a high-throughput SB mutagenesis screen and identified 28 thousand SB insertions in non-TA sites. We recovered some of these non-TA sites using PCR and confirmed that at least a subset of the insertions at non-TA sites are real integrations. The consensus sequence of these non-TA sites shows an asymmetric pattern distinct from the symmetric pattern of the canonical TA sites. Perfect similarity between the downstream flanking sequence and SB transposon ends indicates there may be interaction between the transposon DNA binding domain of transposase and the target DNA.

**Conclusion:** The TA-preference of SB transposon is not as strict as what people had thought. And the SB integrations at non-TA sites might be guided by the interaction between the transposon DNA binding domain of SB transposase and the target DNA.

**Keywords:** Sleeping beauty, Transposon, Integration, Non-TA dinucleotide, Asymmetric, Non-palindromic

## Background

The Sleeping Beauty transposon (SB) is a DNA transposon of the Tc1/mariner family, which was constructed from a consensus transposable element sequence in the genome of the Salmonid subfamily of fish [1]. SB is capable of transposing in mammalian systems and has become a popular genetic tool for generating genome-wide mutations [2–5]. DNA transposons often have strong preferences for their target sites. For example, piggyBac strictly integrates into TTAA sites [6], while Hermes prefers T at the second position and A at the seventh position of its target site duplication (TSD) [7–9]. It has long been accepted that SB, along with all other transposons of the Tc1/mariner family, integrate only into TA dinucleotides, based on previous studies with limited integration events [10]. In 2005, Yant et al. identified more than 1300 SB insertions, which all targeted to TA dinucleotides [11], further confirming the strict TA

preference of SB integration. With the advent of next generation sequencing, even more SB integration sites have been sequenced in the context of transposon-mediated mutagenesis assays. However, rather than examining for other preferences in SB integration, most studies have rejected non-TA integration sites as probable artifacts, discarding reads lacking TA at the start [3, 4]. In 2013, Li et al. identified 71 SB insertions at non-TA sites (~1.6% of the total insertions), raising the idea that TA dinucleotides may not be the sole targets of SB [12]. Recently, the largest-to-date *ex vivo* SB mutagenesis screen was performed [5], in which more than 1100 integration libraries were sequenced, 600 million pairs of sequence reads were obtained, and 2 million SB target positions were identified, providing a uniquely rich dataset to mine for rare integration events. In this study, we re-analyzed these sequence reads and found that SB does in fact integrate into non-TA sites at a frequency of ~1.4%. Further analysis suggests that the SB insertions at non-TA sites might be a result of side reaction of the canonical integration.

\* Correspondence: [guoyb9@sysu.edu.cn](mailto:guoyb9@sysu.edu.cn)

Guangdong Provincial Key Laboratory of Malignant Tumor Epigenetics and Gene Regulation, Medical Research Center, Sun Yat-sen Memorial Hospital, Sun Yat-sen University, Guangzhou 510120, China



## Results

### Twenty-eight thousand integrations at non-TA sites were identified by re-analyzing previous sequencing data

To search for possible integrations in all genomic contexts besides TA sites, we re-trimmed the sequence reads from a previous study that sequenced 600 million read pairs from SB integration libraries [5], including reads that began with non-TA sequences as well as TA dinucleotides. After aligning to the mouse genome (mm10), 2,018,489 unique sites were identified, of which 28,794 insertions were at non-TA dinucleotides (Additional file 1: Table S1 and Figure S1). Sequence reads from the same libraries that were aligned to the same coordinate and same strand were considered duplicates as previously described [13]. Briefly, duplicates comprise sequence reads from independent insertions, cell duplications, as well as PCR amplifications. The insertions at non-TA sites were roughly 1.4% of the total insertions, which is similar to the frequency reported previously [12]. These insertions were termed *general matches* (Table 1 and Fig. 1a). To be more stringent, we assumed that those non-TA dinucleotides could be sequencing errors of TA dinucleotides. We replaced the first two nucleotides of the sequence reads with TA, and aligned them to the mouse genome again. All the sequences that aligned successfully were removed from the *general matches* with the remaining inserts resulting in a new set termed *high stringency matches*. The *high stringency matches* are 92.8% of *general matches* (Table 1), suggesting that sequencing error cannot account for most detected insertions in non-TA contexts.

**Table 1** Number of SB insertions identified at all 15 non-TA dinucleotides

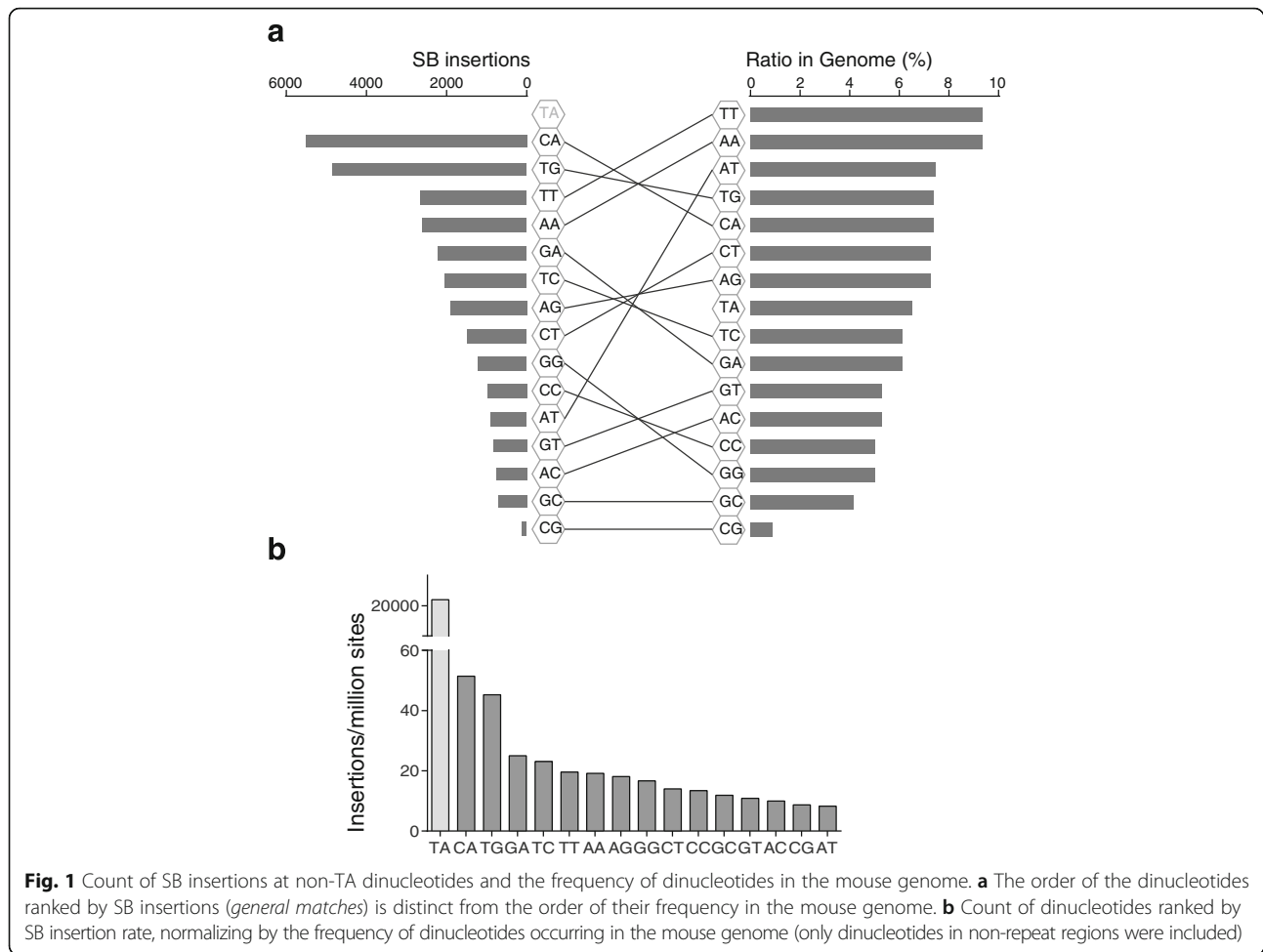
Dinucleotides	General match	High stringency match
CA	5511	4916
TG	4853	3963
TT	2659	2366
AA	2600	2536
GA	2214	2168
TC	2046	1929
AG	1913	1907
CT	1478	1464
GG	1213	1211
CC	977	970
AT	894	890
GT	836	832
AC	767	750
GC	719	717
CG	114	112
Total	28,794	26,731

We found that the integration frequencies at different dinucleotides are distinct, ranging from 0.0055% to 0.28% (Table 1). To identify dinucleotides enriched in SB insertions, we compared the distribution of insertions with the distribution of dinucleotides in the mouse genome (Fig. 1a, Additional file 1: Table S2). For some dinucleotides, the frequency of insertions mirrored the dinucleotide frequency. For example, CG dinucleotides are very rare in mammalian genomes, and CG dinucleotides had the fewest insertions. However, the order of dinucleotides ranked by SB insertions is distinct from that ranked by their occurrences in the mouse genome (Fig. 1a), indicating that some dinucleotides are more preferred by SB integration than others. Normalizing SB insertions by the frequency of the dinucleotides in the mouse genome, we observed TG/CA dinucleotides as the most preferred non-TA sites of SB integration (Fig. 1b), which may be because they are the most similar dinucleotides to TA (only one transition from TA).

### A subset of integrations at non-TA site were validated using PCR

To confirm whether the SB insertions at non-TA sites are real integrations or artifacts introduced by experimental design or data analysis, we picked nine insertion sites for PCR amplification. Since each library is a pool of cell clones with different integrations, only insertions highly represented in the population (insertions with high duplicates, Additional file 1: Table S1) were possible to be recovered (Additional file 1: Tables S1 and S3). For each insertion, two PCR reactions were performed, one for each strand orientation. Four out of nine insertions showed clear single bands in agarose gels for both primer pairs, which were then sequenced by Sanger sequencing. Figure 2 shows two sites recovered from lib155.11 and lib133.13. Junctions between transposon and genomic sequences were found in both directions and perfect TSD was identified for each site (Fig. 2). This result indicates at least some of the non-TA sites found in sequence analysis are bona fide integrations.

However, TSDs were not identified at other sites (Additional file 1: Figure S2). For example, the TA dinucleotides are preserved to the right, but not to the left of the SB insertion at a site in lib165.5 and lib160.12, thus no TSD was formed. These sites may be the result of aberrant integrations, which has been described in HIV-1 previously [14]. Because the LM-PCR for Illumina sequencing only detects the SB left end [5], our analysis cannot distinguish non-TA sites integrations from these aberrant integrations. To see if there are many sequences with patterns resembling what shown in Additional file 1: Figure S2, We examined the genomic sequences at all the non-TA sites and found that 1922 sequences have TA immediately after the target



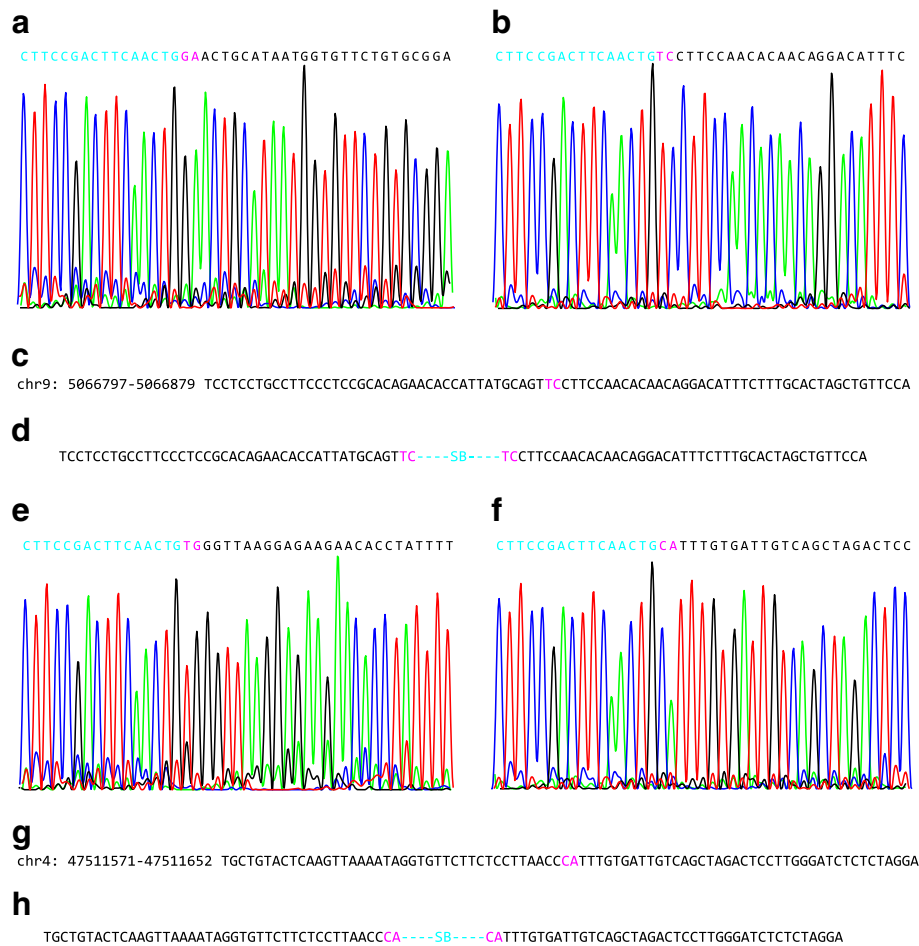
dinucleotides, while 1204 sequences have TA at the second position after the target dinucleotides. Even if the combined 3126 sites were the result of aberrant integrations, they are still only 0.1% of the total insertions at non-TA sites. Therefore, aberrant integrations do not contribute significantly to the insertions at non-TA sites identified in our analysis.

**The target site sequences of non-TA sites have an asymmetric pattern**

To find the sequence pattern of SB target sites, we extracted the SB target site sequences from the mouse genome by their coordinate and strand, and displayed the sequence preferences as sequence logos (Fig. 3). Like many transposons/retrotransposons, the SB target site sequences show a perfect symmetric pattern (palindrome) (Fig. 3a). Recently, Kirk et al. stated that the palindromic consensus sequence at the target sites of some retroviruses is a result of integrations occurring “in approximately equal proportions on the plus strand and the minus strand of the host genome” [15]. However, the symmetric sequence logos in some previous studies on transposon/

retrotransposon [8, 16], as well as the present study, were all made of sequences with fixed orientation (i.e. reverse complement sequences were taken for integrations in the minus strand). Actually, in Sleeping Beauty and Hermes transposon, or Tf1 retrotransposon, the consensus sequences at target sites are always palindromic, even if the sequence logos are made of sequences from plus or minus strand separately (data not shown).

We then generated sequence logos with target site sequences only from non-TA sites of *general matches*(Table 1). Strikingly, a distinct asymmetric pattern was revealed (Fig. 3b). The upstream sequence flanking the insertion is essentially unchanged from that of the canonical TA sites (Fig. 3a), whereas the downstream sequence shows a conserved motif, CAGTTGAA. Interestingly, this consensus sequence is exactly the same as the sequence of the SB transposon ends (Fig. 4). We also made sequence logos with sequences from different dinucleotides separately, and they all showed a similar pattern (Additional file 1: Figures S3-S5). To our knowledge, this is the first time that a target site consensus sequence has been shown to replicate a transposon sequence.



**Fig. 2** Recovery of SB integrations at non-TA sites. The SB integration site sequences were amplified in both orientations, using PCR and Sanger sequenced for integrations from lib155.11 (**a, b**) and lib133.13 (**e, f**). **c** and **g** The genomic sequences of the target sites. **d** and **h** The sequence patterns after integration. The cyan characters are the SB ends (**a**) and (**e**) are right ends; (**b**) and (**f**) are left ends. The black characters are genomic sequences and; the pink characters are the TSDs

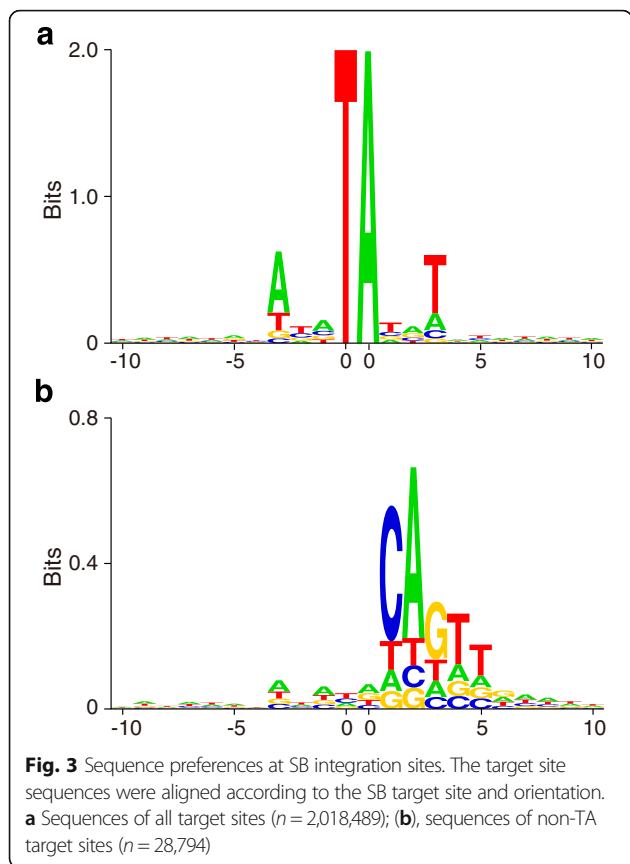
## Discussion

We analyzed 600 million pairs of Illumina sequence reads, allowing for reads with all other dinucleotides as well as TA dinucleotides, and identified 28 thousand SB insertions at non-TA dinucleotides, which is around 1.4% of the total insertions. These insertions were not randomly distributed at non-TA dinucleotides. On the contrary, they showed preferences for different dinucleotides, CA/TG being the most preferred ones (Fig. 1). To confirm that these integrations are real, we recovered some integration sites using PCR from the genomic DNA of certain integration libraries. Sanger sequencing showed that two out of four sites have perfect TSDs, indicating they are bona fide integrations (Fig. 2). We also identified two aberrant integrations (Additional file 1: Figure S2). Accounting for aberrant integrations and possible sequencing errors, we infer that the frequency of SB integration into non-TA sites is at least 1%.

We confirmed that Sleeping Beauty transposon can integrate into non-TA dinucleotides using large number of

insertions and a PCR recovery assay, thus modifying the well-established strict TA-preference of SB integration. Although ignoring these integration events, which account for ~1% of total integrations, does not change the conclusions of previous studies of SB integration, these low-frequency events have important consequences nonetheless. Recent studies have explored the use of SB in gene therapy [17, 18]. Rare integrations in non-TA sites need to be considered in these experiments to minimize the possibility of unexpected insertions disrupting gene functions.

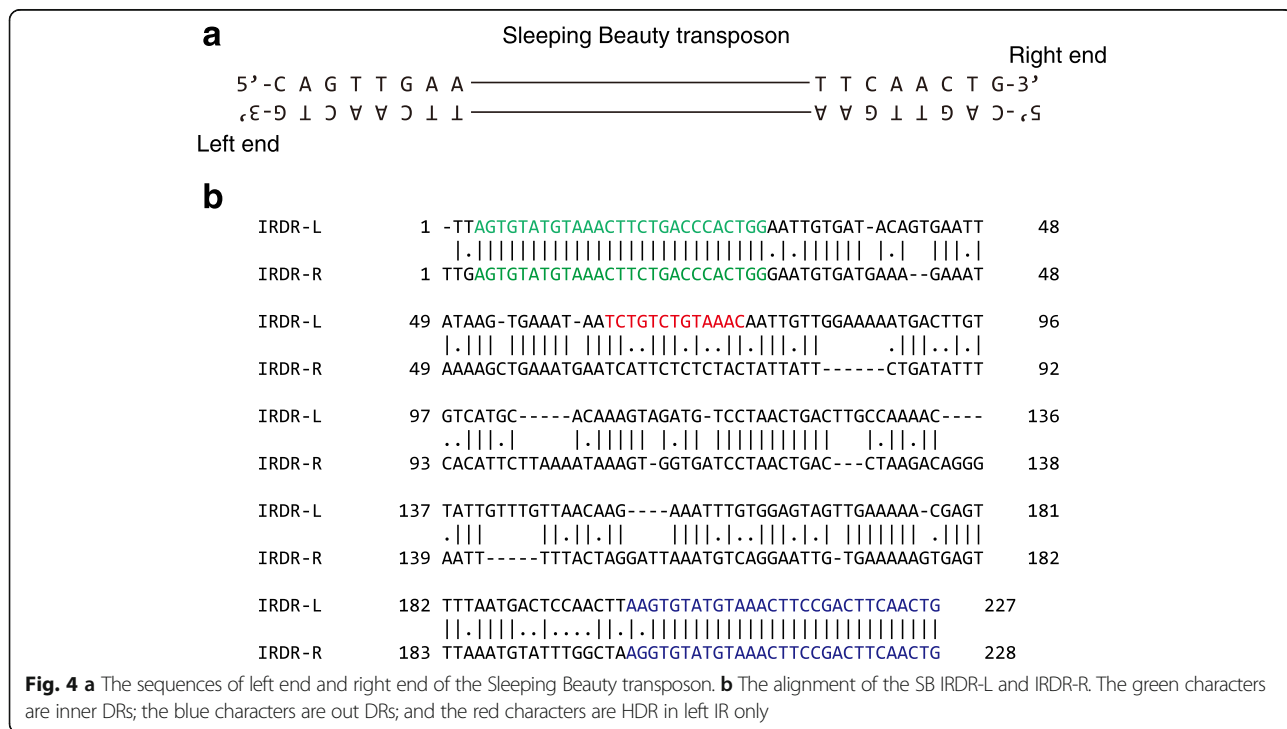
Unlike the general SB target sites, which have a palindromic consensus sequence, these non-TA sites show a distinct non-palindromic pattern. The consensus sequence downstream of the target site sequence (5'-CAGTTGAA-3') is exactly the same sequence as the SB transposon end. We considered if the identical sequence pattern of the consensus sequence at these target sites and the transposon end is an artifact: 1) the LM-PCR detects the



junction between SB left end and genomic DNA, but the consensus sequence pattern is to the right side of the target site; 2) the target site sequences are extracted from the mouse genome, but not from sequence reads of Illumina sequencing; 3) there are no homologous sequences of SB in the mouse genome; thus, even if the SB sequences were amplified in LM-PCR, they still could not be aligned to the mouse genome (although there are CAGTTGAA sequences in the mouse genome, they are not long enough for alignment); 4) the consensus sequence is a sequence of most frequent nucleotides, but is not necessarily a real sequence. Therefore, it is highly unlikely the detected sequence pattern is an artifact.

Moreover, the insertions at non-TA sites are not due to homologous recombination, because: 1) the consensus sequence is not long enough for homologous recombination; 2) the insertions have strong orientation bias (strand bias) (Fig. 3b); and 3) the TSDs found in PCR (Fig. 2) indicate true integration events.

We then focus on the 8-nucleotide box adjacent to the target dinucleotide at the right side of the target sites (called the R8 box), where the consensus sequence is located. Among the 28,794 non-TA sites, 12,732 (44%) R8 boxes are CAGnnnnn, 6276 (22%) R8 boxes are CAGTTnnn, and 694 (2.4%) R8 boxes are CAGTTGAA. When the insertion numbers were normalized by the occurrence of the sequences in the mouse genome, we observed that the integration efficiency increased dramatically as the similarity between the R8 box sequence and the transposon end





increased (Additional file 1: Figure S6). The frequency of SB integration in the context of the CAGTTGAA motif is >12,000/million sites, more than half of the integration frequency at canonical TA sites (Fig. 1b). It is possible that the low number of integrations at non-TA sites is not due to low integration efficiency at non-TA sites, but because there are far fewer CAGTTGAA sites than TA dinucleotides in the mouse genome.

Since the occurrence frequency for an 8-nucleotide sequence is roughly  $4^{-8}$ , the parity between these two sequences could not be a coincidence. Instead, it is far more likely to be a result of specific interaction. Therefore, we hypothesize that when certain genomic DNA sequences are similar to the SB end sequences, the transposon DNA binding domain of one SB transposase molecule in the pre-integration complex (PIC) may bind these genomic DNA strands as if they are transposon DNA strands, thus guiding the PIC to integrate into these positions, even there are no TA dinucleotides. This process might be a side reaction resembling the aberrant integration described recently by Wang et al. [19], in which the transposon DNA is circularized with one end attached to the other. Due to the small contribution of the side reaction to the entire integration collection, only when it is viewed separately from the canonical integration, can we notice its different property.

Why is there consensus sequence only at the downstream of the target dinucleotides, but not at both sides? SB is a transposon of the inverted repeat direct repeat (IRDR) subfamily [10, 20]. IRs are located at the two ends of the transposon. Each IR contains two DRs for binding transposase. However, the two IRs are asymmetric. A half direct repeat motif (HDR) which also can binds transposase was only found in the left IR (Fig. 4b). This could be the reason that only the transposon DNA binding domain of the transposase at the one side is capable of interacting with genomic DNA strands.

Notably, CA/TG are the most similar dinucleotides to TA, while the sequence pattern flanking them are the weakest (Additional file 1: Figures S3-S5), which indicates that there is no absolute barrier between the side reaction and the canonical reaction. The insertions at non-TA sites should be a pool of both canonical reaction and side reaction. When the target dinucleotides are more different from TA, the integration will rely more on the interaction between the DNA binding domain of transposase and the target DNA.

Finally, the SB transposase used for generating the present integrations is the hyperactive SB100X [21], and the SB transposase used by Li et al. is another hyperactive version, HSB16 [12, 22]. Probably, hyperactive versions of transposase tend to have less strict preference, which is to be answered by future studies.

## Conclusion

We have shown that SB transposon integrates into non-TA sites in addition to TA sites and suggest that these integrations are guided by interactions between SB transposase and genomic DNA sequences resembling the sequence of transposon end. Our finding improved the knowledge on the strict TA-preference of SB transposon.

## Methods

### Data source

The generation of the SB integration libraries in mouse BaF3 cells with SB100X transposase [21] and T2/Onc vector were described previously [5]. The Illumina sequencing results were deposited in NCBI Short Read Archive, <http://www.ncbi.nlm.nih.gov/sra>. Accession no. SRX1491647.

### Bioinformatic analyses

Scripts for sequence trimming and dinucleotide frequency counting were written in Perl. The trimmed sequences were aligned to the mouse genome (mm10) using Bowtie 2 [23]. The alignment output was filtered using a Perl script.

The target site sequences were extracted from the mouse genome using a Perl script. Reverse complement sequences were taken when the integration orientations are right-to-left (i.e. at minus strand). The target site sequence logos were generated using an application called DNALogo [24], which has been described in previous studies [8, 16]. The output PostScript (.ps) vector maps were converted to .pdf format in Adobe Illustrator.

### Recovering the SB target sites by PCR

Genomic DNA samples which had been extracted from the cell pools of SB integration libraries using DNeasy Blood & Tissue Kit (Qiagen) were obtained from Dr. Kathryn O'Donnell's lab at UT Southwestern Medical Center. In this study, the genomic DNA samples were used as templates. Primers were designed according to the genomic sequences flanking the SB target sites and the SB left/right ends. The primer pairs are [Primer 5, SB-left] and [SB-right, Primer 3] for insertions at plus strand, or [Primer 5, SB-right] and [SB-left, Primer 3] for insertions at minus strand (Additional file 1: Table S3). PCR reactions were performed using CloneAmp HiFi PCR Premix (Clontech). The PCR products were then sequenced by Sanger sequencing.

## Additional file

**Additional file 1:** Supplementary tables and figures. (ZIP 3928 kb)

### Abbreviations

IRDR: Inverted repeat direct repeat; PIC: Pre-integration complex; TSD: Target site repeat

**Acknowledgements**

We thank Dr. Kathryn O'Donnell of UT Southwestern Medical Center for her kind help. We thank Mr. Stephen Hung of Case Western Reserve University for his generous help. And we thank Dr. Henry Levin of NICHD, NIH for his valuable advices.

**Funding**

This work was supported by the Sun Yat-sen Memorial Hospital, Sun Yat-sen University. Finance No. 1320317002.

**Availability of data and materials**

The Illumina sequencing results are applicable in NCBI Short Read Archive, <http://www.ncbi.nlm.nih.gov/sra>. Accession no. SRX1491647.

**Authors' contributions**

YG conceived the idea for the project, performed the bioinformatics analyses and experiments, and wrote the manuscript. YZ analyzed the sequencing data and wrote the manuscript with YG. KH did the PCR assay with YG. All authors read and approved the final manuscript.

**Ethics approval and consent to participate**

Not applicable.

**Consent for publication**

Not applicable.

**Competing interests**

The authors declare that they have no competing interests.

**Publisher's Note**

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 21 December 2017 Accepted: 30 January 2018

Published online: 07 February 2018

**References**

- Ivics Z, Hackett PB, Plasterk RH, Izsvák Z. Molecular reconstruction of Sleeping Beauty, a Tc1-like Transposon from fish, and its transposition in human cells. *Cell*. 1997;91:501–10.
- Dupuy AJ, Akagi K, Largaespada DA, Copeland NG, Jenkins NA. Mammalian mutagenesis using a highly mobile somatic Sleeping Beauty transposon system. *Nature*. 2005;436:221–6.
- Keng VW, Villanueva A, Chiang DY, Dupuy AJ, Ryan BJ, Matisse I, et al. A conditional transposon-based insertional mutagenesis screen for hepatocellular carcinoma-associated genes in mice. *Nat Biotechnol*. 2009;27:264–74.
- O'Donnell KA, Keng VW, York B, Reineke EL, Seo D, Fan D, et al. PNAS plus: a Sleeping Beauty mutagenesis screen reveals a tumor suppressor role for Ncoa2/Src-2 in liver cancer. *Proc Natl Acad Sci*. 2012;109:E1377–86.
- Guo Y, Updegraff BL, Park S, Durakoglugil D, Cruz VH, Maddux S, et al. Comprehensive ex vivo transposon mutagenesis identifies genes that promote growth factor independence and leukemogenesis. *Cancer Res*. 2016;76:773–86.
- Cary LC, Goebel M, Corsaro BG, Wang HG, Rosen E, Fraser MJ. Transposon mutagenesis of baculoviruses: analysis of Trichoplusia Ni transposon IFP2 insertions within the FP-locus of nuclear polyhedrosis viruses. *Virology*. 1989;172:156–69.
- Gangadharan S, Mularoni L, Fain-Thornton J, Wheelan SJ, Craig NL. DNA transposon Hermes inserts into DNA in nucleosome-free regions in vivo. *Proc Natl Acad Sci U S A*. 2010;107:21966–72.
- Guo Y, Park JM, Cui B, Humes E, Gangadharan S, Hung S, et al. Integration profiling of gene function with dense maps of transposon integration. *Genetics*. 2013;195:599–609.
- Evertts AG, Plymire C, Craig NL, Levin HL. The hermes transposon of *Musca Domestica* is an efficient tool for the mutagenesis of *Schizosaccharomyces pombe*. *Genetics*. 2007;177:2519–23.
- Plasterk RH, Izsvák Z, Ivics Z. Resident aliens: the Tc1/mariner superfamily of transposable elements. *Trends Genet*. 1999;15:326–32.
- Yant SR, Wu X, Huang Y, Garrison B, Burgess SM, Kay MA. High-resolution genome-wide mapping of Transposon integration in mammals. *Mol Cell Biol*. 2005;25:2085–94.
- Li X, Ewis H, Hice RH, Malani N, Parker N, Zhou L, et al. A resurrected mammalian hAT transposable element and a closely related insect element are highly active in human cell culture. *Proc Natl Acad Sci*. 2013;110:E478–87.
- Guo Y, Levin HL. High-throughput sequencing of retrotransposon integration provides a saturated profile of target activity in *Schizosaccharomyces pombe*. *Genome Res*. 2010;20:239–48.
- Varadarajan J, McWilliams MJ, Hughes SH. Treatment with suboptimal doses of raltegravir leads to aberrant HIV-1 integrations. *Proc Natl Acad Sci*. 2013;110:14747–52.
- Kirk PDW, Huvet M, Melamed A, Maertens GN, Bangham CRM. Retroviruses integrate into a shared, non-palindromic DNA motif. *Nat Microbiol*. 2016;2:16212.
- Chatterjee AG, Esnault C, Guo Y, Hung S, McQueen PG, Levin HL. Serial number tagging reveals a prominent sequence preference of retrotransposon integration. *Nucleic Acids Res*. 2014;42:8449–60.
- Aronovich EL, Mcivor RS, Hackett PB. The Sleeping Beauty transposon system: a non-viral vector for gene therapy. *Hum Mol Genet*. 2011;20:14–20.
- Hou X, Du Y, Deng Y, Wu J, Cao G. Sleeping Beauty transposon system for genetic etiological research and gene therapy of cancers Sleeping Beauty transposon system for genetic etiological research and gene therapy of cancers. *Cancer Biol Ther*. 2015;16:8–16.
- Wang Y, Pryputniewicz-Dobrzinska D, Nagy EE, Kaufman CD, Singh M, Yant S, et al. Regulated complex assembly safeguards the fidelity of Sleeping Beauty transposition. *Nucleic Acids Res*. 2017;45:311–26.
- Izsvák Z, Khare D, Behlke J, Heinemann U, Plasterk RH, Ivics Z. Involvement of a bifunctional, paired-like DNA-binding domain and a transpositional enhancer in Sleeping Beauty transposition. *J Biol Chem*. 2002;277:34581–8.
- Mates L, Chuah MK, Belay E, Jerchow B, Manoj N, Acosta-Sanchez A, et al. Molecular evolution of a novel hyperactive Sleeping Beauty transposase enables robust stable gene transfer in vertebrates. *Nat Genet*. 2009;41:753–61.
- Baus J, Liu L, Heggstad AD, Sanz S, Fletcher BS. Hyperactive transposase mutants of the Sleeping Beauty transposon. *Mol Ther*. 2005;12:1148–56.
- Langmead B, Salzberg SL. Fast gapped-read alignment with bowtie 2. *Nat Methods*. 2012;9:357–9.
- Guo Y. DNALogo: a smart mini application for generating DNA sequence logos. *bioRxiv* [Internet]. 2016; Available from: <http://biorxiv.org/content/early/2016/12/27/096933>

Submit your next manuscript to BioMed Central and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at  
[www.biomedcentral.com/submit](http://www.biomedcentral.com/submit)

