Supplemental Information

Shifting Retroviral Vector Integrations Away

from Transcriptional Start Sites via DNA-Binding

Protein Domain Insertion into Integrase

Jung-soo Nam, Ji-eun Lee, Kwang-hee Lee, Yeji Yang, Soo-hyun Kim, Gyu-un Bae, Hohsuk Noh, and Kwang-il Lim

Tables

Table S1. Amino acid sequences of promising mutant integrases.

Mutant	Amino acid sequences
A2	IENSSPYTSEHFHYTVTDIKDLTKLGAIYDKTKKYWVYQGKPVMPDQFTFELLDFLHQ LTHLSFSKMKALLERSHSPYYMLNRDRTLKNITETCKACAQVNASKSAVKQGTRVRG HRPGTHWEIDFTEIKPGLYGYKYLLVFIDTFSGWIEAFPTKKETAKVVTKKLLEEIFPRF GMPQVLGTDNGPAFVSKVSQTVADLLGIDWKLHCAYRPQSSGQVERMNRTIKETLT KLTLATGSRDWVLLLPLALYRARNTPGPHGLTPYEILYGAPPLSATGEKPYKCPECG KSFSRSDHLAEHQRTHTGEKPYACPECGKSFSQSGDLRRHQRTHTGEKPYKCPEC GKSFSRKDNLKNHQRTHTGEKPYKCPECGKSFSDPGALVRHQRTHTGKKTSGQAG QATGEKPASPPLVNFPDPDMTRVTNSPSLQAHLQALYLVQHEVWRPLAAAYQEQLD RPVVPHPYRVGDTVWVRRHQTKNLEPRWKGPYTVLLTTPTALKVDGIAAWIHAAHV KAADPGGGPSSRLTWRVQRSQNPLKIRLTREAP
B2	IENSSPYTSEHFHYTVTDIKDLTKLGAIYDKTKKYWVYQGKPVMPDQFTFELLDFLHQ LTHLSFSKMKALLERSHSPYYMLNRDRTLKNITETCKACAQVNASKSAVKQGTRVRG HRPGTHWEIDFTEIKPGLYGYKYLLVFIDTFSGWIEAFPTKKETAKVVTKKLLEEIFPRF GMPQVLGTDNGPAFVSKVSQTVADLLGIDWKLHCAYRPQSSGQVERMNRTIKETLT KLTLATGSRDWVLLLPLALYRARNTPGPHGLTPYEILYGAPPLSATGEKPMAERPFQ CRICMRNFSRSDALSRHIRTHTGEKPFACDICGRKFAQSGDLTRHTKIHTGGQRPFQ CRICMRNFSQSGDLTRHIRTHTGEKPFACDICGRKFATSGHLSRHTKIHTGGGGSAS PPLVNFPDPDMTRVTNSPSLQAHLQALYLVQHEVWRPLAAAYQEQLDRPVVPHPYR VGDTVWVRRHQTKNLEPRWKGPYTVLLTTPTALKVDGIAAWIHAAHVKAADPGGGP SSRLTWRVQRSQNPLKIRLTREAP

The sequence of integrase is shown in black and the sequence of the inserted protein domains is shown in blue.

Table S2. Total numbers of analyzed integration sites.

ZFD insertion mutants		PBS mutants		Double mutants	
Wild-type	197	PBS2	477	5GA + A2	126
A1	84	PBS3	432	5GA + B2	97
A2	100	PBS4	239	9CG + A2	144
A3	76	PBS8	176	9CG + B2	174
A4	79	PBS9	288	PBS4 + A2	35
A5	83	5GA	557		
B1	93	5GC	538		
B2	120	5GT	375		
B3	43	9CA	489		
B4	76	9CG	450		
B5	75	9CT	471		

Table S3. List of human oncogenes near retroviral integration sites.

	Wild-type	A3	B1	B4	B5
Total number of integrations	197	76	93	76	75
Number of integrations into oncogenic spots	6	1	2	1	1
Frequency of integrations into oncogenic spots (%)	3.05	1.32	2.15	1.32	1.33
Oncogenes adjacent integration sites	HOXA9	PBRM1	HOXD11	PBRM1	ELK4
	HOXA11				
	CCDC6				
	Myc				
	PSIP1				
	PTPN11				

Mutant vector B1 was integrated twice into the genomic region near *HOXD11*.

Table S4. Frequency of removal of the 5′ overhang dinucleotides in the 3′ LTR ends of proviruses.

	Frequency (%)	P-value (compared with Wild-type)
Wild-type	94.4	
A1	41.7	< 2.22E-16
A2	39.0	< 2.22E-16
A3	35.5	< 2.22E-16
A4	55.7	6.10E-15
A5	54.2	6.23E-16
B1	61.3	9.01E-13
B2	39.2	< 2.22E-16
B3	67.4	5.10E-06
B4	36.8	< 2.22E-16
B5	42.7	< 2.22E-16

Statistical significance of the difference in frequency of the removal of the 5′ overhang dinucleotides between wild-type and mutant vectors was determined by the chi-square test. The *P*-value of 2.22E-16 is the smallest value that can be numerically obtained in the statistical package R. *P*-values lower than this value are all displayed as < 2.22E-16 in the package.

Figures

Figure S1. FACS plots showing GFP expression by HEK 294T cells transduced with wild-type and the mutant retroviral vectors at seven days post transduction
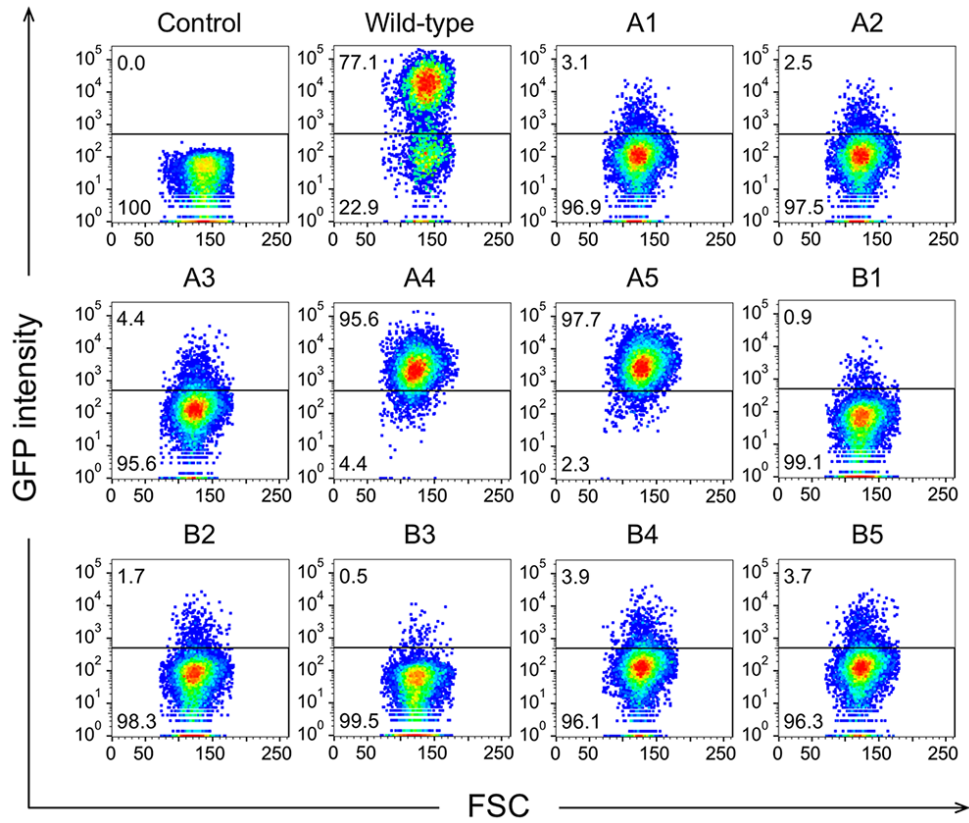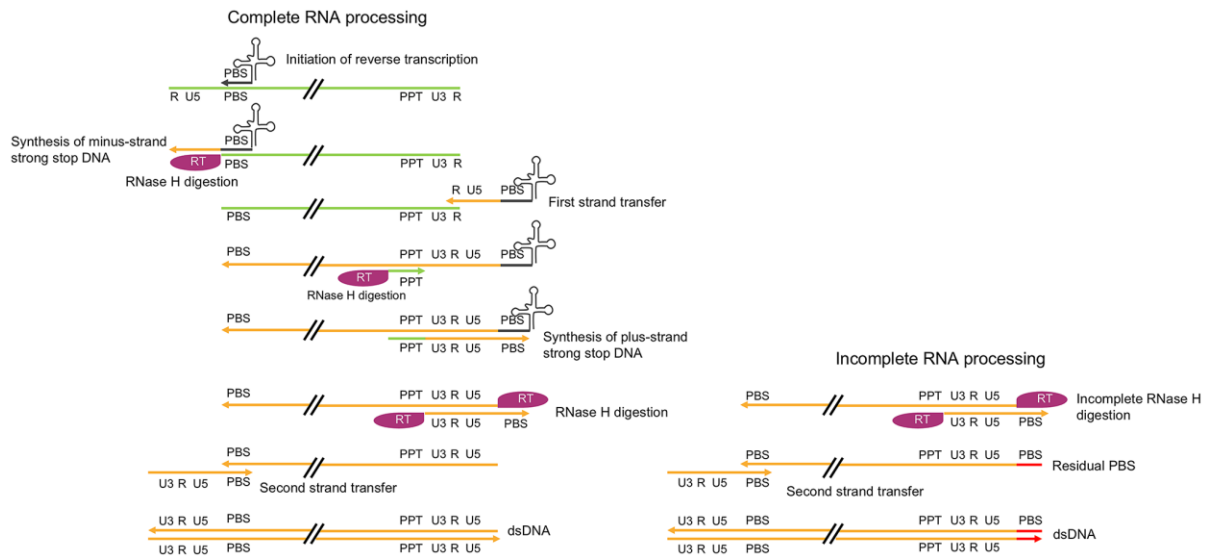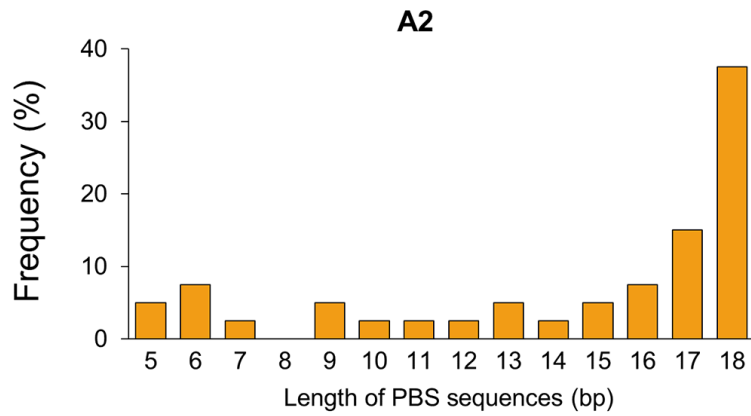
Figure S2. Schematic representation of reverse transcription and RNA processing during retroviral infection.
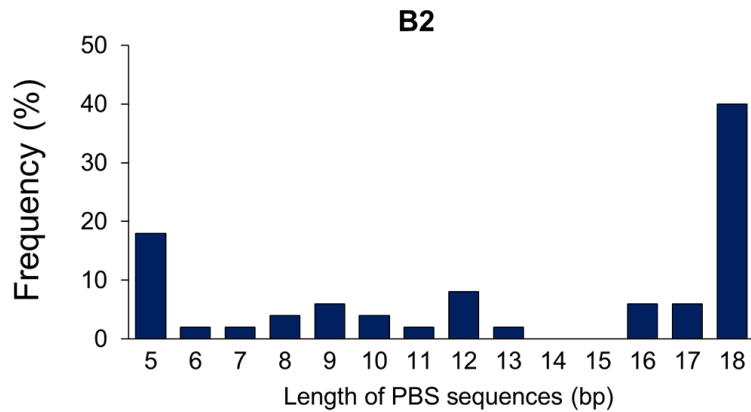

Incomplete RNA processing by the RNase H domain of reverse transcriptase can leave residual PBS sequences next to the end of U5 of the 3′ LTR (right panel). These residual PBS sequences are indicated in red. Host tRNA functioning as a primer for reverse transcription is indicated in black. Viral RNA and DNA molecules are indicated in green and orange, respectively. The left panel shows complete RNA processing, and the right panel shows incomplete RNA processing.

Figure S3. Frequency of the formation of residual PBS sequences in the 3´ LTR of proviruses for the A2 mutant.

This figure shows the frequency of the occurrence of residual PBS sequences of different length in the 3´ LTR ends of proviruses for the A2 mutant vector. The numbers in the x-axis denote the length (in bp) of residual PBS sequences.

Figure S4. Frequency of formation of residual PBS sequences in the 3´ LTR of proviruses for the B2 mutant.
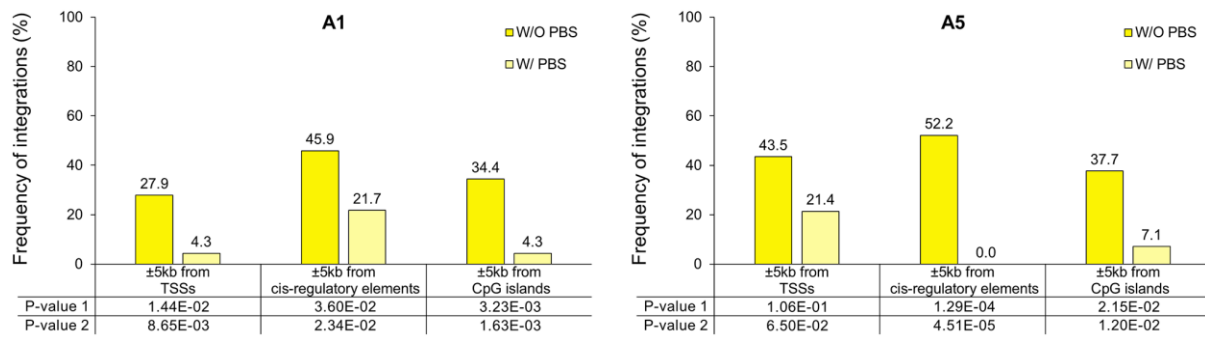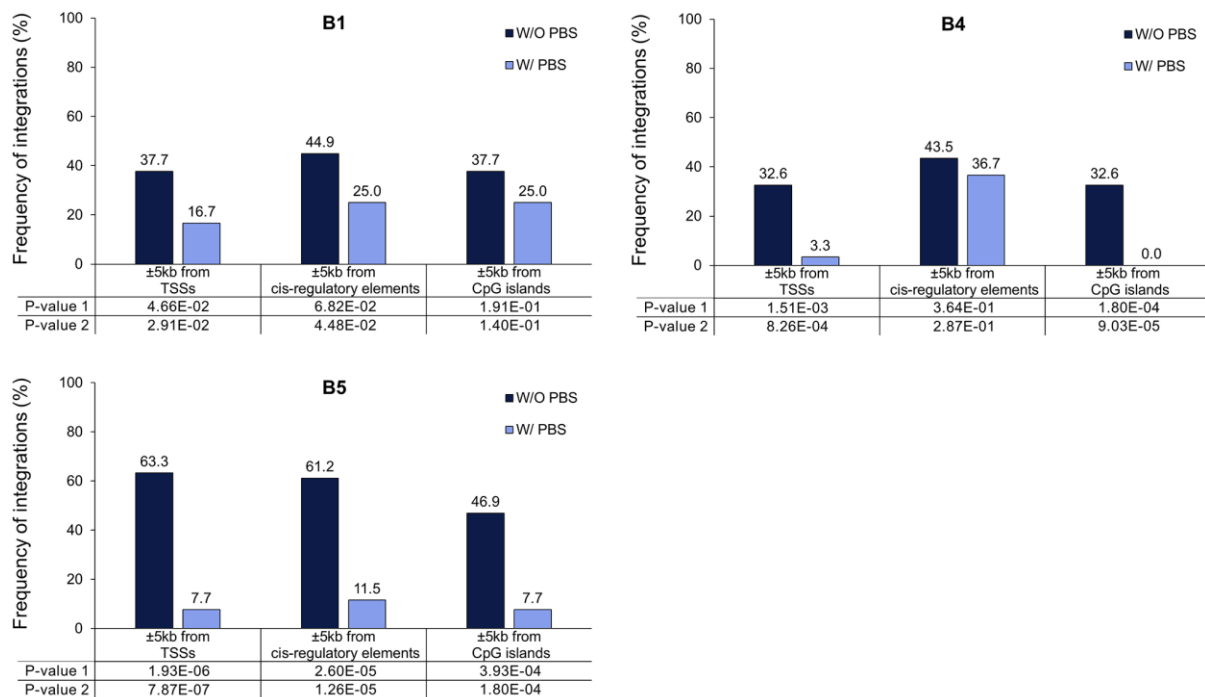

This figure shows the frequency of the occurrence of residual PBS sequences of different length in the 3´ LTR ends of proviruses for the B2 mutant vector case. The numbers in the x-axis denote the length (in bp) of residual PBS sequences.

Figure S5. Frequency of retroviral integrations into different human genomic regions in the presence and absence of residual PBS sequences in the provirus 3´ end for mutants with group A ZFDs.

Whether the presence of residual PBS sequences is significantly associated with a lower frequency of vector integrations into each type of genomic regions was tested using Fisher's exact test (*P*-value 1) and Boschloo's test (*P*-value 2). The two resulting P-values are presented in the frequency plot.

Figure S6. Frequency of retroviral integrations into different human genomic regions in the presence and absence of residual PBS sequences in the provirus 3´ end for the mutants with group B ZFDs.
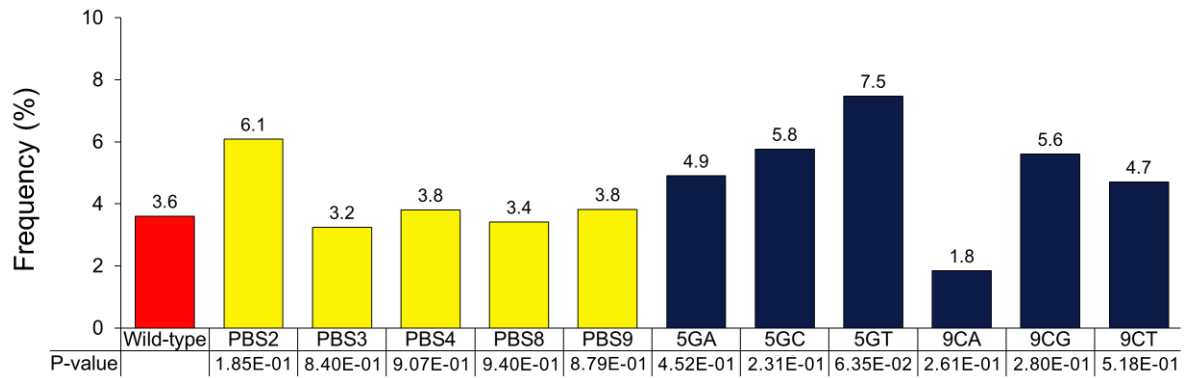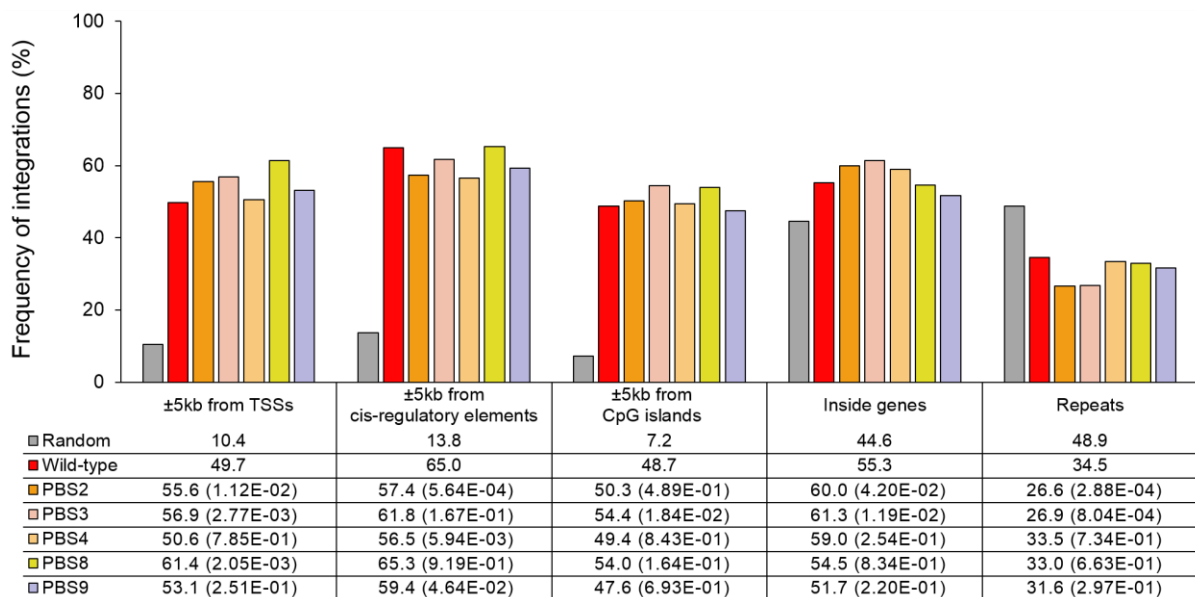

Whether the presence of residual PBS sequences is significantly associated with a lower frequency of vector integrations into each type of genomic regions was tested using Fisher's exact test (*P*-value 1) and Boschloo's test (*P*-value 2). The two resulting P-values are presented in the frequency plot.

Figure S7. Frequency of proviruses with residual PBS sequences.

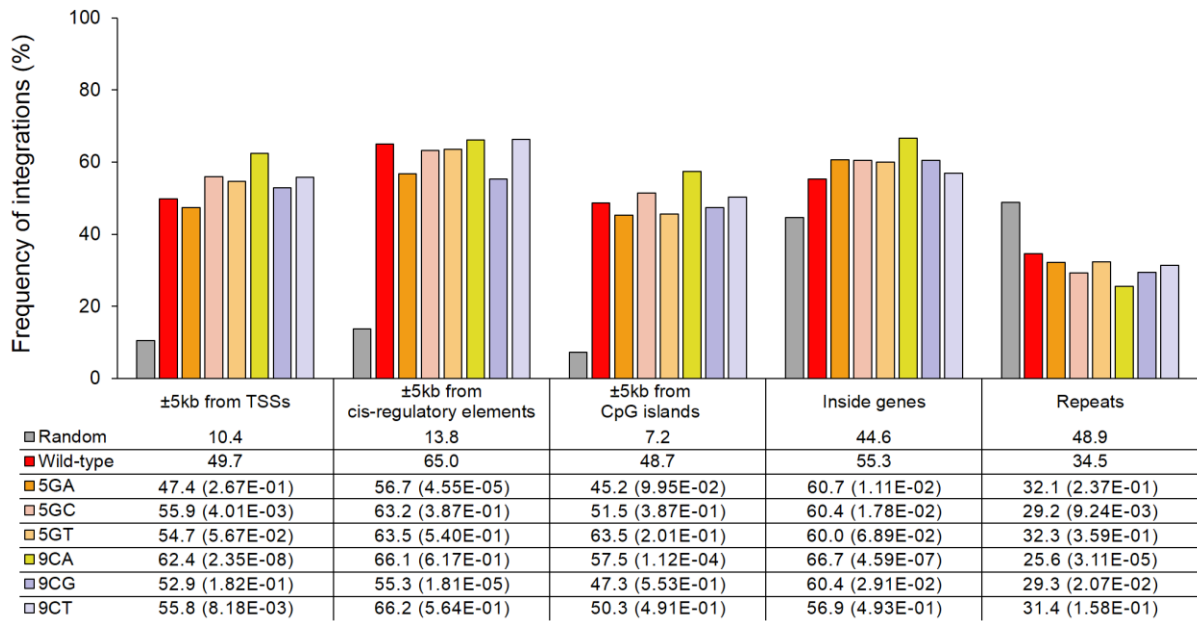| P-value	Wild-type	PBS2	PBS3	PBS4	PBS8	PBS9	5GA	5GC	5GT	9CA	9CG	9CT
		1.85E-01	8.40E-01	9.07E-01	9.40E-01	8.79E-01	4.52E-01	2.31E-01	6.35E-02	2.61E-01	2.80E-01	5.18E-01

Statistical significance for the difference in frequency of residual PBS sequences between wild-type and mutant vectors is indicated by *P*-values that were obtained by the chi-square test.

Figure S8. Frequency of retroviral integrations into different human genomic regions for mutants with additional PBSs.

	±5kb from TSSs	±5kb from cis-regulatory elements	±5kb from CpG islands	Inside genes	Repeats
■ Random	10.4	13.8	7.2	44.6	48.9
■ Wild-type	49.7	65.0	48.7	55.3	34.5
▢ PBS2	55.6 (1.12E-02)	57.4 (5.64E-04)	50.3 (4.89E-01)	60.0 (4.20E-02)	26.6 (2.88E-04)
▢ PBS3	56.9 (2.77E-03)	61.8 (1.67E-01)	54.4 (1.84E-02)	61.3 (1.19E-02)	26.9 (8.04E-04)
▢ PBS4	50.6 (7.85E-01)	56.5 (5.94E-03)	49.4 (8.43E-01)	59.0 (2.54E-01)	33.5 (7.34E-01)
▢ PBS8	61.4 (2.05E-03)	65.3 (9.19E-01)	54.0 (1.64E-01)	54.5 (8.34E-01)	33.0 (6.63E-01)
▢ PBS9	53.1 (2.51E-01)	59.4 (4.64E-02)	47.6 (6.93E-01)	51.7 (2.20E-01)	31.6 (2.97E-01)

Random integrations were also computationally generated using the QuickMap tool. The relevant frequencies of the random and experimental retroviral integrations into different genomic regions were quantified using the QuickMap tool. Statistical significance of the differences between wild-type and mutant integration patterns is indicated by *P*-values that were obtained by the chi-square test. P-values for the comparison with wild-type are shown in parentheses.

Figure S9. Frequency of retroviral integrations into different human genomic regions for mutants with PBS point mutation.

	±5kb from TSSs	±5kb from cis-regulatory elements	±5kb from CpG islands	Inside genes	Repeats
□ Random	10.4	13.8	7.2	44.6	48.9
■ Wild-type	49.7	65.0	48.7	55.3	34.5
□ 5GA	47.4 (2.67E-01)	56.7 (4.55E-05)	45.2 (9.95E-02)	60.7 (1.11E-02)	32.1 (2.37E-01)
□ 5GC	55.9 (4.01E-03)	63.2 (3.87E-01)	51.5 (3.87E-01)	60.4 (1.78E-02)	29.2 (9.24E-03)
□ 5GT	54.7 (5.67E-02)	63.5 (5.40E-01)	63.5 (2.01E-01)	60.0 (6.89E-02)	32.3 (3.59E-01)
□ 9CA	62.4 (2.35E-08)	66.1 (6.17E-01)	57.5 (1.12E-04)	66.7 (4.59E-07)	25.6 (3.11E-05)
□ 9CG	52.9 (1.82E-01)	55.3 (1.81E-05)	47.3 (5.53E-01)	60.4 (2.91E-02)	29.3 (2.07E-02)
□ 9CT	55.8 (8.18E-03)	66.2 (5.64E-01)	50.3 (4.91E-01)	56.9 (4.93E-01)	31.4 (1.58E-01)

Random integrations were also computationally generated using the QuickMap tool. The relevant frequencies of the random and experimental retroviral integrations into different genomic regions were determined using the QuickMap tool. Statistical significance of the differences between wild-type and mutant integration patterns is indicated by *P*-values that were obtained by the chi-square test. *P*-values for the comparison with wild-type are shown in parentheses.