Molecular Therapy
Oncolytics
Original Article

Tagmentation-based analysis reveals the clonal behavior of CAR-T cells in association with lentivector integration sites

Jaeryuk Kim,[1,2,3,12] Miyoung Park,[4,5,6,12] Gyungwon Baek,[4,5,6] Joo-Il Kim,[8,9] Euna Kwon,[9] Byeong-Cheol Kang,[8,9,10,11] Jong-Il Kim,[1,2,3,4] and Hyoung Jin Kang[4,5,6,7]

[1]Genomic Medicine Institute, Medical Research Center, Seoul National University, Seoul, Republic of Korea; [2]Department of Biomedical Sciences, Seoul National University Graduate School, Seoul, Republic of Korea; [3]Department of Biochemistry and Molecular Biology, Seoul National University College of Medicine, Seoul, Republic of Korea; [4]Seoul National University Cancer Research Institute, Seoul, Republic of Korea; [5]Department of Pediatrics, Seoul National University College of Medicine, Seoul, Republic of Korea; [6]Seoul National University Children's Hospital, Seoul, Republic of Korea; [7]Wide River Institute of Immunology, Hongcheon, Republic of Korea; [8]Graduate School of Translational Medicine, Seoul National University College of Medicine, Seoul, Republic of Korea; [9]Department of Experimental Animal Research, Biomedical Research Institute, Seoul National University Hospital, Seoul, Republic of Korea; [10]Biomedical Center for Animal Resource and Development; Seoul National University College of Medicine, Seoul, Republic of Korea; [11]Designed Animal Resource Center, Institute of GreenBio Science Technology, Seoul National University, Pyeongchang-gun, Gangwon-do, Republic of Korea

Integration site (IS) analysis is essential in ensuring safety and efficacy of gene therapies when integrating vectors are used. Although clinical trials of gene therapy are rapidly increasing, current methods have limited use in clinical settings because of their lengthy protocols. Here, we describe a novel genome-wide IS analysis method, "detection of the integration sites in a time-efficient manner, quantifying clonal size using tagmentation sequencing" (DIStinct-seq). In DIStinct-seq, a bead-linked Tn5 transposome is used, allowing the sequencing library to be prepared within a single day. We validated the quantification performance of DIStinct-seq for measuring clonal size with clones of known IS. Using *ex vivo* chimeric antigen receptor (CAR)-T cells, we revealed the characteristics of lentiviral IS. We then applied it to CAR-T cells collected at various times from tumor-engrafted mice, detecting 1,034–6,233 IS. Notably, we observed that the highly expanded clones had a higher integration frequency in the transcription units and vice versa in genomic safe harbors (GSH). Also, in GSH, persistent clones had more frequent IS. Together with these findings, the new IS analysis method will help to improve the safety and efficacy of gene therapies.

INTRODUCTION

Integrating vectors are routinely used to allow permanent expression during gene therapies. However, hematologic malignancies induced by gammaretroviral vectors have raised safety issues concerning insertional mutagenesis.[1,2] Accordingly, integration site (IS) analysis has become essential for monitoring unusual clonal expansion events in gene therapies using integrating vectors.[3,4] Beside the gene therapy fields, IS analysis has also been used to investigate the clonal dynamics of HIV-infected cells, which could be crucial for treating HIV latent reservoirs.[5–7]

Because the advent of next-generation sequencing (NGS) technology, the throughput of IS detection has significantly increased while decreasing labor compared with that in colony formation and the Sanger sequencing method.[8] However, current methods, such as linear-amplification mediated PCR (LAM-PCR)[9] or ligation-mediated PCR (LM-PCR),[10] are restricted by complex and lengthy protocols spending 3–7 days. Furthermore, these methods require a large amount of DNA, ranging from 500 ng to 3 μg. That is often problematic in clinical trials with limited DNA, where blood samples should be allocated for various laboratory tests. To decrease the time required, several studies have used transposases to perform DNA fragmentation and adapter ligation simultaneously. A quantitative IS detection method using Mu transposase has been reported,[11] but it requires a relatively large input DNA of 2 μg or more, which limits its application. A recent study has demonstrated that a tag-PCR method based on the Tn5 transposase can detect IS with only 50 ng of DNA.[12] However, this method has not been applied to quantitative analyses. Therefore, to address the expansion of gene therapy trials, including chimeric antigen receptor (CAR)-T cell therapies, a more clinically applicable method for IS analysis with quantitative capabilities is required.

Received 24 August 2022; accepted 12 May 2023;
https://doi.org/10.1016/j.omto.2023.05.004.

[12]These authors contributed equally

Correspondence: Jong-Il Kim M.D., Ph.D. Department of Biomedical Sciences, Seoul National University Graduate School 103 Daehak-ro, Jongno-gu, Seoul 03080, Republic of Korea
E-mail: jongil@snu.ac.kr
Correspondence: Hyoung Jin Kang, M.D., Ph.D. Department of Pediatrics, Seoul National University College of Medicine, 101 Daehak-ro, Jongno-gu, Seoul 03080, Republic of Korea
E-mail: kanghj@snu.ac.kr

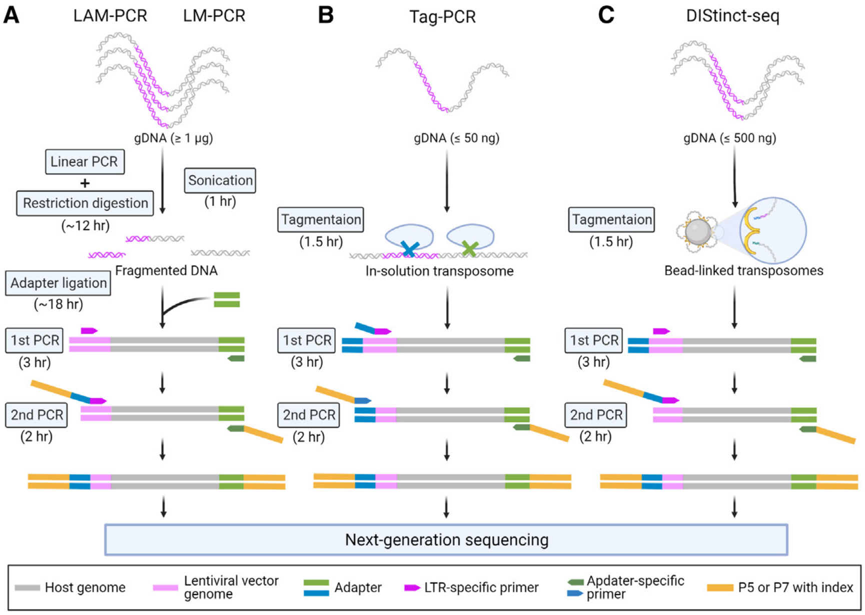


Figure 1. Schematic comparison of DIStinct-seq with other ISs analysis methods

(A) Conventional IS analysis methods (A) start with linear PCR followed by DNA fragmentation using restriction enzymes for LAM-PCR and sonication for LM-PCR using 1 µg or more DNA. Adapters are ligated to fragmented DNA to provide additional sites for PCR primer annealing. After that, nested PCR is performed to specifically amplify host-vector chimeric fragments and add the adapter and index sequences required for NGS in the Illumina platform. (B) In Tag-PCR, genomic DNA of 50 ng is tagmented with in-solution transposomes to achieve simultaneous fragmentation and adapter ligation. Two rounds of PCR amplification are performed to amplify host-vector chimeric fragments and attach the required NGS sequences for the Illumina platform. (C) In DIStinct-seq, up to 500 ng genomic DNA per sample is tagmented by bead-linked transposomes. Nested PCR is carried out to specifically amplify host-vector chimeric fragments and attach adapter and index sequences. All three methods require NGS to detect IS, which corresponds with junctions between the host and the vector genome.

In recent years, IS analysis has helped to evaluate the clonal dynamics in CAR-T therapies, where persistent clones play an instrumental role in therapeutic efficacy.[13,14] For example, IS analysis of CAR-T cells revealed that integration into the TET2 locus caused them to become dominant clones, leading to complete remission in chronic lymphocytic leukemia.[15] Therefore, in terms of efficacy, it is crucial to reveal genomic regions associated with the clonal behavior of CAR-T cells. These genomic regions will be potential targets for gene editing, with rapidly advancing technologies such as CRISPR-Cas9.[16] In this regard, a few studies on CAR-T therapy have identified potential targets mainly at the gene level.[15,17,18] However, to gain a deeper understanding of how IS interacts with clonal behavior, a comprehensive approach that is not limited to specific genes would be needed.

In this study, we describe "detection of the integration sites in a time-efficient manner, quantifying clonal size using tagmentation sequencing" (DIStinct-seq). For genome-wide IS detection, DIStinct-seq uses a bead-linked Tn5 transposome, enabling library preparation within a single day with less than 500 ng DNA. We validated its quantification accuracy using clones with known single IS. We then analyzed genome-wide integration patterns of lentiviral-transduced CAR-T cells, demonstrating the features of lentiviral integration. Moreover, we applied our approach to CAR-T cells derived from *in vivo* murine models, providing insights into clonal behavior in association with IS throughout the genome.

RESULTS

On-bead tagmentation enabled straightforward library preparation for IS detection

The most used methods, such as LAM-PCR or LM-PCR, have lengthy steps, including DNA fragmentation and adapter ligation, followed by the amplification of the host-vector chimeric fragments (Figure 1A). Performing DNA fragmentation and adapter ligation separately is time consuming and induces DNA loss during sample transfer. Using tagmentation, Tag-PCR simultaneously fragments DNA in random positions while binding sequencing adapters, resulting in substantial savings in labor, time, and input DNA (Figure 1B). However, the kit used in Tag-PCR is optimized for a maximum input DNA of 50 ng as being in-solution-based, which could result in insufficient recovery for quantitative analyses. Furthermore, because of the in-solution-based reaction, it is necessary to ensure the correct enzyme-to-DNA ratio to generate the desired fragment size for sequencing. This can cause inconveniences requiring precise quantification of the input DNA during sample preparation. Critically, the tagmentation kit used in Tag-PCR (Nextera DNA Library Preparation Kit, Illumina) is no longer commercially available. Accordingly, we speculated that the bead-linked Tn5 transposome kit (Illumina DNA prep, Illumina) could be used to detect IS. In contrast with the in-solution-based kit, bead-linked Tn5 transposome normalizes library yields through on-bead tagmentation, leading to consistent recovery.[19] In addition, we speculated that a substantial amount of input DNA, ranging up to 500 ng in the bead-linked transposome, could facilitate quantitative IS analysis.

As a result, we developed DIStinct-seq, a method for detecting IS using beads-linked transposomes (Figure 3C). An overview of the library preparation step is as follows (further details are provided in the Materials and methods). First, adapter ligation and fragmentation were simultaneously performed through on-bead tagmentation for input DNA up to 500 ng, which is the maximum allowed amount of the kit. In the following steps, host-vector chimeric fragments were specifically amplified via nested PCR while attaching sequences for subsequent sequencing. In a single day, we were able to prepare

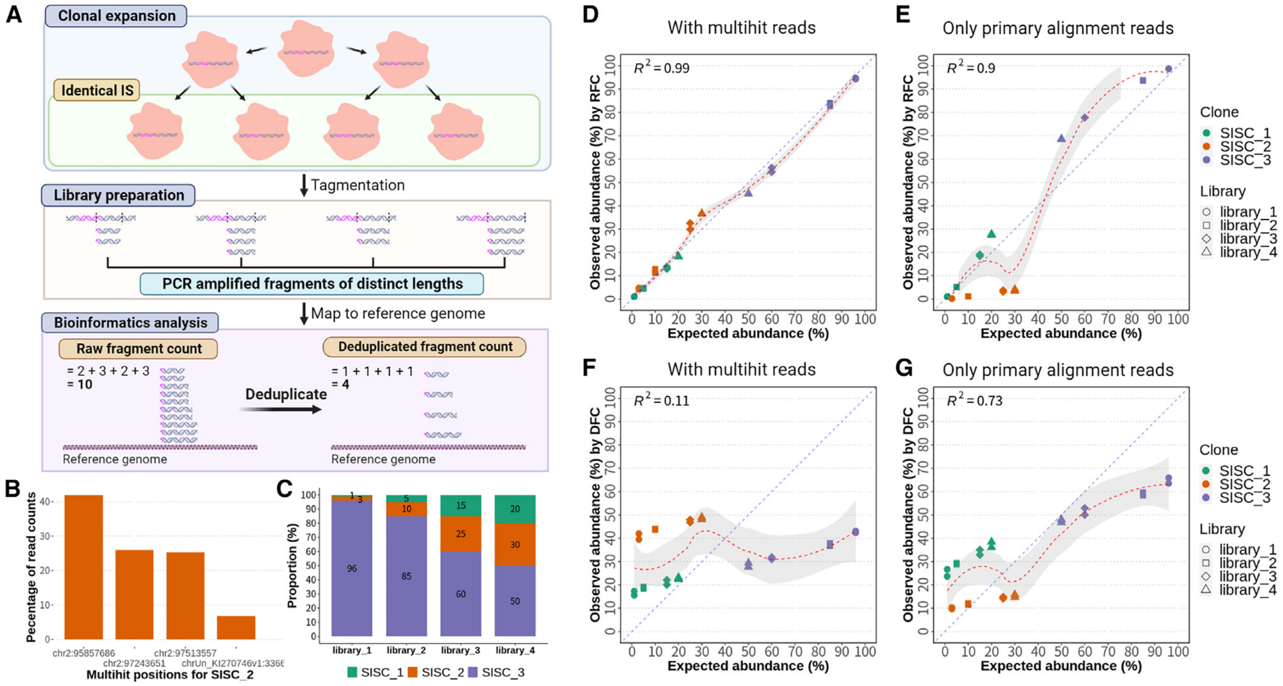


Figure 2. DIStinct-seq can quantify clonal size using ISs as molecular barcodes

(A) The theoretical explanation of quantitative analysis methods for IS. All progeny cells from the same parent cell have the same IS. In the process of preparing the library, DNA from each cell is fragmented into distinct lengths and amplified by PCR. After NGS, the number of raw fragments mapped to an identical location (RFC) can be calculated as a measure of the clone's size. It is also possible to use deduplicated fragments with distinct lengths (DFC) as a measure of clonal size after removing reads with identical IS and fragment lengths. (B) We confirmed that SISCs had unique ISs using whole-genome sequencing, although one clone (SISC_2) had multiple alignment sites due to integration into repetitive genomic regions. In SISC_2, the most frequent multiple alignment sites accounted for approximately 40%. (C) Genomic DNA extracted from three SISCs was mixed in specific proportions to reflect various concentration ranges. Using DIStinct-seq, we produced four libraries in duplicate. The numbers on the bar graph represent the percentage of DNA in each SISC. (D) When merging multiple alignment reads of SISC_2 into the single most frequent site, the observed clonal abundance by RFC was directly proportional to the expected abundance with a small bias. (E) When only primary alignments with mapping quality exceeding a certain threshold were counted, the observed clonal abundance of SISC_2 was significantly lower than predicted. (F and G) In the case of DFC, observed clonal abundance was generally proportionate to the expected value. However, the clonal abundance of SISC_2 was significantly higher or lower than expected when multiple alignment reads were merged (F) or only primary alignment reads were used (G), respectively. The red dashed lines are trend lines plotted with local polynomial regression fitting. The gray zone around the trend line represents confidence intervals.

the libraries. After that, we generated approximately 4 Gb of sequencing data for each sample.

To detect IS accurately while filtering out false positives rigorously, we developed a bioinformatics pipeline (Figure S1). As a first step, we obtained reads that contained long terminal repeat (LTR) sequences, removed the LTR portion, and mapped the reads to the human (hg38)/vector fusion reference genome. Several steps were then taken to filter out artifactual reads. First, we filtered reads with improper orientations for paired-end sequencing. Second, putative chimeric fragments produced by PCR recombination were removed. We filtered the reads when paired reads were detected on different chromosomes or if the estimated size of the fragment was greater than 2,000 base pairs (bp), which we set as the maximum allowed fragment length for the short-read sequencing platform. The IS were then identified as vector-human genome junction sites after several putative false positives had been corrected. In cases where reads aligned with multiple sites due to integration into a repetitive genome region, these ISs were merged into the single most frequent IS. Additionally, ISs from a family of reads with a few bp discrepancies due to PCR or sequencing errors, called "fuzz,"[20] were assigned to the original IS. Details of the software and methodology used in each step can be found in the Materials and methods. The code for this bioinformatics pipeline is available on GitHub (https://github.com/jaeryuk/DIStinct-seq).

DIStinct-seq successfully quantified the relative size of lentiviral-transduced clones

The IS analysis measures clonal abundance by using stable vector integration as a footprint for the clone (Figure 2A). Since all progeny cells derived from the same parent cell have the same IS, the count of the IS can be used to determine the size of the clone. Thus, the total number of DNA fragments contributing to an IS, which we call raw fragment counts (RFC) can be considered as a quantitative measure of the clone. However, it could be affected by PCR amplification bias because of the different lengths or GC contents between

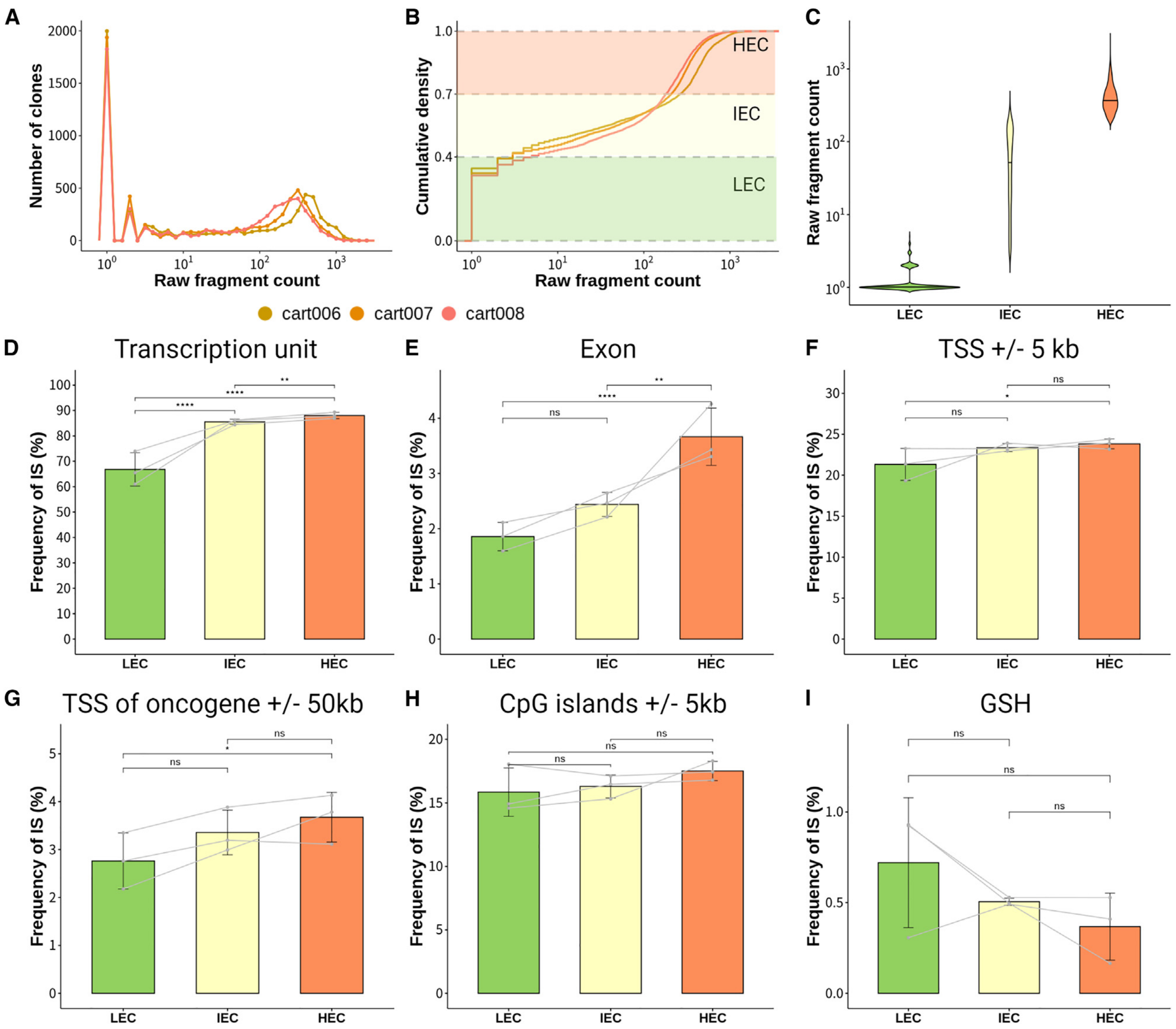


Figure 3. Lentiviral ISs are associated with *ex vivo* clonal expansion of CAR-T cells

(A) The clones with RFCs of 1 exhibited the greatest dominance when we measured the clonal size of three CAR-T products. (B) The clones were divided into three groups based on their size. To include clones with RFC of 1 which comprised nearly 40% of all clones, LEC was defined as the lower 40th percentile. IEC and HEC were defined as the 40th–70th and upper 70th percentile, respectively. (C) Using the violin plot, RFCs are presented for each clonal group, along with kernel probability densities representing the data's proportion. Black horizontal line represents the median value. (D–I) The frequency of ISs in regions that may be associated with clonal selection. In (D) transcription unit, (E) exon, (F) regions within ±5 kb of the TSS, and (G) regions within ±50 kb of the TSS of oncogenes, more expanded clones showed significantly higher frequency of IS. Vertical lines represent standard errors. Gray lines indicate clone groups in the same sample. The statistical analysis was performed using the R package, lmer4, with a generalized linear mixed-effect model fitted by maximum likelihood. ns, not significant. *p < 0.05, **p < 0.01, ***p < 0.001, and ****p < 0.0001.

fragments. Instead, it is possible to quantify IS using the count of distinct fragment lengths arising from random shearing of DNA after removing duplicated PCR fragments,[21] which we call deduplicated fragment counts (DFC).

To experimentally evaluate the quantification performance of DIStinct-seq, we used samples with known IS mixed in certain proportions. As a first step, we transduced lentiviral vectors into HEK293FT cells and generated clones that were expanded from a single cell by the serial dilution method, which we refer to as single IS clones (SISCs) (Figure S2). We then performed whole genome sequencing for three SISCs and mapped the reads to the human-vector fusion reference genome. Based on paired reads mapped to both human and vector genomes simultaneously, we were able to identify IS. We confirmed that all clones had a single IS by inspecting the binary alignment map (BAM) file; however, we observed that SISC_2 displayed multiple alignment sites within repetitive regions caused by ambiguous mapping (Figure 2B). Then we generated four libraries in duplicate with mixtures of DNA from SISCs (Figure 2C). After that, the size of the clone was quantified using two approaches: RFC and DFC. In RFC, when multiple alignment reads were merged into a single IS, the observed clonal abundance was directly proportional to the expected abundance with small biases (Figure 2D). To estimate biases in RFC introduced by differences in PCR amplification efficiency between model clones, we calculated IS detection efficiencies using observed read counts for each clone. They varied among clones within a sample, with ratios between clones ranging from 18% to 60% (Table S2). Meanwhile, when only primary alignment reads were considered after removing multiple alignment reads with low mapping quality, SISC_2 showed substantially less clonal abundance (Figure 2E). In the case of DFC, the observed clonal abundance tended to be proportionate to the expected abundance (Figures 2F and 2G). However, the abundance of the SISC_2 clone was erroneously high or low depending on whether multiple alignment reads were considered or not. In DFC, an overestimation of clonal abundance may result from separately counting fragment lengths from multiple alignment sites, which are saturated independently. In addition, when we investigated the correlation of absolute fragment counts with expected abundance (Figure S3), DFC reached saturation approximately above 300. This represented a limitation of DFC that could underestimate the size of enriched clones because of the limited diversity of fragment lengths on short-read sequencing platforms. Accordingly, although small biases may arise because of inconsistent PCR efficiencies, we used the relative clonal abundance based on RFC incorporating multiple alignment reads into a single IS to compare clonal abundance in subsequent analyses.

DIStinct-seq revealed characteristic features of genome-wide lentiviral ISs

We applied DIStinct-seq to analyze the distribution of lentiviral vector IS in three CAR-T cell products (cart006, cart007, and cart008) that we produced from the blood of three independent healthy donors (see the materials and methods for further details). We detected a total of 17,695 IS (5,786, 6,008, and 5,859, respectively) from three CAR-T products using 500 ng DNA per sample (Table S1). For comparison, we created *in silico* data by generating 5,000 random IS across the genome in 1,000 iterations, resulting in 5 million IS in total. To verify the accuracy of the detection, we compared the DNA motifs surrounding the IS, which we revealed for CAR-T products with a known lentiviral integration motif (Figure S4). The DNA motif identified by our analysis coincided with the DNA motif identified by LM-PCR for 13,442 IS of HIV-1[22] (Figure S4D).

To gain insight into the characteristics of lentiviral integration, we examined the distribution of IS across diverse genomic regions (Figure S5). We found that IS was distributed across all chromosomes and tended to be proportional to the frequency of random IS that reflects the size of each chromosome (Figure S5A). However, the frequency of integration on some chromosomes was proportional to the number of genes per chromosome rather than its size. This indicates that lentiviral integration was more likely to occur in gene-rich regions than in gene-poor regions. Furthermore, we examined the distribution of IS across genomic regions related to the expression of nearby genes, which may affect clonal behavior (Figures S5B–S5F). We also investigated the distribution of IS on genomic safe harbors (GSH), sites that support stable and efficient transgene expression without detrimentally altering cellular functions (Figure S5G). The frequency of IS on CAR-T products was higher than that of random IS in the transcription unit, exons, regions within 5 kb of the transcription start sites (TSS), regions within 50 kb of oncogene TSS, and regions within 5 kb of CpG islands. In contrast, in GSH, the frequency of IS was lower than that of random IS. To further investigate the distribution of IS around the TSS in high resolution, we analyzed the frequency of IS at 500-bp intervals (Figure S5H). Overall, IS were enriched around the TSS but depleted within approximately 1 kb. Our results were in line with those of the previous analysis,[12,23–25] thus demonstrating the robustness of our approach.

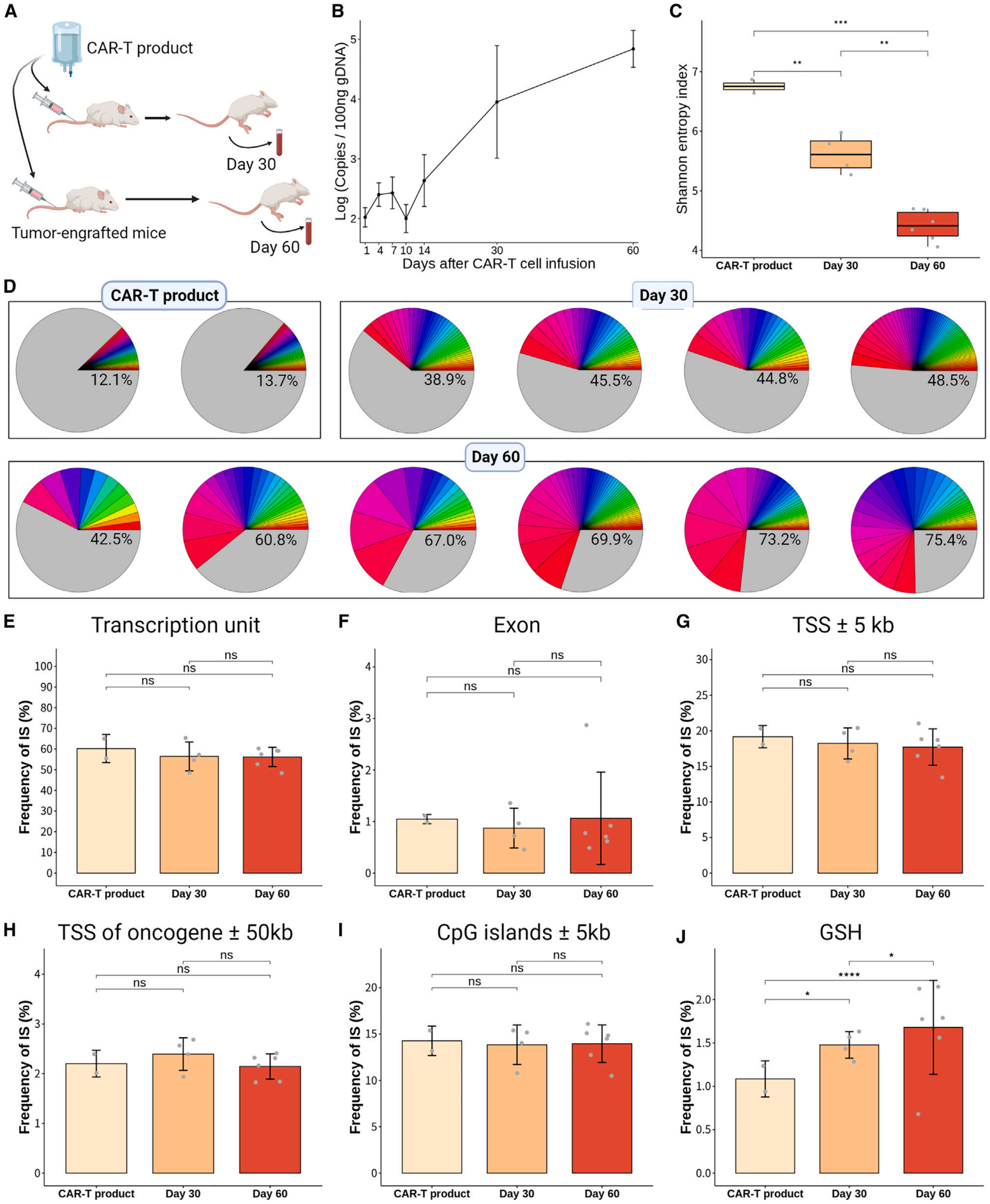
The clonal size of CAR-T products was relevant to lentiviral ISs

As clonal selection may occur through integration into genomic regions where cellular function may be affected,[17] we hypothesized that the frequency of IS in these regions would differ based on clonal size. As a first step, we measured the clonal size of three CAR-T products and found that clones with RFC of 1 were the most dominant (Figure 3A). We then grouped the clones of each sample based on their relative clonal abundance measured with RFC. For the inclusion of clones with RFC of 1, which comprise nearly 40% of all clones, less expanded clones (LEC) were defined as the lower 40th percentile for each sample (Figures 3B and 3C). Intermediately expanded clones (IEC) and highly expanded clones (HEC) were defined as the 40th–70th and upper 70th percentiles, respectively. We found that the frequency of IS varies significantly depending on clonal size in various genomic regions (Figures 3D–3I). An increased frequency of IS was observed in more expanded clones in the transcription unit, exon, regions within 5 kb of the TSS, regions within 5 kb of the oncogene TSS, and regions within 5 kb of CpG islands. By contrast, in GSH, the frequency of IS was negatively correlated with clonal size, although not statistically significant. Overall, these data demonstrate that the integration of the lentiviral vector into specific genomic regions results in clonal selection in *ex vivo* cultured CAR-T cells.

We then investigated whether clonal expansion was associated with specific pathways disrupted by IS (Figure S6). The functional enrichment analysis of genes integrated with vectors was conducted separately for LEC and merged IEC/HEC. Our analysis revealed enriched pathways including cellular metabolism, RNA processing, T cell receptor signaling, and HIV-1 infection. However, we could not find significant differences between the two clone groups.

Lentiviral integration into GSH was associated with clonal persistence in CAR-T cells *in vivo*

Next, to examine the relationship between IS and the clonal behavior of *in vivo* CAR-T cells, we used immunocompromised mouse models



(legend on next page)

that were generated for a preclinical study. Mice engrafted with acute lymphoblastic leukemia cell lines were divided into 6 groups of 10 mice per group according to the day of sacrifice (days 1, 4, 6, 14, 30, and 60). They were sacrificed to collect blood samples on the specified date after the infusion of CAR-T products targeting CD19 (Figure 4A). Using quantitative PCR targeting the vector-specific sequence, we observed the expansion of CAR-T cells *in vivo*, despite a temporary contraction on day 10 (Figure 4B). Up to day 30, the CAR-T cell concentration was generally inversely correlated with tumor size measured by the *in vivo* imaging system. Specifically, we observed a decrease until day 7, followed by an increase until day 14, and then a subsequent decrease until day 30 with a final increase at day 60 (data not shown). Tumor size was significantly decreased in all mice infused with CAR-T cells compared with control mice (data not shown). Most mice (7 of 10) infused with CAR-T cells survived until day 60, whereas all mice infused with mock T cells developed clinical symptoms including paraplegia and weight loss by day 30 and were, therefore, euthanized according to internal Institutional Animal Care and Use Committee (IACUC) euthanasia guidelines.

To compare relative clonal abundance following CAR-T cell infusion, we applied DIStinct-seq to blood collected at different time points. Because of the allocation of blood for other pre-clinical study tests, we were only able to collect 100 ng DNA from *in vivo* samples. We also measured the relative clonal abundance of infused CAR-T products using the same amount of DNA. As a result, we detected 4,055–4,473 IS for CAR-T products, 2,650–6,233 IS for day 30, and 1,034–4,895 IS for day 60 (Table S1). As a means of investigating changes in clonal diversity over time, we calculated the Shannon entropy index (Figure 4C). The Shannon entropy index indicated that CAR-T products had the greatest diversity compared with *in vivo* samples. Additionally, the *in vivo* samples displayed a decrease in diversity from day 30 to day 60. Furthermore, when we evaluated the proportion of clones in the top 1 percentile by size (Figure 4D), we found that the CAR-T product had the lowest proportion (11.3%–13.2%), while the *in vivo* samples exhibited higher proportions on day 30 (37.8%–45.6%) and day 60 (38.4%–71.2%). Although the results were not from serial observations on the same individuals, these findings suggest a decreasing trend in clonal diversity of CAR-T cells for *in vivo* samples over time after the infusion of CAR-T products.

To better understand the *in vivo* behavior of clones originating from CAR-T products, we investigated the clones present in both *in vivo* samples and CAR-T products, and analyzed their original clone group in the latter (Table S3). Our analysis revealed that only a small proportion of clones from the CAR-T products were present in the *in vivo* samples, possibly because of sampling bias or negative selection against clones that were sufficiently abundant to be detected. Notably, LEC clones, which constituted 40% of all clones in the original CAR-T product, accounted for most of the shared clones, suggesting that the expanded clones may have undergone negative selection. However, because of the lack of serial observations on the same individuals in our experimental design, a comprehensive investigation of clonal dynamics was limited.

To explore whether clonal persistence was linked with specific pathways affected by IS, we conducted a functional enrichment analysis of genes integrated with vectors (Figure S7) separately for CAR-T product, day 30, and day 60. Our analysis showed no significant differences between the three groups in Gene Ontology: biological process. However, we observed distinct pathways enriched for each group in WikiPathways. It is possible that there might be an association between the disruption of certain pathways by vector integration and clonal persistence; however, further validation through functional studies to confirm this association was not performed.

Next, we examined the frequency of IS in different genomic regions for samples collected at different time points (Figures 4E–4J). There were no statistically significant differences in the frequency of IS between each time point for most genomic regions we investigated. Only in GSH did we observe a significant increase in the frequency of IS over time (Figure 4J). Collectively, these results imply that while clonal expansion occurs as a physiological response to the tumor, lentiviral integration, especially into GSH, could lead to clonal persistence *in vivo*.

Lentiviral IS were associated with clonal expansion *in vivo* as well as *ex vivo*

We wondered whether clonal expansion was related to IS *in vivo* as well as *ex vivo*. To begin with, we divided the clones into three groups according to their clonal size as determined by RFC. As clones with RFC of 1 accounted for up to 70% of all clones in some samples (Figure 5A), we grouped the lower 70th percentile as LEC, the 70th–85th percentile as IEC, and the upper 85th percentile as HEC (Figures 5B and 5C). The pattern of integration frequency by clonal size was similar between *ex vivo* and *in vivo* samples (Figures 5D–5I). The

Figure 4. CAR-T cells expand polyclonally with decreased diversity *in vivo*, and their persistence is associated with integration into GSH

(A) The CAR-T product was infused into tumor-engrafted mice and blood was collected after 1, 4, 7, 10, 14, 30, and 60 days of infusion. Having limited samples, we could only analyze ISs at 30 and 60 days post-infusion. (B) qPCR targeting vector-specific sequences in blood from 7–10 mice at different time points after CAR-T product infusion demonstrated an increase in the concentration of cells harboring lentiviral vectors. Vertical line represent standard errors. (C) Shannon entropy index showed that clonal diversity decreased with time. Boxes indicate the median and the quartiles, while vertical lines indicate the value of multiplying 1.5 by the interquantile range. (D) As indicated by the colors of the rainbow scale, the proportion of top 1 percentile clones by size increased over time with decreasing diversity. (E–J) The frequency of IS over time in regions that may be associated with clonal selection. (J) In GSH, IS frequency was significantly higher in clones that were more expanded, suggesting that clones with IS in this region are more likely to persist. Each dot represents each sample (2 for CAR-T products, 4 for day 30, 6 for day 60). Height of the bar indicates the mean of the value in each clone group. Vertical black lines represent standard errors. The statistical analysis was performed using the R package, lmer4, with a generalized linear mixed-effect model fitted by maximum likelihood. ns, not significant. *p < 0.05, **p < 0.01, ***p < 0.001, and ****p < 0.0001.

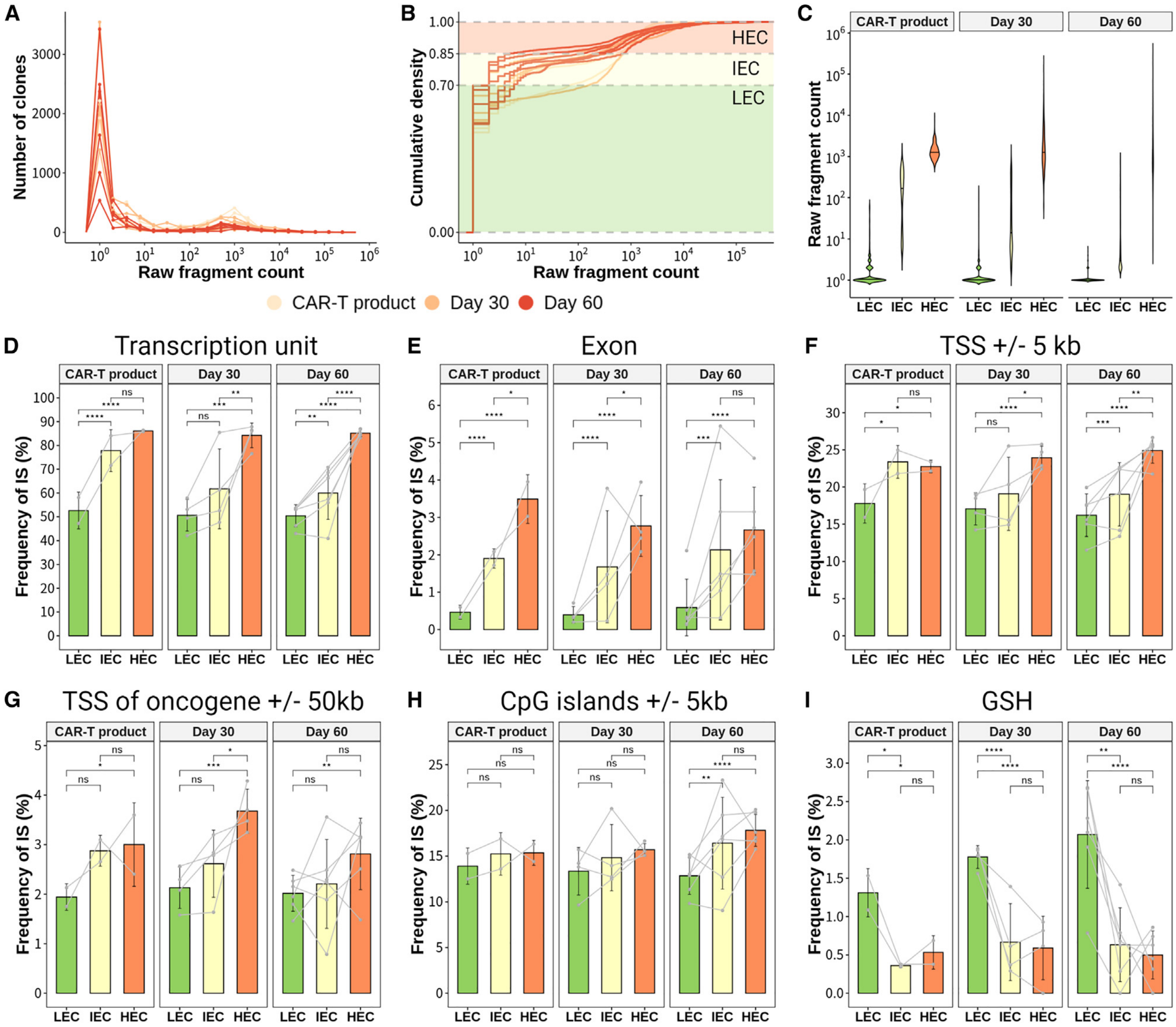


Figure 5. Lentiviral ISs are associated with clonal expansion of CAR-T cells in vivo

(A) We measured the clonal size of CAR-T products and *in vivo* samples. (A) Clones with RFCs of 1 exhibited the greatest dominance. (B) The clones in each sample were divided into three groups based on their size. To include clones with RFC of 1 which comprised nearly 70% of all clones, LEC were defined as the lower 70th percentile. IEC and HEC were defined as the 70th–85th and upper 85th percentile, respectively. (C) Using the violin plot, RFCs are presented for each clone group, along with kernel probability densities representing the data's proportion. Black horizontal line represents the median value. (D–I) The frequency of ISs in regions that may be associated with clonal selection by time point. (D) In transcription unit, (E) exon, (F) regions within ±5 kb of the TSS, and (G) regions within ±50 kb of the TSS of oncogenes, more expanded clones showed significantly higher frequency of IS in all time points. (H) In regions within ±5 kb of CpG islands, the frequency of IS was significantly higher only in samples collected on day 60. (I) In GSH, the frequency of IS was significantly lower in expanded clones across all time points. Height of the bar indicates the mean of the value in each clone group. Vertical black lines represent standard errors. Gray lines indicate clone groups in the same sample. The statistical analysis was performed using the R package, lmer4, with a generalized linear mixed-effect model fitted by maximum likelihood. ns, not significant. *p < 0.05, **p < 0.01, ***p < 0.001, and ****p < 0.0001.

frequency of IS was significantly higher in more expanded clones in the transcription unit (Figure 5D), exon (Figure 5E), regions within 5 kb of the TSS (Figure 5F), and regions within 50 kb of the oncogene TSS (Figure 5G). By contrast, in GSH, the frequency of IS was significantly higher in LEC (Figure 5I). Similar results were obtained when we grouped the clones into three excluding clones with RFC of 1 (Figure S8). These results demonstrate that lentiviral IS is associated with clonal expansion *in vivo* as well as *ex vivo*.

DISCUSSION

Quantifying clonal abundance by vector IS has been essential for assessing genotoxicity in gene therapy trials using integrating vectors.[3,4,8,26]

Furthermore, for CAR-T cell therapies, clonal tracking has helped to evaluate the persistence of immune surveillance.[13,14] Clinically, LAM-PCR or LM-PCR have been used for this purpose, but their use has been limited because of technical bias,[27] lengthy protocols, and the need for expensive instruments and relatively high DNA input. By developing DIStinct-seq, we have improved the clinical feasibility of IS analysis. As in previous studies,[11,12] tagmentation was used, but a bead-linked Tn5 transposome simplified the experimental process in addition to allowing the quantification of IS. In addition, with it, we revealed a relationship between IS and clonal behavior, which could be crucial for the development of safe and efficacious CAR-T cells.

We demonstrated that DIStinct-seq reliably quantified clonal abundance using cell lines with known IS. We quantified clonal abundance with two measures, RFC and DFC. As tagmentation by Tn5 transposase is nearly random across the genome,[28] we could apply fragment length data to measure clonal abundance in DFC. Nevertheless, because of the finite complexity of fragment lengths on the short-read sequencing platform, saturation was observed above approximately 300 of DFC (Figure S3D). This could lead to an underestimation of the clonal abundance of the dominant clone exceeding this threshold. As well, when integration occurs in repetitive genomes, DFC results in an overestimation of clonality because of separately counting fragmentation sites in multiple alignment sites (Figure 2F). To address these issues, Sherman et al.[10] constructed a pipeline based on statistical modeling with maximum likelihood estimation,[21] which was optimized for LM-PCR coupled with sonication generated with MiSeq platform. However, this method requires sophisticated statistical modeling for an accurate estimation, which has not been applied to tagmentation. Instead, we used RFC as a quantitative measure based on our results with samples with known IS. However, as RFC can be biased by PCR amplification[29,30] and is not an absolute representation of clonal size, we grouped the clones based on their relative abundance in each sample for comparison.

These limitations of DIStinct-seq may be overcome by alternative methods. A unique molecular identifier (UMI) consisting of random nucleotides would provide an accurate estimate of clonal abundance, eliminating PCR amplification biases.[20,31] However, there are not yet commercially available tagmentation kits with UMI, which limits their application in clinical settings. A recent IS analysis using long-read based nanopore sequencing showed a potential for precise quantification of clonal size with fragmentation length data.[32] However, the need for large amount of DNA, up to 10 μg, restricts its use in clinical settings. Thus, despite a few limitations, DIStinct-seq has the potential to be a valuable tool in clinical settings, given its simplicity, speed, and availability with the commercial kit at a relatively inexpensive cost ($US43 per sample). Meanwhile, we found that the amount of input DNA was related to recovery of the number of IS, that is, approximately 4,000 IS with 100 ng and approximately 6,000 IS with 500 ng in the cart006 sample (Table S1). Regarding this, the maximum DNA input, 500 ng, may limit its use in highly polyclonal samples. This could be alleviated by fully representing the given amount of DNA in the library preparation step. For instance, minimizing DNA loss during library preparation will be possible by using the full volume of the first PCR product for subsequent nested PCR.

We found characteristics of lentiviral integration that were consistent with prior findings.[12,23–25] Additionally, we discovered that lentiviral IS may contribute to clonal expansion. A variety of mechanisms can be involved in this phenomenon.[33] One such mechanism is the activation of oncogenes by the internal enhancer or promoter of lentiviral vectors integrated nearby.[34] In addition, several other potential mechanisms have been revealed. Two such mechanisms are i) as in the *TET2* gene, inactivation of the catalytic domain of tumor suppressor genes by insertional disruption[15] and ii) as in the *HMG2A* gene, activation via mRNA 3′ end substitution that suppresses RNA degradation.[35] In this study, we focused on the certain genomic regions rather than examining the insertional mutagenesis mechanisms for individual genes in detail. This approach provided a comprehensive insight into the relationship between IS and clonal behavior from a genome-wide perspective.

Meanwhile, the enriched pathways for genes containing the vector did not show significant differences between clone size groups in CAR-T products. Notably, these pathways included T cell receptor signaling and HIV-1 infection, irrespective of clone size. Therefore, considering that lentivirus preferentially integrates into actively transcribed genes,[24] we speculate that highly expressed genes during transduction may have compromised the enrichment of genes related to clonal expansion, We observed polyclonal expansion of CAR-T cells with decreasing diversity *in vivo*. Since tumor size decreased after the infusion of CAR-T cells, the main mechanism driving CAR-T cell expansion seems to be a physiological response to tumors. However, during this process, our results indicate that vector integration into certain genomic regions also contributed to clonal behavior. Especially in GSH, surviving clones had a higher frequency of IS. In combination with the finding that IS frequency in GSH was inversely correlated with clonal expansion, this suggests that GSH could be an ideal target for transgene insertion for CAR-T cell therapy. Targeting GSH may enhance long-term efficacy while decreasing the risk of malignant transformation by preventing abnormal expansion. Indeed, the insertion of transgenes into GSH regions using CRISPR-Cas9 for CAR-T cell therapy has shown promising results demonstrating sustained expression of the transgene with high anti-leukemic efficacy.[36,37] Further research will be required to identify additional specific loci within GSHs that can be optimized to balance the safety and efficacy of CAR-T cells.

To conclude, we have developed DIStinct-seq, which uses on-bead tagmentation that offers several practical advantages over existing protocols. Straightforward library preparation with relatively low DNA input makes it feasible for clinical applications. Even though we developed DIStinct-seq based on lentiviral vectors, it can also be adapted to any type of integrating vector by modifying the vector-specific primers. Furthermore, by using it, we gained a better understanding of the clonal behavior of CAR-T cells in association with vector integration into certain genomic regions, such as GSH. Along with the methodological advancement of IS analysis, we anticipate that these findings will contribute to enhanced safety and efficacy of gene therapy.

MATERIALS AND METHODS

Starting blood products for manufacturing CAR-T cells

Clinical leukapheresis products were obtained from healthy volunteers (n = 3). Mononuclear peripheral blood cell apheresis products were processed without cryopreservation within 24 h after receipt. The institutional review board of Seoul National University Hospital approved the study for human research (SNUH-IRB, H-1606-033-768).

Anti-CD19 CAR-T cells using lentiviral vector transduction

All experiments described in this article used a new CD19 CAR vector, LTG1563, developed and provided by Lentigen, Miltenyi Biotec. The vector contains single-chain variable fragment FMC63-based targeting domain, CD8-derived hinge region, TNFRSF19-derived transmembrane region, 4-1BB/CD137 costimulatory domain, and CD3-zeta chain intracellular signaling domain. According to the manufacturer's process, lentiviral vectors expressing the anti-CD19 CAR transgene were produced using a four-plasmid packaging system (third generation).

Selected $CD4^+$ and $CD8^+$ human primary T cells from healthy donors were cultured in TexMACS medium supplemented with 3% human AB serum (Life Science Production) in the presence of IL-7 (12.5 ng/mL) and IL-15 (12.5 ng/mL) and activated with CD3/CD28 MACS GMP TransAct reagent (Miltenyi Biotec). On day 1, activated T cells were transduced with lentiviral vectors encoding CAR constructs, and media was changed on day 3. On day 6, cultures were transferred to TexMACS medium (serum free) supplemented with 12.5 ng/mL of IL-7 and IL-15 each and propagated until harvest on day 12. These processes were performed on an automated CliniMACS Prodigy production equipment (Miltenyi Biotec).

Single cell-derived clones transduced with lentiviral vector

We produced single clones with a unique IS for validating quantification performance. To produce a cell with a single IS, the CAR-harboring vector had to be further modified to include selection markers such as fluorescence proteins. In addition, since our method relies on the LTR regions present in all lentiviral vectors, whether the CAR is incorporated or not would not affect the quantification measure. Thus, we used empty lentiviral vector harboring selection marker, EmGFP (Addgene #113884), instead of the CAR vector. First, we developed a stable HEK293FT cell line transduced with a lentiviral vector expressing EmGFP at an MOI of 0.4, as previously described.[38] Cells expressing EmGFP were sorted using FACS (BD FACSAria III Cell Sorter) to purify transduced cells. EmGFP-harboring cells were cultured in a 96-well plate after 10-fold serial dilutions. We selected three wells containing a single cell-derived clone by observing fluorescence signals under a microscope. The single cell-derived clones were passaged to cover a 100-mm plate. Subsequently, DNA was extracted and the whole genome was sequenced to a depth of 30×, confirming that each clone had a unique IS, although one of the clones, SISC_2, had reads with multiple alignment sites.

Library preparation for NGS

For tagmentation, we used Illumina DNA prep (Illumina) with up to 500 ng DNA. We performed tagmentation and post-tagmentation steps according to the manufacturer's instructions, followed by nested PCR. First, DNA was treated with bead-bound Tn5 transposome that fragmented DNA and attached adapter sequences simultaneously at both ends. We added a tagmentation stop buffer to stop the reaction and washed the beads. To enrich the vector-host chimeric fragment, we performed first-round PCR, using a forward primer complementary to the 3' LTR sequence and a reverse primer complementary to the adapter sequence with PrimSTAR GXL DNA polymerase (Takara Bio). We performed a second-round semi-nested PCR with the products of the first-round PCR to increase specificity and attach the required sequences for NGS. The forward primer was complementary to the inner 3′ LTR and extended with P5, index, and Rd1 SP sequences. The reverse primer was complementary to Rd2 SP sequences and was extended with index and P7 sequences. Finally, the PCR product was purified and a size of 150–1,500 bp was selected to be suitable for downstream NGS using Ampure beads. The libraries were sequenced with NovaSeq 6000 at Theragen Bio (Seongnam-si), resulting in about 4 Gb of FASTQ for each library. Detailed primer sequences and protocols are provided in the Supplementary material.

Bioinformatics pipeline for IS detection

We designed a bioinformatics pipeline to accurately detect IS as illustrated in Figure S1. First, reads containing vector-genome junctions were extracted with SeqKit[39] (version 0.14.0) from adapter-trimmed FASTQ files. We then used Cutadapt[40] (version 1.18) to remove 3′ LTR-specific sequences from each read. The reads were then aligned with the human/vector fusion reference genome using the BWA[41] (version 0.7.17) mem option. PCR duplicates were subsequently marked via Picard[42] (version 2.24.0) Markduplicates. The reads were then filtered with SAMtools[43] (version 1.3.1) according to the following criteria to ensure the quality of the analysis: properly paired reads represented by SAM flag 0 × 2, and excluding reads aligned with the genome of the lentiviral vector. To exclude putative chimeric fragments generated from PCR recombination, paired reads found on different chromosomes or the estimated size of the fragment of more than 2,000 bp were discarded. Reads with soft-clipped bases were filtered based on the orientation of the reads. Next, we identified IS using in-house Python, eliminating potential artifacts resulting from multiple alignments or incorrect base calls. Multiple alignment reads due to ambiguous mapping were merged into a single IS using the information from the XA tags representing secondary alignments in BAM files. Last, we annotate IS to genomic regions with ANNOVAR[44] using a database that we curated from public resources. The above analysis was performed on the computing server at the Genomic Medicine Institute Research Service Center in Seoul, Korea.

The database for annotation

Using curated public databases, we annotated IS to specific genomic regions. We used RefSeq genes (GCF_000001405.39_GRCh38.p13) to define the areas that flank the TSS and to annotate the orientation of the genes. We used the Cancer Gene Census database from COSMIC[45] to identify oncogenes. The CpG island regions were determined with the UCSC genome browser CpG islands Track. We collected computationally defined GSH data[46] that satisfied the following criteria.

i) 50 kb away from known genes
ii) 300 kb away from known oncogenes
iii) 300 kb away from miRNAs; 150 kb away from long non-coding RNAs and tRNAs
iv) 300 kb away from the telomeres and centromeres
v) 20 kb away from known enhancer regions

DNA motif analysis

We used WebLogo[47] to analyze the DNA motifs around the lentiviral IS. A logo plot represents consensus sequences around IS, as well as the conservation of nucleotides at those positions represented as height.

Functional enrichment analysis

We performed a functional enrichment analysis of Gene Ontology,[48] Kyoto Encyclopedia of Genes and Genomes,[49] Reactome,[50] and WikiPathways[51] using gProfileR[52] with default options.

Mouse experiment

A total of 60 immunodeficient NOD.Cg-*Prkdc*^{scid} *Il2rg*^{tm1Wjl}/SzJ mice, aged 77 weeks, were injected with 1.0×10^5 cells/mouse of Luc-NALM-6 cells (Imanis Life Science) via the tail vein. These cells were derived from B acute lymphoblastic leukemia cells and have been engineered to express luciferase. Mice was group into 6 groups of 10 per group based on the day of sacrifice (1, 4, 7, 10, 14, 30, and 60 days after CAR-T cell injection). Three days after tumor cell inoculation, saline-suspended CD19 CAR-T cells were injected at 4.0×10^6 cells/mouse, and control groups received an equivalent volume of saline. All experiments were approved by the IACUC in Seoul National University Hospital (SNUH-IACUC, 20–0177).

For each mice group defined by the day of sacrifice, whole blood was collected from the abdominal vein at each specified date, under deep anesthesia. Buffy coat was isolated from whole blood after centrifugation and stored at −80°C until analysis. DNA was isolated from buffy coats using DNeasy Blood & Tissue Kits (QIAGEN) according to the manufacturer's instructions. DNA concentration was determined using a Nanodrop (Epoch, BioTek).

qPCR for CAR-T cell detection

qPCR was performed using DNA extracted from blood samples from mice. Primers and probes for CD19 CAR-T cells were synthesized by Bosung Scientific Co.; the sequences were FAM-ACT TGG AAC AAG AGG ACA TCG CCA-QSY for probe, AAA CTG CTG ATC TAC CAT AC for forward primer, and TCC TTG TTG ACA GAA GTA AG for reverse primer. All reactions were performed using the ViiA7 Real-Time PCR System (Applied Biosystems), and the parameters for the PCR cycles were as follows: 50°C for 2 min, 95°C for 10 min, followed by 40 cycles at 95°C for 15 s and 60°C for 1 min.

Statistics

To compare IS frequencies between clone groups, we used the R package, lmer4,[53] to build a generalized linear mixed-effect model fitted by maximum likelihood (Laplace approximation). Population sizes of clone groups was input as prior weights factoring in the differences in population sizes of each clone group. Since random effects may arise from correlation within variables, the sample was assigned as a random intercept, while the clone group was assigned as a random slope. When comparing Shannon entropy index, Student's t test was used. Statistical significance was defined as a p value of less than 0.05.

DATA AND CODE AVAILABILITY

Sequencing data (FASTQ files) are available on NCBI with BioProject ID PRJNA824541. Code to detect IS from raw sequencing data using DIStinct-seq is available at https://github.com/jaeryuk/DIStinct-seq.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.omto.2023.05.004.

ACKNOWLEDGMENTS

This research was supported by a grant from the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number: HR18C0006010021); Basic Science Research Program through the National Research Foundation of Korea (NRF), funded by the Ministry of Education (2020R1A6A1A03047972); Korea Basic Science Institute (National Research Facilities and Equipment Center) grant funded by the Ministry of Education (2021R1A6C101A445); and a grant of the MD-PhD/Medical Scientist Training Program through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea. Bukyung Cha, a laboratory technician of Genomic Medicine Institute, Medical Research Center, Seoul National University, Seoul, Republic of Korea, assisted in performing sequencing library preparation. The authors thank Miltenyi Biotec for providing lentiviral vectors and the CAR-T cell production process used in this study.

AUTHOR CONTRIBUTIONS

H.J.K. and J.-I.K. conceived the project and guided the study. M.P. and G.B. designed and carried out CAR-T cell experiments and contributed to manuscript writing. J.K. designed the study and performed library preparation, bioinformatic analysis, data interpretation, figure design, and manuscript writing. B.-C.K., J.-I.K., and E.K. conducted the animal experiments and contributed to manuscript writing.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

1. Braun, C.J., Boztug, K., Paruzynski, A., Witzel, M., Schwarzer, A., Rothe, M., Modlich, U., Beier, R., Göhring, G., Steinemann, D., et al. (2014). Gene therapy for Wiskott-Aldrich syndrome—long-term efficacy and genotoxicity. Sci. Transl. Med. *6*, 227ra33. https://doi.org/10.1126/scitranslmed.3007280.
2. Hacein-Bey-Abina, S., Von Kalle, C., Schmidt, M., McCormack, M.P., Wulffraat, N., Leboulch, P., Lim, A., Osborne, C.S., Pawliuk, R., Morillon, E., et al. (2003). LMO2-associated clonal T cell proliferation in two patients after gene therapy for SCID-X1. Science *302*, 415–419. https://doi.org/10.1126/science.1088547.
3. Magrin, E., Semeraro, M., Hebert, N., Joseph, L., Magnani, A., Chalumeau, A., Gabrion, A., Roudaut, C., Marouene, J., Lefrere, F., et al. (2022). Long-term outcomes of lentiviral gene therapy for the β-hemoglobinopathies: the HGB-205 trial. Nat. Med. *28*, 81–88. https://doi.org/10.1038/s41591-021-01650-w.
4. Magnani, A., Semeraro, M., Adam, F., Booth, C., Dupré, L., Morris, E.C., Gabrion, A., Roudaut, C., Borgel, D., Toubert, A., et al. (2022). Long-term safety and efficacy of lentiviral hematopoietic stem/progenitor cell gene therapy for Wiskott–Aldrich syndrome. Nat. Med. *28*, 71–80. https://doi.org/10.1038/s41591-021-01641-x.
5. Abdel-Mohsen, M., Richman, D., Siliciano, R.F., Nussenzweig, M.C., Howell, B.J., Martinez-Picado, J., Chomont, N., Bar, K.J., Yu, X.G., Lichterfeld, M., et al. (2020). Recommendations for measuring HIV reservoir size in cure-directed clinical trials. Nat. Med. *26*, 1339–1350. https://doi.org/10.1038/s41591-020-1022-1.
6. Bruner, K.M., Wang, Z., Simonetti, F.R., Bender, A.M., Kwon, K.J., Sengupta, S., Fray, E.J., Beg, S.A., Antar, A.A.R., Jenike, K.M., et al. (2019). A quantitative approach for measuring the reservoir of latent HIV-1 proviruses. Nature *566*, 120–125. https://doi.org/10.1038/s41586-019-0898-8.
7. Maldarelli, F., Wu, X., Su, L., Simonetti, F.R., Shao, W., Hill, S., Spindler, J., Ferris, A.L., Mellors, J.W., Kearney, M.F., et al. (2014). Specific HIV integration sites are linked to clonal expansion and persistence of infected cells. Science *345*, 179–183. https://doi.org/10.1126/science.1254194.
8. Biasco, L. (2017). Integration site analysis in gene therapy patients: expectations and reality. Hum. Gene Ther. *28*, 1122–1129. https://doi.org/10.1089/hum.2017.183.
9. Schmidt, M., Schwarzwaelder, K., Bartholomae, C., Zaoui, K., Ball, C., Pilz, I., Braun, S., Glimm, H., and von Kalle, C. (2007). High-resolution insertion-site analysis by linear amplification-mediated PCR (LAM-PCR). Nat. Methods *4*, 1051–1057. https://doi.org/10.1038/nmeth1103.
10. Sherman, E., Nobles, C., Berry, C.C., Six, E., Wu, Y., Dryga, A., Malani, N., Male, F., Reddy, S., Bailey, A., et al. (2017). INSPIIRED: a pipeline for quantitative analysis of sites of new DNA integration in cellular genomes. Mol. Ther. Methods Clin. Dev. *4*, 39–49. https://doi.org/10.1016/j.omtm.2016.11.002.
11. Brady, T., Roth, S.L., Malani, N., Wang, G.P., Berry, C.C., Leboulch, P., Hacein-Bey-Abina, S., Cavazzana-Calvo, M., Papapetrou, E.P., Sadelain, M., et al. (2011). A method to sequence and quantify DNA integration for monitoring outcome in gene therapy. Nucleic Acids Res. *39*, e72. https://doi.org/10.1093/nar/gkr140.
12. Hamada, M., Nishio, N., Okuno, Y., Suzuki, S., Kawashima, N., Muramatsu, H., Tsubota, S., Wilson, M.H., Morita, D., Kataoka, S., et al. (2018). Integration mapping of piggyBac-Mediated CD19 chimeric antigen receptor T cells analyzed by novel tagmentation-assisted PCR. Ebiomedicine *34*, 18–26. https://doi.org/10.1016/j.ebiom.2018.07.008.
13. Melenhorst, J.J., Chen, G.M., Wang, M., Porter, D.L., Chen, C., Collins, M.A., Gao, P., Bandyopadhyay, S., Sun, H., Zhao, Z., et al. (2022). Decade-long leukaemia remissions with persistence of CD4+ CAR T cells. Nature *602*, 503–509. https://doi.org/10.1038/s41586-021-04390-6.
14. Biasco, L., Izotova, N., Rivat, C., Ghorashian, S., Richardson, R., Guvenel, A., Hough, R., Wynn, R., Popova, B., Lopes, A., et al. (2021). Clonal expansion of T memory stem cells determines early anti-leukemic responses and long-term CAR T cell persistence in patients. Nat. Cancer *2*, 629–642. https://doi.org/10.1038/s43018-021-00207-7.
15. Fraietta, J.A., Nobles, C.L., Sammons, M.A., Lundh, S., Carty, S.A., Reich, T.J., Cogdill, A.P., Morrissette, J.J.D., DeNizio, J.E., Reddy, S., et al. (2018). Disruption of TET2 promotes the therapeutic efficacy of CD19-targeted T cells. Nature *558*, 307–312. https://doi.org/10.1038/s41586-018-0178-z.
16. Dimitri, A., Herbst, F., and Fraietta, J.A. (2022). Engineering the next-generation of CAR T-cells with CRISPR-Cas9 gene editing. Mol. Cancer *21*, 78. https://doi.org/10.1186/s12943-022-01559-z.
17. Shah, N.N., Qin, H., Yates, B., Su, L., Shalabi, H., Raffeld, M., Ahlman, M.A., Stetler-Stevenson, M., Yuan, C., Guo, S., et al. (2019). Clonal expansion of CAR T cells harboring lentivector integration in the CBL gene following anti-CD22 CAR T-cell therapy. Blood Adv. *3*, 2317–2322. https://doi.org/10.1182/bloodadvances.2019000219.
18. Nobles, C.L., Sherrill-Mix, S., Everett, J.K., Reddy, S., Fraietta, J.A., Porter, D.L., Frey, N., Gill, S.I., Grupp, S.A., Maude, S.L., et al. (2020). CD19-targeting CAR T cell immunotherapy outcomes correlate with genomic modification by vector integration. J. Clin. Invest. *130*, 673–685. https://doi.org/10.1172/Jci130144.
19. Bruinsma, S., Burgess, J., Schlingman, D., Czyz, A., Morrell, N., Ballenger, C., Meinholz, H., Brady, L., Khanna, A., Freeberg, L., et al. (2018). Bead-linked transposomes enable a normalization-free workflow for NGS library preparation. BMC Genomics *19*, 722. https://doi.org/10.1186/s12864-018-5096-9.
20. Wells, D.W., Guo, S., Shao, W., Bale, M.J., Coffin, J.M., Hughes, S.H., and Wu, X. (2020). An analytical pipeline for identifying and mapping the integration sites of HIV and other retroviruses. BMC Genomics *21*, 216. https://doi.org/10.1186/s12864-020-6647-4.
21. Berry, C.C., Gillet, N.A., Melamed, A., Gormley, N., Bangham, C.R.M., and Bushman, F.D. (2012). Estimating abundances of retroviral insertion sites from DNA fragment length data. Bioinformatics *28*, 755–762. https://doi.org/10.1093/bioinformatics/bts004.
22. Kirk, P.D.W., Huvet, M., Melamed, A., Maertens, G.N., and Bangham, C.R.M. (2016). Retroviruses integrate into a shared, non-palindromic DNA motif. Nat. Microbiol. *2*, 16212. https://doi.org/10.1038/nmicrobiol.2016.212.
23. Biasco, L., Ambrosi, A., Pellin, D., Bartholomae, C., Brigida, I., Roncarolo, M.G., Di Serio, C., von Kalle, C., Schmidt, M., and Aiuti, A. (2011). Integration profile of retroviral vector in gene therapy treated patients is cell-specific according to gene expression and chromatin conformation of target cell. EMBO Mol. Med. *3*, 89–101. https://doi.org/10.1002/emmm.201000108.
24. Milone, M.C., and O'Doherty, U. (2018). Clinical use of lentiviral vectors. Leukemia *32*, 1529–1541. https://doi.org/10.1038/s41375-018-0106-0.
25. Gogol-Döring, A., Ammar, I., Gupta, S., Bunse, M., Miskey, C., Chen, W., Uckert, W., Schulz, T.F., Izsvák, Z., and Ivics, Z. (2016). 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. Mol. Ther. *24*, 592–606. https://doi.org/10.1038/mt.2016.11.
26. Biasco, L., Baricordi, C., and Aiuti, A. (2012). Retroviral integrations in gene therapy trials. Mol. Ther. *20*, 709–716. https://doi.org/10.1038/mt.2011.289.
27. Giordano, F.A., Appelt, J.-U., Link, B., Gerdes, S., Lehrer, C., Scholz, S., Paruzynski, A., Roeder, I., Wenz, F., Glimm, H., et al. (2015). High-throughput monitoring of integration site clonality in preclinical and clinical gene therapy studies. Mol. Ther. Methods Clin. Dev. *2*, 14061. https://doi.org/10.1038/mtm.2014.61.
28. Shevchenko, Y., Bouffard, G.G., Butterfield, Y.S.N., Blakesley, R.W., Hartley, J.L., Young, A.C., Marra, M.A., Jones, S.J.M., Touchman, J.W., and Green, E.D. (2002). Systematic sequencing of cDNA clones using the transposon Tn5. Nucleic Acids Res. *30*, 2469–2477. https://doi.org/10.1093/nar/30.11.2469.
29. Gabriel, R., Eckenberg, R., Paruzynski, A., Bartholomae, C.C., Nowrouzi, A., Arens, A., Howe, S.J., Recchia, A., Cattoglio, C., Wang, W., et al. (2009). Comprehensive genomic access to vector integration in clinical gene therapy. Nat. Med. *15*, 1431–1436. https://doi.org/10.1038/nm.2057.
30. Wang, G.P., Garrigue, A., Ciuffi, A., Ronen, K., Leipzig, J., Berry, C., Lagresle-Peyrou, C., Benjelloun, F., Hacein-Bey-Abina, S., Fischer, A., et al. (2008). DNA bar coding and pyrosequencing to analyze adverse events in therapeutic gene transfer. Nucleic Acids Res. *36*, e49. https://doi.org/10.1038/nm.2057.10.1093/nar/gkn125.
31. Dawes, J.C., Webster, P., Iadarola, B., Garcia-Diaz, C., Dore, M., Bolt, B.J., Dewchand, H., Dharmalingam, G., McLatchie, A.P., Kaczor, J., et al. (2020). LUMI-PCR: an Illumina platform ligation-mediated PCR protocol for integration site cloning, provides molecular quantitation of integration sites. Mob. DNA *11*, 7. https://doi.org/10.1186/s13100-020-0201-4.

32. van Haasteren, J., Munis, A.M., Gill, D.R., and Hyde, S.C. (2021). Genome-wide integration site detection using Cas9 enriched amplification-free long-range sequencing. Nucleic Acids Res. *49*, e16. https://doi.org/10.1093/nar/gkaa1152.

33. Bushman, F.D. (2020). Retroviral insertional mutagenesis in humans: evidence for four genetic mechanisms promoting expansion of cell clones. Mol. Ther. *28*, 352–356. https://doi.org/10.1016/j.ymthe.2019.12.009.

34. Cesana, D., Ranzani, M., Volpin, M., Bartholomae, C., Duros, C., Artus, A., Merella, S., Benedicenti, F., Sergi Sergi, L., Sanvito, F., et al. (2014). Uncovering and dissecting the genotoxicity of self-inactivating lentiviral vectors in vivo. Mol. Ther. *22*, 774–785. https://doi.org/10.1038/mt.2014.3.

35. Cavazzana-Calvo, M., Payen, E., Negre, O., Wang, G., Hehir, K., Fusil, F., Down, J., Denaro, M., Brady, T., Westerman, K., et al. (2010). Transfusion independence and HMGA2 activation after gene therapy of human beta-thalassaemia. Nature *467*, 318–322. https://doi.org/10.1038/nature09328.

36. Kararoudi, M.N., Likhite, S., Elmas, E., Yamamoto, K., Schwartz, M., Sorathia, K., de Souza Fernandes Pereira, M., Devin, R.D., Lyberger, J.M., Behbehani, G.K., et al. (2021). CRISPR-targeted CAR gene insertion using Cas9/RNP and AAV6 enhances anti-AML activity of primary NK cells. Preprint at bioRxiv. https://doi.org/10.1101/2021.03.17.435886.

37. Odak, A., Yuan, H., Feucht, J., Mansilla - Soto, J., Eyquem, J., Leslie, C., and Sadelain, M. (2020). Targeted integration of a CAR at a novel genomic safe harbor directs potent therapeutic outcomes. Blood *136*, 28. https://doi.org/10.1182/blood-2020-141967.

38. Elegheert, J., Behiels, E., Bishop, B., Scott, S., Woolley, R.E., Griffiths, S.C., Byrne, E.F.X., Chang, V.T., Stuart, D.I., Jones, E.Y., et al. (2018). Lentiviral transduction of mammalian cells for fast, scalable and high-level production of soluble and membrane proteins. Nat. Protoc. *13*, 2991–3017. https://doi.org/10.1038/s41596-018-0075-9.

39. Shen, W., Le, S., Li, Y., and Hu, F. (2016). SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. PLoS One *11*, e0163962. https://doi.org/10.1371/journal.pone.0163962.

40. Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet. J. *17*, 10–12. https://doi.org/10.14806/ej.17.1.200.

41. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760. https://doi.org/10.1093/bioinformatics/btp324.

42. Picard Toolkit. (2019). Broad Institute, GitHub repository.

43. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The sequence alignment/map format and SAMtools. Bioinformatics *25*, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352.

44. Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. *38*, e164. https://doi.org/10.1093/nar/gkq603.

45. Sondka, Z., Bamford, S., Cole, C.G., Ward, S.A., Dunham, I., and Forbes, S.A. (2018). The COSMIC cancer gene census: describing genetic dysfunction across all human cancers. Nat. Rev. Cancer *18*, 696–705. https://doi.org/10.1038/s41568-018-0060-1.

46. Aznauryan, E., Yermanos, A., Kinzina, E., Devaux, A., Kapetanovic, E., Milanova, D., Church, G.M., and Reddy, S.T. (2022). Discovery and validation of human genomic safe harbor sites for gene and cell therapies. Cell Rep. Methods *2*, 100154. https://doi.org/10.1016/j.crmeth.2021.100154.

47. Crooks, G.E., Hon, G., Chandonia, J.M., and Brenner, S.E. (2004). WebLogo: a sequence logo generator. Genome Res. *14*, 1188–1190. https://doi.org/10.1101/gr.849004.

48. Gene Ontology Consortium, Douglass, E., Good, B.M., Unni, D.R., Harris, N.L., Mungall, C.J., Basu, S., Chisholm, R.L., Dodson, R.J., Hartline, E., et al. (2021). The Gene Ontology resource: enriching a GOld mine. Nucleic Acids Res. *49*, D325–D334. https://doi.org/10.1093/nar/gkaa1113.

49. Kanehisa, M., Furumichi, M., Sato, Y., Ishiguro-Watanabe, M., and Tanabe, M. (2021). KEGG: integrating viruses and cellular organisms. Nucleic Acids Res. *49*, D545–D551. https://doi.org/10.1093/nar/gkaa970.

50. Fabregat, A., Sidiropoulos, K., Viteri, G., Forner, O., Marin-Garcia, P., Arnau, V., D'Eustachio, P., Stein, L., and Hermjakob, H. (2017). Reactome pathway analysis: a high-performance in-memory approach. BMC Bioinformatics *18*, 142. https://doi.org/10.1186/s12859-017-1559-2.

51. Slenter, D.N., Kutmon, M., Hanspers, K., Riutta, A., Windsor, J., Nunes, N., Mélius, J., Cirillo, E., Coort, S.L., Digles, D., et al. (2018). WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. Nucleic Acids Res. *46*, D661–D667. https://doi.org/10.1093/nar/gkx1064.

52. Raudvere, U., Kolberg, L., Kuzmin, I., Arak, T., Adler, P., Peterson, H., and Vilo, J. (2019). g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). Nucleic Acids Res. *47*, W191–W198. https://doi.org/10.1093/nar/gkz369.

53. Bates, D.W., Zimlichman, E., Bolker, B.M., and Walker, S.C. (2015). Fitting linear mixed-effects models using lme4. J. Stat. Softw. *67*, 1–48. https://doi.org/10.18637/jss.v067.i01.