scientific reports

Check for updates

OPEN

# High spontaneous integration rates of end-modified linear DNAs upon mammalian cell transfection

Samuel Lim[1,2]✉, R. Rogers Yocum[3], Pamela A. Silver[1,2] & Jeffrey C. Way[3]✉

**In gene therapy, potential integration of therapeutic transgene into host cell genomes is a serious risk that can lead to insertional mutagenesis and tumorigenesis. Viral vectors are often used as the gene delivery vehicle, but they are prone to undergoing integration events. More recently, non-viral delivery of linear DNAs having modified geometry such as closed-end linear duplex DNA (CELiD) have shown promise as an alternative, due to prolonged transgene expression and less cytotoxicity. However, whether modified-end linear DNAs can also provide a safe, non-integrating gene transfer remains unanswered. Herein, we compare the genomic integration frequency upon transfection of cells with expression vectors in the forms of circular plasmid, unmodified linear DNA, CELiDs with thioester loops, and Streptavidin-conjugated blocked-end linear DNA. All of the forms of linear DNA resulted in a high fraction of the cells being stably transfected—between 10 and 20% of the initially transfected cells. These results indicate that blocking the ends of linear DNA is insufficient to prevent integration.**

Gene therapy aims to treat patients with genetic disease by introducing genetic material into cells to repair defective cellular functions[1,2]. Traditionally, viral vectors such as retrovirus or lentivirus have been preferred as the gene delivery vehicles due to their high transfection efficiency[3]. However, viral vector-based methods suffer from a critical safety concern, regarding potential integration of DNA into a host cell chromosome, which may result in activation of oncogenes or knockout of tumor suppressor genes[4]. Indeed, there were multiple reports of patients developing oncogenesis during the early trials of gene therapy using retroviral vector delivery[5,6]. While adeno-associated virus (AAV) vectors display less propensity to integrate, several studies suggest that the risk may still exist[7–9]. Because of such potential danger, the current scope of gene therapy is limited primarily to the treatment of cancer or rare genetic disorders, where patients and their caregivers are more likely to accept the risks.

Reflecting recent efforts to address the limits of viral vectors, there is a growing interest in non-viral delivery of nucleic acids[10]. Despite its simplicity and ease of production, circular plasmid DNAs may allow chromosomal integration through crossover events. Linear DNAs, including those derived from cleavage of circular DNAs after entering a cell, may undergo non-homologous end joining (NHEJ) to be randomly integrated into the host cell genome, while being prone to exonuclease digestion. Recently, close-ended linear duplex DNA (CELiD), which consists of double-stranded DNA molecules with covalently closed terminal hairpins, has been reported[11]. A type of CELiD is generated as an intermediate of AAV replication in eukaryotic cells, and such structures are in effect a double stranded AAV genome capped with hairpin-forming palindromic terminal regions[12]. When CELiDs and circular plasmids were injected into a mouse, both forms of DNA were trafficked into the liver, but the CELiDs promoted extended and more stable expression of the transgenes[11]. This observation highlights the promising potential of end-modified linear DNA as an alternative non-viral vector for gene therapy. However, the question of whether this strategy can provide a safe, non-integrating gene delivery remains uncertain. While modifying DNA ends is known to hinder random end-joining events[13–15], extended transgenic expression by CELiDs may result from a higher tendency for integration.

Herein, we compare the genomic integration of linear DNAs having different types of modified ends, and investigate whether blocking DNA ends improves the safety of gene delivery. Specifically, we synthesized a novel blocked-end linear DNA using streptavidin–biotin bioconjugation chemistry[16] as well as in vitro synthesized analogue of a CELiD with thioester bonds in the loop, and subsequently compared its propensity for integration to those of unmodified linear and circular DNAs. Upon transfection, we cultured the cells in the absence of a

[1]Department of Systems Biology, Harvard Medical School, Boston, MA 02115, USA. [2]Wyss Institute for Biologically Inspired Engineering, Boston, MA 02115, USA. [3]General Biologics, Inc, 108 Fayerweather Street, Unit 2, Cambridge, MA 02138, USA. ✉email: samuel_lim@hms.harvard.edu; jeff.way@genbiologics.com

drug selection over extended time, during which transiently transfected genes would get diluted away and any remaining stable gene expression could be attributed to integration events.

## Results

**Design and synthesis of the linear DNA constructs.** To investigate the effect of varying the structure of the ends of double stranded linear DNA end region geometry on chromosomal integration, we designed four DNA constructs, each encoding a green fluorescent protein (GFP) reporter and puromycin resistance. We first constructed a circular reporter plasmid and subsequently synthesized plain linear DNA and two types of modified-end constructs from this plasmid, thus minimizing any possible effect originating from differences in sequence or synthesis method. The circular plasmid contained two constitutive expression cassettes for the GFP reporter and a puromycin resistant selection marker, in addition to the WPRE element for increasing mRNA stability, all flanked by *BsaI* restriction sites designed to give different four-base overhangs at the two ends (Fig. 1A). Plain linear DNA with unmodified ends was created simply by digestion with *BsaI* followed by purification (Fig. 1B). Because *BsaI* is a type IIS enzyme that cleaves at a short distance outside of its recognition site, the two sticky ends have different and non-complimentary sequences, avoiding potential self-ligation. CELiDs were synthesized by modifying each end of this plain linear DNA by ligation with hairpin-forming oligonucleotides using a Golden Gate assembly (Fig. 1B, C; Methods).

To construct blocked-end DNAs, we first synthesized biotin end-labeled DNAs using the same *BsaI*-digested linear DNA as for the CELiDs, using two different biotin-labeled duplex oligonucleotide fragments containing specific sticky ends corresponding to each end of the linear DNA (Fig. 1C, 2A). Two different versions were synthesized. In one version biotin was incorporated only at the 5' ends ("single biotin"). In the other version, biotin was incorporated in both the 3' and 5' ends ("double biotin"). We then tested three types of biotin-binding proteins, avidin, streptavidin, and neutravidin; protein-to-DNA molar ratios were varied from 0.5 to 2 in order to determine the optimal reaction condition. Binding of these proteins to the linear DNA was demonstrated in a gel-shift assay, in which DNA migration in an agarose gel is retarded upon protein conjugation (Fig. 2B, S5A). Streptavidin showed the strongest binding capacity, with all of the DNA shifting upon addition of equimolar or 2 × excess streptavidin. In addition, a band running at a much higher apparent molecular weight was observed, which was likely a circular form in which the streptavidin tetramer is bound to both ends of the DNA. Avidin and neutravidin did not bind as strongly, as inferred from the gel shift assay. Therefore we used streptavidin for constructing blocked-end linear DNA.

Next, we tested the stability of our blocked-end constructs by measuring streptavidin dissociation over time in the presence of excess free biotin. We assumed that any streptavidin dissociated from the DNA complex would be bound by free biotin, which would then prevent re-binding to the DNA complex. Streptavidin/biotin-DNA complexes using DNAs that were double-biotinylated (5' and 3' ends) or single-biotinylated (5' end only) were incubated with a 10 × molar excess of free biotin at 37 °C for various times. Agarose gel electrophoresis showed that after 18 h of incubation with excess free biotin, the majority of streptavidin was dissociated from

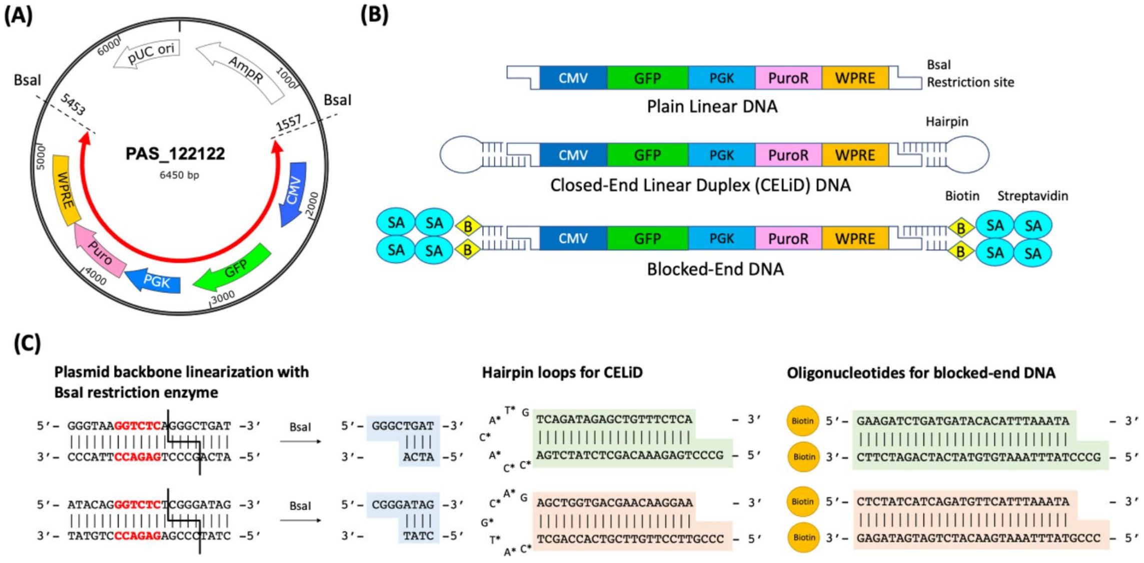
(A)
(B)
(C)
BsaI
BsaI
pUC ori
AmpR
6000
1000
5453
1557
5000
2000
3000
4000
PAS_122122
6450 bp
WPRE
Puro
PGK
GFP
CMV
CMV
GFP
PGK
PuroR
WPRE
BsaI
Restriction site
Plain Linear DNA
Hairpin
Closed-End Linear Duplex (CELiD) DNA
Biotin
Streptavidin
SA
B
Blocked-End DNA
Plasmid backbone linearization with
BsaI restriction enzyme
5'- GGGTAAGGTCTCAGGGCTGAT -3'
3'- CCCATTCCAGAGTCCCGACTA -5'
BsaI
5'- GGGCTGAT -3'
3'- ACTA -5'
5'- ATACAGGGTCTCTCGGGATAG -3'
3'- TATGTCCCAGAGAGCCCTATC -5'
BsaI
5'- CGGGATAG -3'
3'- TATC -5'
Hairpin loops for CELiD
A* T* G C* A* C* C*
TCAGATAGAGCTGTTTCTCA - 3'
AGTCTATCTCGACAAAGAGTCCCG - 5'
C* A* G G* T* A* C*
AGCTGGTGACGAACAAGGAA - 3'
TCGACCACTGCTTGTTCCTTGCCC - 5'
Oligonucleotides for blocked-end DNA
Biotin
5'- GAAGATCTGATGATACACATTTAAATA - 3'
3'- CTTCTAGACTACTATGTGTAAATTTATCCCG - 5'
5'- CTCTATCATCAGATGTTCATTTAAATA - 3'
3'- GAGATAGTAGTCTACAAGTAAATTTATGCCC - 5'

**Figure 1.** Design of DNA constructs used in this study. **(A)** Circular plasmid used as the common backbone for constructing various end-modified linear DNAs. The plasmid consisted of two constitutive expression cassettes for the GFP reporter and puromycin resistant selection marker, in addition to the WPRE element, flanked by the two *BsaI* restriction sites. The red arrow indicates a portion of the plasmid corresponding to the linear DNAs. **(B)** Structure of the linear DNAs. The end regions of the CELiD consisted of closed hairpin loop structures. The ends of blocked-end DNA contained biotin-labeled oligonucleotides, which were further non-covalently bound to streptavidin tetramers. **(C)** Detailed sequences of the sticky ends created by the plasmid backbone linearization by the *BsaI* restriction enzyme, as well as the hairpin loops and oligonucleotides complementary to each end. Starred bases indicate positions of phosphorothioate linkages on the 5' side.

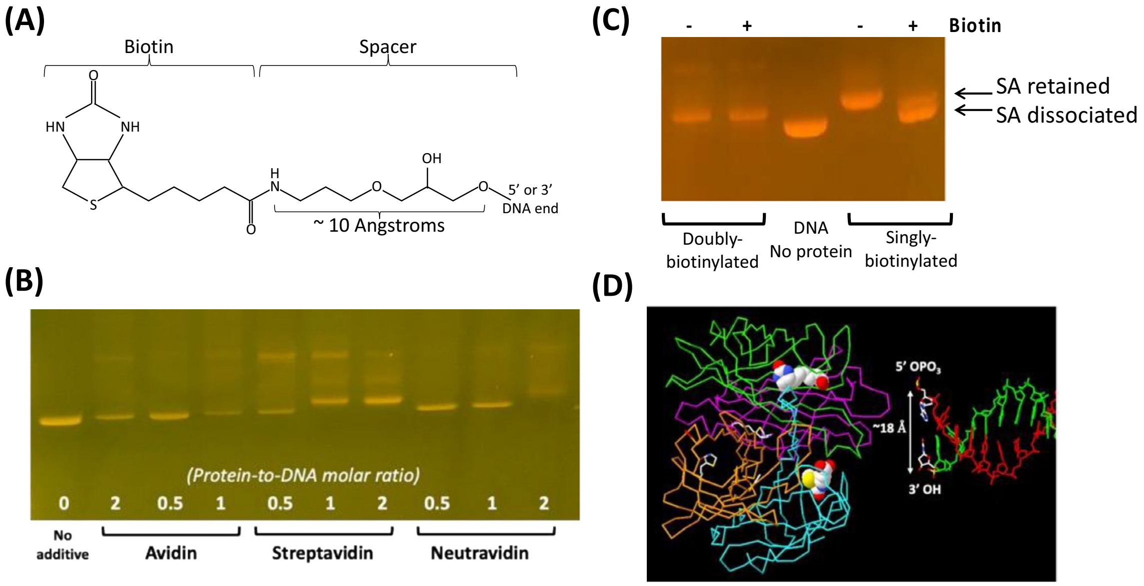

**Figure 2.** Synthesis and characterization of blocked-end linear DNA. (**A**) Structure of the 5' end of biotin-labeled linear DNA[17], (**B**) Agarose gel electrophoresis showing the formation of DNA–protein complexes using three different biotin-binding proteins: avidin, streptavidin and neutravidin. Bound complexes migrate slower compared to unbound DNA. The numbers indicate protein-to-DNA molar ratio, which varied from 0.5 to 2 for each protein. In the streptavidin-bound complexes, the DNA species that migrates slightly slower than the linear DNA alone (leftmost lane) presumably has a streptavidin tetramer bound to each end, while the most slowly migrating species may be a relaxed circle that is held together by a single streptavidin tetramer. (**C**) An agarose gel showing the dissociation of streptavidin (SA) from blocked-end DNA complexes where the DNA has either a biotin at only the 5' ends or biotins at both the 5' and 3' ends, after overnight incubation with (+) and without (−) excess free biotin. Arrows indicate bands corresponding a species with a streptavidin bound to one of the two ends ("SA retained") and a completely dissociated species, respectively. (**D**) Structural illustration of the end of a 'blocked-end linear' DNA, showing a juxtaposition of the end of a DNA fragment with the tetrameric biotin-streptavidin complex (PDB file 1STP)[18]. The 5' and 3' ends of the DNA backbone are about 18 Angstroms apart, and the distance between the carboxyl groups in the biotins is also about 18 Angstroms. Supporting Figure S5 shows the complete gel pictures.

the single-biotinylated DNA, whereas the double-biotinylated construct did not show any noticeable dissociation (Fig. 2C, S5B). Both complexes remained tightly bound in the absence of free biotin. This observation suggested that double-biotinylated linear DNA forms a significantly tighter complex with streptavidin than its single-biotinylated counterpart. This may be attributed to the favorable geometry of the double-biotinylated DNA complex, in which the distance between 5' and 3' ends (~ 18 Angstroms) is approximately similar to the distance between the carboxyl groups of adjacent biotins bound to the streptavidin tetramer according to the known crystal structure[18] (Fig. 2D). Doubly biotinylated DNA was used to construct blocked-end construct for the rest of this study.

**Comparing integration of DNA constructs upon transfection into a human cell line.** Having confirmed the formation as well as optimal design of the streptavidin-capped blocked-end linear DNA, we then transfected it into a human cell line and compared the frequency of integration to those of circular, plain linear and modified-end DNAs. We transfected the same amount of each DNA construct into human embryonic kidney (HEK) 293 cells using lipofectamine and monitored the changes in percentage of reporter-expressing cells over 3 weeks using flow cytometry (Fig. 3A). Cells were split 1:10 every three days, which essentially maintained the cells at a constant density because the doubling time of HEK293 cells is slightly less than 1 day. As the transfected cells were maintained without an antibiotic selection, the percentage of GFP-expressing cells was expected to decrease (Fig. 3B). Cell division would initially dilute the reporter gene concentration, and subsequently produce new cells that do not contain the reporter gene (Fig. 3C). Thus, after an extended period of cell growth, transiently transfected reporter expression would become negligible, and those still displaying a fluorescence signal would represent the cell population that had incorporated a chromosomal insertion of the reporter gene.

As expected, cells transfected with all four DNA constructs initially showed a decreasing GFP positive population that subsequently reached a plateau, which represented the percentage of the cells having the reporter gene integrated into their chromosome (Fig. 3B). Overall, lipofectamine-mediated transfection with plasmids and various linear DNAs resulted in 6% to 40% of the cells initially expressing GFP. Circular, supercoiled plasmid DNA gave the highest transient transfection frequency, and the relative frequencies were: plasmid > non-modified linear > thioester CELiD > blocked end (Figure S1). After the first split, the proportion of cells expressing GFP was about the same, but the strength of GFP expression was reduced (Figure S2), consistent with the idea that the

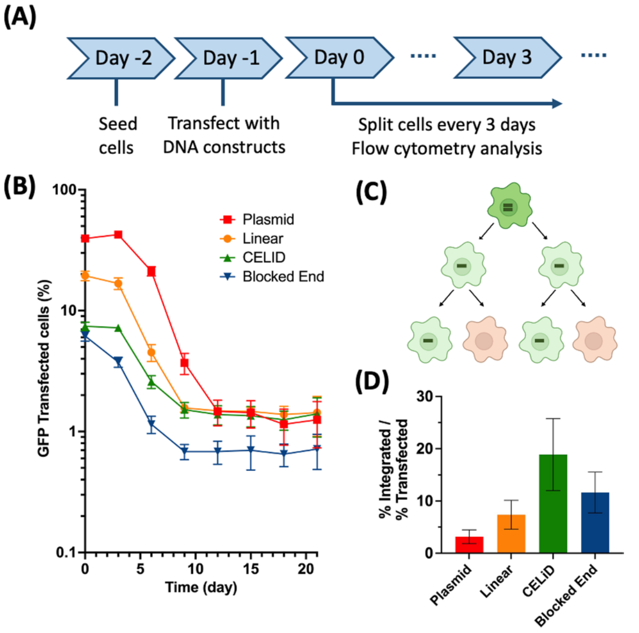

**Figure 3.** Comparing chromosomal integration rate of various DNA structures upon transfection into HEK293 cells. (**A**) Steps in the experiment. (**B**) Decrease in percentage of GFP expressing cells over time. Cells were split and analyzed by flow cytometry every 3 days until day 21. (**C**) Cartoon showing the segregation of transiently transfected DNAs during cell growth in an absence of drug selection. (**D**) Normalized integration frequency represented by (% GFP positive cells at day 21) / (% GFP positive cells at day 0).

cells were transfected with multiple DNA molecules that then segregated from each other during cell division. Upon the 2nd and 3rd splits, the proportion of cells expressing GFP was reduced five to tenfold per split, consistent with the idea that the reporter DNAs are being segregated without replicating, and cells at this stage contain one such DNA. On about day 9–12, the fraction of cells expressing GFP stabilized, formally indicating that the GFP DNAs are replicating as well as not conferring a selective disadvantage to the host cells, and suggesting that these DNAs have integrated. An alternative hypothesis is that a portion of the GFP expression vectors assume an extrachromosomal but replication-competent form; this possibility is addressed below.

Unexpectedly, our observations suggested that the end modification of linear DNA did not significantly decrease the rate of integration. At 21 days after transfection, circular plasmid, non-modified linear DNA and end-modified DNAs all showed integration between 0.5 – 2% of the total cell population, whereas the blocked end construct showed slightly decreased integration of 0.6%. When we normalize the final percentage of GFP-positive cells at day 24 to the initially transfected percentage at day 0, apparent integration potency was: CELiD > blocked end ~ non-modified linear > circular plasmid (Fig. 3D). Specifically, integration rate of the blocked end construct did not show a statistically significant difference compared to that of non-modified linear DNA (both ~ 10%), and the CELiDs showed a slightly increased frequency of integration (~ 20%). We observed a consistent result in a biological repeat experiment (Figure S3).

Data in Fig. 3 suggest that, on average, about 100 copies of transcriptionally active DNA enter a transfected cell. For the plasmid-transfected cells, the proportion of GFP + cells is relatively steady until after the second 1:10 split, suggesting that 100 plasmid DNA molecules present in an initially transfected cell are distributed into 100 cells at this point. According to this idea, the initially transfected cells go through six to seven cell divisions, during which non-replicating DNA is passively segregated into daughter cells. These results further suggest that in the plasmid-transfected cells, the fraction of plasmids that integrate is about $10^{-4}$ of the input plasmids. The quantitation of GFP expression in the transfected cells (Figure S2) does not indicate that silencing of transgene gene expression occurs, although the experiments were not designed to examine silencing per se.

To verify that the GFP-positive cell population after reaching plateau actually contains the reporter sequence in genomic DNA, we isolated the clones of individual fluorescent cells and analyzed their genomes for the presence of the GFP gene using a qPCR-based copy number assay. For each type of transfected DNA, two separate monoclonal cells were isolated from the population of cells stably expressing GFP (Fig. 4A). qPCR reactions specific to the GFP reporter revealed that all clones contained the target sequence, and that their copy numbers varied from 1 to 3 copies (Fig. 4B). In addition, inverse PCR experiment identified at least one integration junction from each monoclonal cell genome to support our result (Fig. 4C). Thus, we confirmed that the lipofectamine-mediated transfection with plasmids and various linear DNAs had indeed led to the stable incorporation of genes into the chromosomes.

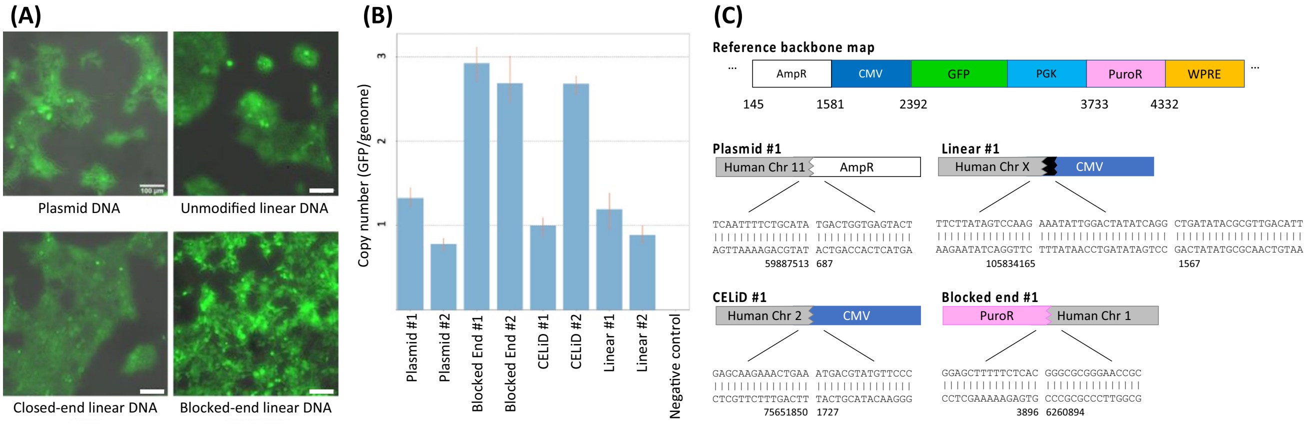

**Figure 4.** Gene integration analysis using cloned cells. GFP + cells from the experiment in Fig. 3B at day > 20 were isolated by FACS and plated at limiting dilution to obtain clones. (**A**) Images of monoclonal cells isolated from HEK293 cells transfected with each type of DNA construct. Images taken from brightfield and GFP channel were overlapped. Scale bars indicate 100 µm. (**B**) Copy number analysis of monoclonal cells transfected with different DNA types. For each DNA type, genomic DNA from two separate clonal cell lines was extracted and analyzed by qPCR using probes specific to the GFP gene. (**C**) Detailed sequences of the integration junctions identified for each type of monoclonal cells using an inverse PCR method.

## Discussion

The frequency at which transgenes spontaneously integrate into a mammalian chromosome is important in understanding the safety of gene therapy systems. Retroviruses were the first gene therapy delivery system, but the high integration frequency and ability to transactivate nearby genes, including oncogenes, has led to leukemia in patients[3]. Subsequent retroviral vectors have been engineered to eliminate transactivation, but this still leaves the potential problem of insertion into tumor suppressors. An alternative strategy has been to use "non-integrating" vectors such as Adeno-Associated Virus (AAV), or Closed-End Linear DNAs. Unlike retroviruses, AAV DNAs are not obligated to integrate during the virus life cycle[7], so this virus is dubbed "non-integrating." However, the dose of AAV used in gene therapy clinical trials is quite high, so even a low rate of integration could pose a risk of insertion into tumor suppressor genes. Moreover, in a mouse model, administration of AAV leads to oncogenic integration events[8,9].

We quantitated the frequency of spontaneous unselected chromosomal integration of transgenes that were delivered into cultured mammalian cells as covalently closed plasmids or in three different linear formats: DNA with exposed 5' and 3' ends, Closed-End Linear DNAs (CELiDs) with thioester loops, and DNA with biotinylated 5' and 3' ends further capped with streptavidin (Fig. 2). We developed the DNAs with non-natural ends to try to minimize the potential for non-homologous end joining (NHEJ). Within streptavidin, the biotin binding sites are the same distance apart as the 3' and 5' ends of DNA (~ 18 Angstroms), and dissociation of streptavidin from doubly biotin capped DNA was undetectable (Fig. 2C).

Several lines of evidence indicate that spontaneous, unselected integration occurred at a high rate. First, after transfection, the fraction of cells expressing GFP stabilized at 1–20% of the initially transfected cells; this fraction was stable over at least three to four tenfold splits. This is consistent with either integration or conversion into a replication-competent form, such as a long concatemer. Second, when clones of GFP + cells were isolated after FACS and extreme limiting dilution, quantitative PCR indicated that only one to three copies of transgene DNA are present. Finally, characterization of several cell clones identified junctions between transgene DNA and human chromosomal DNA.

Different forms of DNA integrate at different frequencies. After transfection into HEK293 cells, about 1–10% of the transiently transfected cells became stably transfected (Fig. 3). The circular plasmid DNA had a lower rate of integration than all of the linear forms, which showed comparable integration frequencies regardless of the configuration of the DNA ends. We also estimated that transfected cells received about 100 functional expression units per cell, based on the fact that the proportion of GFP-expressing cells stays roughly constant through two tenfold splits before decreasing. Thus, the integration frequency was about $10^{-4}$/DNA for circular plasmids and about $10^{-3}$ for linear DNAs. These results are consistent with a previous report[11], in which mice in vivo-transfected with AAV-derived CELiDs by hydrodynamic injection demonstrated more stable expression of the transgene in the liver, compared to expression in livers in vivo-transfected with a circular plasmid.

The linear DNAs all integrated at about the same frequency, and when we could identify junctions between human and plasmid DNA, the ends of the plasmid DNA were quite far from the end of the transfected DNA where it was cut by the restriction enzyme for linearization (Fig. 4C). The integration of these diverse forms of linear DNA suggests that the cell has a mechanism for eliminating non-natural DNA ends and then rescuing the DNA by integration.

These results may have implications that for AAV gene therapy, wherein viral particles deliver a mostly single-stranded DNA with closed hairpin ends. (AAV DNA ends typically have two hairpins instead of one.) Upon delivery to the nucleus, such DNAs are converted into a CELiD form[19,20]. Clinical trials of AAV-mediated delivery of genes for hemophilia A and B (Factor VIII and Factor IX deficiency, respectively) illustrate that the doses are quite high. George et al. used doses of $5 \times 10^{11}$ to $1.5 \times 10^{12}$ vector genomes per kg body weight in a Phase I–II study of Factor VIII delivery by AAV[21,22]. These correspond to absolute doses of $3.5 \times 10^{13}$ to $10^{14}$ vector genomes for a 70-kg person. The adult human liver contains about $1.5 \times 10^{11}$ cells ($10^8$ cells/gm; Liver mass is ~ 1.5 kg)[23,24]. Thus, the ratio of AAV transducing genomes to liver cells is about 100–1,000:1, and if the integration rate is the same as what we observed for linear DNAs in this work, the fraction of liver cells experiencing an integration event would be as high as 10% or more, assuming every AAV particle infects a cell. One important difference between our experiments and human AAV gene therapy is that HEK293 cells are rapidly dividing and liver cells are not, which may lead to differences in integration rates. Our analysis is consistent with observations from studies reporting transgene integration and tumorigenesis in animals injected with AAV vectors[7–9,25–29].

Currently, gene therapy aims to treat cancer or rare genetic diseases where long-term risks are acceptable. Further development of non-integrating gene delivery methods may need to precede wider application of gene therapy.

## Methods

### Cell culture and maintenance.

Dulbecco's Modified Eagle's Medium (DMEM) (ATCC 30-2002), Heat-inactivated Fetal Bovine Serum (FBS) (Thermo Fisher Scientific 10082147), and Penicillin–Streptomycin (Corning 30-002-Cl) were all vacuum sterilized by filtration (0.22 µm pore size, Corning, 430767) and used for maintaining cell growth. Homo sapiens HEK-293 T cells (ATCC, CRL-3216) were maintained in DMEM supplemented by FBS and Penicillin–Streptomycin at 37 °C. Once the cells reached 70% confluency, they were split into new flasks at one tenth of the density using trypsin (Trypsin EDTA 1X 0.25% Trypsin/ 2.21 mM EDTA in HBSS without sodium bicarbonate, calcium and magnesium, VWR 45000-664).

### Circular and Linear DNA synthesis.

The circular plasmid DNA in Fig. 1 (GenBank repository, accession number OQ117120), containing a GFP reporter expressed from the CMV promoter, and a Puromycin resistance cassette with a PGK promoter and WPRE sequence flanked by BsaI sites, was ordered from IDT. Unmodified plain linear DNA was synthesized by first digesting this plasmid with the *BsaI* restriction enzyme (New England Biolabs), followed by a purification using the PCR cleanup kit (Qiagen).

### CELiD DNA synthesis.

Two hairpin oligos were ordered from Integrated DNA Technologies (IDT). Their sequences were:

5'-GCCCTGAGAAACAGCTCTATCTGAC*C*A*C*A*T*GTCAGATAGAGCTGTTTCTCA-3'

5'-CCCGTTCCTTGTTCGTCACCAGCTC*A*T*G*C*A*GAGCTGGTGACGAACAAGGAA-3' where the asterisks indicate thioester linkages in the loop sections. Plain linear DNA was incubated with a tenfold molar excess of each hairpin oligo in a Golden Gate Assembly reaction (NEB Catalog #E1601). The Golden Gate reaction mix was then digested using Exo III (NEB Catalog #M0206S), and the reaction terminated by addition of EDTA and heating to 70 °C in accordance with the manufacturer's instructions. This eliminated vector DNA fragments and undesired ligation products and left a single 3900-bp DNA species corresponding to the end-modified insert (Fig. 1). The resulting synthesized CELiDs were purified using a PCR cleanup kit (Qiagen) to remove remaining oligonucleotide fragments. The resulting CELiD molecule was resistant to Exo III digestion at 37 °C for 30 min, a condition that resulted in a complete digestion of unmodified linear DNA (Supporting Figure S6). This result is consistent with the previous report that in vivo-synthesized CELiDs show exonuclease resistance[11].

### Blocked-end DNA synthesis.

Four Biotinylated oligonucleotides forming two double-stranded biotinylated fragments were ordered from IDT. Their sequences were:

5'Bio-CTCTATCATCAGATGTTCATTTAAATA-3'.

5'-CCCGTATTTAAATGAACATCTGATGATAGAG-Bio3'.

5'Bio-GAAGATCTGATGATACACATTTAAATA-3'.

5'-GCCCTATTTAAATCTGTATCATCAGATCTTC-Bio3'.

Each pair was mixed, heated at 95 °C and cooled slowly to anneal the two strands. Similar to the CELiD synthesis, plain linear DNA was incubated with 10 × molar excess of each biotinylated fragment in a Golden Gate Assembly. Resulting biotinylated DNA was then further incubated with 2 × molar excess of streptavidin for 10 min, followed by Exo III treatment to remove vector sequences and inserts that did not have streptavidin tetramers at each end. The reaction was terminated by addition of EDTA and heating to 70 °C, and then purified with a Qiagen PCR cleanup kit.

### Cell seeding.

Prior to seeding, 96-well cell culture plates were coated with a Purecol collagen solution (Advanced Biomatrix) diluted 1:30 with Phosphate-Buffered Saline (PBS) (Thermo Fisher Scientific, AM9625). Each well was incubated with 100 µL collagen for 1 h and rinsed with 150 µL PBS. HEK-293 T cells were washed with Phosphate-Buffered Saline (PBS) (Thermo Fisher Scientific, AM9625) and trypsinized (Trypsin EDTA 1X 0.25% Trypsin/ 2.21 mM EDTA in HBSS without sodium bicarbonate, calcium and magnesium, VWR 45000-664). Cell density was determined by optical density with an automated cell counter (Bio-Rad TC20), cell counting slides (Bio-Rad, 1450015) and a 0.4% solution of Trypan Blue (Bio-Rad, 1450021). After a live cell count was obtained, cells were diluted to $2.7 \times 10^5$ cells per mL in DMEM supplemented with 10% FBS, and 150 µL of cells were added to each well to seed $4 \times 10^4$ cells per well. Plates were then incubated at 37 °C for 24 h.

### Transfection of cells with various DNA constructs.

HEK-293 T cells were transfected 24 h after plating in 96-well plates and were carried out in DMEM supplemented with 10% FBS. Per well, 50 ng of DNA construct was transfected using lipofectamine 3000 kit (Thermo Fisher Scientific, L3000015). Each well was transfected with 10 µL of the Opti-MEM (Thermo 31985062) medium mixture containing 50 ng of DNA, 0.15 µL lipofectamine reagent, and 0.2 µL p3000 reagent. For each construct, a total of six wells were transfected as replicates. Medium was exchanged to fresh DMEM supplemented with 10% FBS 24 h after the addition of transfection reagent. Subsequently, 24 h later, cells were split 1:10 using trypsin–EDTA into a new 96-well plate, and remaining cells were used for the flow cytometry analysis. Similarly, cells were split 1:10 every three days during the 24-day experiment period.

### Flow cytometry.

Cells in a 96-well plate were trypsinized and resuspended in flow cytometry buffer (Invitrogen). Fluorescence was measured on a LSR II flow cytometer equipped with a HTS sampler (BD Biosciences) using the following filter configuration: excitation, 488 nm; emission, 530/30.

### Microscopy imaging.

Images were taken using a Nikon Eclipse Ti-2 widefield inverted microscope equipped with Hamamatsu Flash 4.0 LT camera at 20 × objective. For imaging green fluorescence, the following filter configuration was used: excitation, 466/40; emission, 525/50.

### Isolation of monoclonal cell population.

To generate individual clones of the fluorescent cells, transfected cells at day 24 were first sorted using a FACSAria II cell sorter (BD Biosciences) using the same filter configuration as above. Then, monoclonal cells were isolated from the sorted GFP-positive population using limiting dilution. Specifically, homogenized cells were diluted to concentration of 2.5 cells/mL, and 100 µL of this cell solution was transferred into each well of a 96-well plate. At this ratio, one out of four wells contained a single cell that should grow into monoclonal population. Seeded cells were cultured for 14 days, checked for fluorescence, and expanded into larger culture dishes. On average, 5–10 clones per 96-well plate were successfully isolated, which were less than expected; this result may be to low cell viability at a very low density.

### Copy number assay and inverse PCR.

Genomic DNA was extracted from monoclonal cells using DNeasy Blood and Tissue kit (Qiagen) following the manufacturer's instructions. Subsequently, qPCR-based copy number assay was carried out using Taqman Copy Number Assay and RNAse P Reference Assay (Thermo Fisher Scientific). qPCR probes specific to the GFP reporter gene was custom ordered from Thermo Fisher Scientific. Reactions were run using Quantstudio 7 Real-Time PCR system (Applied Biosystems), and subsequently analyzed for copy number using CopyCaller software (Applied Biosystems). For the inverse PCR, extracted genomic DNA was digested using a range of restriction enzymes including *EcoRI, BamHI, SphI* and *NcoI* (NEB) and ligated using T4 DNA ligase (NEB). Then, ligated circular DNAs were amplified using GFP reporter-specific primers (IDT) and Q5 hot start DNA polymerase (NEB). Amplified DNA was sequenced (Azenta) to identify the integration junction. Acquired sequences were compared with human genome sequence provided by UCSC Genome Browser (http://genome.uscs.edu)[30], using a genome assembly released in Dec 2013.

## Data availability

The datasets generated during and/or analyzed during the current study are either included in this published article (supporting information), or available from the corresponding author on reasonable request. Plasmid sequence is available in the GenBank repository, accession number OQ117120.

Received: 18 December 2022; Accepted: 20 April 2023
Published online: 26 April 2023

## References

1. Kitada, T., Diandreth, B., Teague, B. & Weiss, R. Programming gene and engineered-cell therapies with synthetic biology. *Science* **359**, eaad1067 (2018).
2. Khalil, A. S. & Collins, J. J. Synthetic biology: Applications come of age. *Nat. Rev. Genet.* **11**, 167–370 (2010).
3. Bulcha, J. T., Wang, Y., Ma, H., Tai, P. W. L. & Gao, G. Viral vector platforms within the gene therapy landscape. *Signal Transduct. Target. Ther.* **6**, 1–24 (2021).
4. Ramamoorth, M., Narvekar, A. Non-viral vectors in gene therapy-an overview. *J. Clin. Diagnostic Res.* **9**, GE01–06 (2015).
5. Hacein-Bey-Abina, S. *et al.* Insertional oncogenesis in 4 patients after retrovirus-mediated gene therapy of SCID-X1. *J. Clin. Investig.* **118**, 3132–3142 (2008).
6. Hacein-Bey-Abina, S. *et al.* LMO2-associated clonal T cell proliferation in two patients after gene therapy for SCID-X1. *Science* **302**, 415–419 (2003).
7. Dismuke, D. J., Tenenbaum, L. & Samulski, R. J. Biosafety of recombinant adeno-associated virus vectors. *Curr. Gene Ther.* **13**, 434–452 (2013).
8. Hanlon, K. S. *et al.* High levels of AAV vector integration into CRISPR-induced DNA breaks. *Nat. Commun.* **10**, 4439 (2019).
9. Nguyen, G. N. *et al.* A long-term study of AAV gene therapy in dogs with hemophilia A identifies clonal expansions of transduced liver cells. *Nat. Biotechnol.* **39**, 47–55 (2021).
10. Brooks, P. J., Yang, N. N. & Austin, C. P. Gene therapy: The view from NCATS. *Hum. Gene Ther.* **27**, 7–13 (2016).
11. Li, L. *et al.* Production and characterization of novel recombinant adeno-associated virus replicative-form genomes: A eukaryotic source of DNA for gene transfer. *PLoS ONE* **8**, e69879 (2013).
12. Samulski, R. J. & Muzyczka, N. AAV-mediated gene therapy for research and therapeutic purposes. *Annu. Rev. Virol.* **1**, 427–451 (2014).
13. Chang, X. B. & Wilson, J. H. Modification of DNA ends can decrease end joining relative to homologous recombination in mammalian cells. *Proc. Natl. Acad. Sci. U. S. A.* **84**, 4959–4963 (1987).

14. Ghanta, K. S. *et al.* 5′-Modifications improve potency and efficacy of DNA donors for precision genome editing. *eLife* **10**, e72216 (2021).
15. Canaj, H. *et al.* Deep profiling reveals substantial heterogeneity of integration outcomes in CRISPR knock-in experiments. *BioRxiv* 841098 (2019).
16. Chivers, C. E. *et al.* A streptavidin variant with slower biotin dissociation and increased mechanostability. *Nat. Methods* **7**, 391–393 (2010).
17. Integrated DNA Technologies web site. "Which biotin modification to use." https://www.idtdna.com/pages/education/decoded/article/which-biotin-modification-to-use-.
18. Weber, P. C., Ohlendorf, D. H., Wendolowski, J. J. & Salemme, F. R. Structural origins of high-affinity biotin binding to streptavidin. *Science* **243**, 85–88 (1989).
19. Berns, K. I. The unusual properties of the AAV inverted terminal repeat. *Hum. Gene Ther.* **31**, 518–523 (2020).
20. Zhong, L. *et al.* Single-polarity recombinant adeno-associated virus 2 vector-mediated transgene expression In Vitro and In Vivo: Mechanism of transduction. *Mol. Ther.* **16**, 290–295 (2008).
21. George, L. A. *et al.* Multiyear factor VIII expression after AAV gene transfer for hemophilia A. *N. Engl. J. Med.* **385**, 1961–1973 (2021).
22. Konkle, B. A. *et al.* BAX 335 hemophilia B gene therapy clinical trial results: Potential impact of CpG sequences on gene expression. *Blood* **137**, 763–774 (2021).
23. Wilson, Z. E. *et al.* Inter-individual variability in levels of human microsomal protein and hepatocellularity per gram of liver. *Br. J. Clin. Pharmacol.* **56**, 433–440 (2003).
24. Molina, D. K. & DiMaio, V. J. M. Normal organ weights in men: Part II-the brain, lungs, liver, spleen, and kidneys. *Am. J. Forensic Med. Pathol.* **33**, 368–372 (2012).
25. Bell, P. *et al.* Analysis of tumors arising in male B6C3F1 mice with and without AAV vector delivery to liver. *Mol Ther.* **14**, 34–44 (2006).
26. Donsante, A. *et al.* Observed incidence of tumorigenesis in long-term rodent studies of rAAV vectors. *Gene Ther.* **8**, 1343–1346 (2001).
27. Donsante, A. *et al.* AAV vector integration sites in mouse hepatocellular carcinoma. *Science* **317**, 477–477 (2007).
28. Deyle, D. R. & Russell, D. W. Adeno-associated virus vector integration. *Curr. Opin. Mol. Ther.* **11**, 442–447 (2009).
29. Chandler, R. J., Sands, M. S. & Venditti, C. P. Recombinant adeno-associated viral integration and genotoxicity: Insights from animal models. *Hum. Gene Ther.* **28**, 314–322 (2017).
30. Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).

## Acknowledgements

This research was supported by a DARPA (DARPA-STTR Phase I Contract 140D0420C0057 and DARPA-STTR Phase II Contract HR001121C0142). We thank Paul Sheehan and Edward Mandell for constructive input on the experiments.

## Author contributions

S. L. designed the research, performed the experiments, analyzed the data and wrote the paper. R. Y. designed, synthesized and characterized the DNA constructs. J.W. and P.S. analyzed the data, contributed to experiment design and edited the paper.

## Competing interests

J. W. and R. Y are employed by General Biologics, Inc., a for-profit corporation, and P. S. is a founder of General Biologics, Inc. S. L. declares no competing financial interests. The authors declare that no experiments on live humans were carried out in this study.

## Additional information

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-33862-0.

**Correspondence** and requests for materials should be addressed to S.L. or J.C.W.

**Reprints and permissions information** is available at www.nature.com/reprints.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.