Molecular Therapy
Methods & Clinical Development
Original Article

Quantification of transgene expression in GSH AAVS1 with a novel CRISPR/Cas9-based approach reveals high transcriptional variation

Anne Inderbitzin,[1,2,3,4] Tom Loosli,[1,2,3,4] Roger D. Kouyos,[1,2] and Karin J. Metzner[1,2]

[1]Department of Infectious Diseases and Hospital Epidemiology, University Hospital Zurich, Rämistrasse 100, 8091 Zurich, Switzerland; [2]Institute of Medical Virology, University of Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland; [3]Life Science Zurich Graduate School, University of Zurich, Zurich, Switzerland

Genomic safe harbors (GSH) are defined as sites in the host genome that allow stable expression of inserted transgenes while having no adverse effects on the host cell, making them ideal for use in basic research and therapeutic applications. Silencing and fluctuations in transgene expression would be highly undesirable effects. We have previously shown that transgene expression in Jurkat T cells is not silenced for up to 160 days after CRISPR-Cas9-mediated insertion of reporter genes into the adeno-associated virus site 1 (AAVS1), a commonly used GSH. Here, we studied fluctuations in transgene expression upon targeted insertion into the GSH AAVS1. We have developed an efficient method to generate and validate highly complex barcoded plasmid libraries to study transgene expression on the single-cell level. Its applicability is demonstrated by inserting the barcoded transgene Cerulean into the AAVS1 locus in Jurkat T cells via the CRISPR-Cas9 technology followed by next-generation sequencing of the transcribed barcodes. We observed large transcriptional variations over two logs for transgene expression in the GSH AAVS1. This barcoded transgene insertion model is a powerful tool to investigate fluctuations in transgene expression at any GSH site.

INTRODUCTION

Transgene insertion is predominantly conducted through lentiviral and gamma-retroviral vectors, resulting in an almost random insertion into the human genome.[1–5] The insertion can lead to unpredictable interaction of the transgene with the host genome, such as attenuation or complete silencing of the transgene[6–11] or, more critical, leading to dysregulated expression of host genes.[11,12] The application of a gene-editing tool for therapeutic use with site-specific transgene insertion in genomic safe harbor (GSH) may reduce the risk of insertional mutagenesis.[11,13,14] Thereof, there is a great necessity to identify and characterize GSH sites. A GSH is defined as a genomic locus to support stable transgene expression while not interfering with endogenous gene functions or structure.[11,15,16] To date, no genomic site has been shown to qualify as *bona fide* GSH for safe therapeutic transgene insertion.[11] Nevertheless, the adeno-associated virus site 1 (AAVS1) is often used as a GSH for research purposes,[11,14,17–19] assumed to provide stable transgene expression[20] due to flanking insulator regions.[21] AAVS1 is situated on chromosome 19 between exon 1 and intron 1 of the protein phosphatase 1-regulatory subunit 12C (PPP1R12C) gene and is the preferred site for integration of adeno-associated virus (AAV) DNA.[17,22] Lombardo et al.[14] showed that the GSH AAVS1 supported transgene expression in several human cell types while using multiple promoters. The protein PPP1R12C was shown in most cell types to have an active and open chromatin and to be constitutively transcribed.[14] Even though the GSH AAVS1 is widely used for research purposes, more characterization of its suitability for future clinical applications is needed.

Recent studies have demonstrated that transgene expression, integrated into the GSH AAVS1, is not as stable as previously reported while different promoters and/or cell lines were used.[13,23,24] Ordovàs et al.[23] observed transgenic silencing through DNA methylation in embryonic stem cells and differentiated hepatocytes by using different cell type-specific promoters. Similarly, Bhagwan et al.[24] observed silencing of transgene expression in human induced pluripotent stem cells (hiPSC) reporter cell lines by using two different promoters (pCAG and pTRE promoter).

In our previous studies, HIV-1-based dual fluorophore vectors were generated and originally used to study HIV-1 latency. They consist of two reporter genes, Cerulean under the control of the HIV-1 promoter and mCherry, which is driven by the constitutive human elongation initiation factor 4A1 (heIF4A1) promoter and flanked by two insulators.[25,26] Targeted insertion of our HIV-1-based dual-fluorophore vector LTatCL[M] via CRISPR-Cas9 into the GSH AAVS1 in the same and convergent orientation showed continuous expression of both reporter genes for up to 160 days.[25]

Received 7 January 2022; accepted 3 June 2022;
https://doi.org/10.1016/j.omtm.2022.06.003.

[4]These authors contributed equally

Correspondence: Karin J. Metzner, MD, University Hospital Zurich, Department of Infectious Diseases and Hospital Epidemiology, Rämistrasse 100, 8091 Zurich, Switzerland.
E-mail: karin.metzner@usz.ch

To further characterize the GSH AAVS1, we examined the stochastic fluctuations upon insertion of a barcoded transgene into the GSH AAVS1. We first modified our HIV-1-based dual-fluorophore vector LTatCL[M][25] by inserting a barcode consisting of 10 random nucleotides. We then inserted the vector LTatC10NL[M] into the GSH AAVS1 in Jurkat T cells via CRISPR-Cas9 in the same and convergent orientation relative to the host gene and studied the barcode expression.

RESULTS

Generation of highly complex LTatC10NL[M] AAVS1 barcode libraries

We generated a barcoded version of our previously described HIV-1-based dual-fluorophore vector LTatCL[M][25,26] to transfect Jurkat T cells and integrate the vector in the GSH AAVS1 site in the human genome via CRISPR-Cas9. Targeted insertion of our HIV-1-based dual-fluorophore vector LTatCL[M] into the GSH AAVS1 was extensively studied in our previous study, showing continuous transgene expression for up to 160 days in cell populations. This model allows us to elucidate stochastic fluctuations of the HIV-1 LTR transcriptional activity at the GSH AAVS1 insertion site on a single-cell level by sequencing the barcodes in genomic DNA as well as in reverse-transcribed cellular RNA.

The novel dual-fluorophore vector LTatC10NL[M] contains a barcode of 10 random nucleotides (10N) inserted between Cerulean (C) and the HIV-1 3′ LTR (L) (Figure 1A). Cerulean-barcode expression is controlled by the HIV-1 5′ LTR (L) promoter and supported by the HIV-1 transactivator of transcription (Tat). The vector is flanked by AAVS1 homologous arms in two ways to integrate the vector in the same (s) and convergent (c) transcriptional orientation, respectively, relative to the host gene PPP1R12C (Figure 1A). The original vector LTatCL[M][25] was developed primarily to study HIV-1 latency. The new vector LTatC10NL[M], therefore, contains a second fluorophore (mCherry) driven independently of the HIV-1 promoter by the constitutive heIF4A1 promoter and flanked by two insulators (Figure 1A). This second cassette is not relevant to the current study.

The two generated barcoded plasmid libraries LTatC10NL[M] AAVS1 s and LTatC10NL[M] AAVS1 c were validated with different methods to analyze and confirm the high complexity of both libraries (Figure 1B). First, exemplary single bacterial colonies were sequenced by Sanger sequencing to ascertain the barcode structure.[27,28] Five of six sequences displayed barcodes of different composition, one sequence did not contain the insert. Second, each barcoded plasmid library was independently sequenced by NGS five times, and the fraction of plasmids containing barcodes was quantified by comparing the average coverage of the barcode region with the average coverage of the neighboring region (Figures 1B and S1A). The percentage of 10N-containing plasmids in the LTatC10NL[M] AAVS1 s and LTatC10NL[M] AAVS1 c libraries was 86% (95% CI: 81.4-90.6) and 90.2% (95% CI: 88.2-92.3), respectively (Figure S1B). These findings show that the vast majority of plasmids contained a 10-nucleotide-long barcode.

Next, we examined the frequencies of the four nucleotides at each of the 10 positions within the barcode for each of the five independent NGS samples per library. In a randomly composed barcode, the probability of a nucleotide being incorporated is equal in all positions. For both the LTatC10NL[M] AAVS1 s and LTatC10NL[M] AAVS1 c libraries, there is no position in which a specific nucleotide dominates in any of the 10 independent samples sequenced by NGS (Figures 2A and S2). These results indicate that the structure of the barcode is diverse and balanced for both libraries.

Based on the number of unique barcodes in each library, i.e., barcodes that were present in only one of the five independent NGS samples, a corresponding simulated barcode library was created as a reference. We detected 29,860 and calculated 39,926 unique barcodes for the observed and simulated LTatC10NL[M] AAVS1 s libraries, respectively, and detected 28,352 and calculated 37,925 unique barcodes for the observed and simulated LTatC10NL[M] AAVS1 c libraries, respectively (Figure 2B). 3,084 and 3,533 barcodes were detected in two samples in the observed LTatC10NL[M] AAVS1 s and LTatC10NL[M] AAVS1 c libraries, respectively. 40 and 42 barcodes were found in all five samples in the observed LTatC10NL[M] AAVS1 s and LTatC10NL[M] AAVS1 c libraries, respectively. While there are more unique barcodes and less overlap in the simulated libraries compared with the observed libraries, the latter still show a similar distribution as calculated in the simulated libraries.

It was reported that, in barcoded plasmid libraries, barcode counts follow a Poisson distribution with a small inflation of singlet reads;[29] 95% and 99.3% of barcodes in the LTatC10NL[M] AAVS1 s and LTatC10NL[M] AAVS1 c libraries, respectively, were observed in frequencies consistent with chance. However, in both libraries, we have identified some barcodes that do not correspond to this distribution and therefore occur more frequently than would be expected by chance (Figure 2C). In the LTatC10NL[M] AAVS1 s library, these accounted for 3.3%, 1%, and 0.7% of all barcodes and in the LTatC10NL [M] AAVS1 c library for 0.7% of all barcodes. We assumed varying bacterial growth rates, an effect that propagates exponentially, to be the origin of these overrepresented barcodes.

To validate the complexity of the barcoded plasmid libraries, we adapted the rarefaction technique—used in ecology to assess species diversity—to estimate barcode diversity.[30] Hereby, we fitted logistic growth models to the number of barcodes obtained from each additional sample with carrying capacity equal to the maximally possible number of barcodes and growth rates estimated from exponential models, as the initial growth rate in logistic models is approximately exponential.

Since the samples contained different numbers of unique barcodes, the growth rate estimation in the exponential models depends on

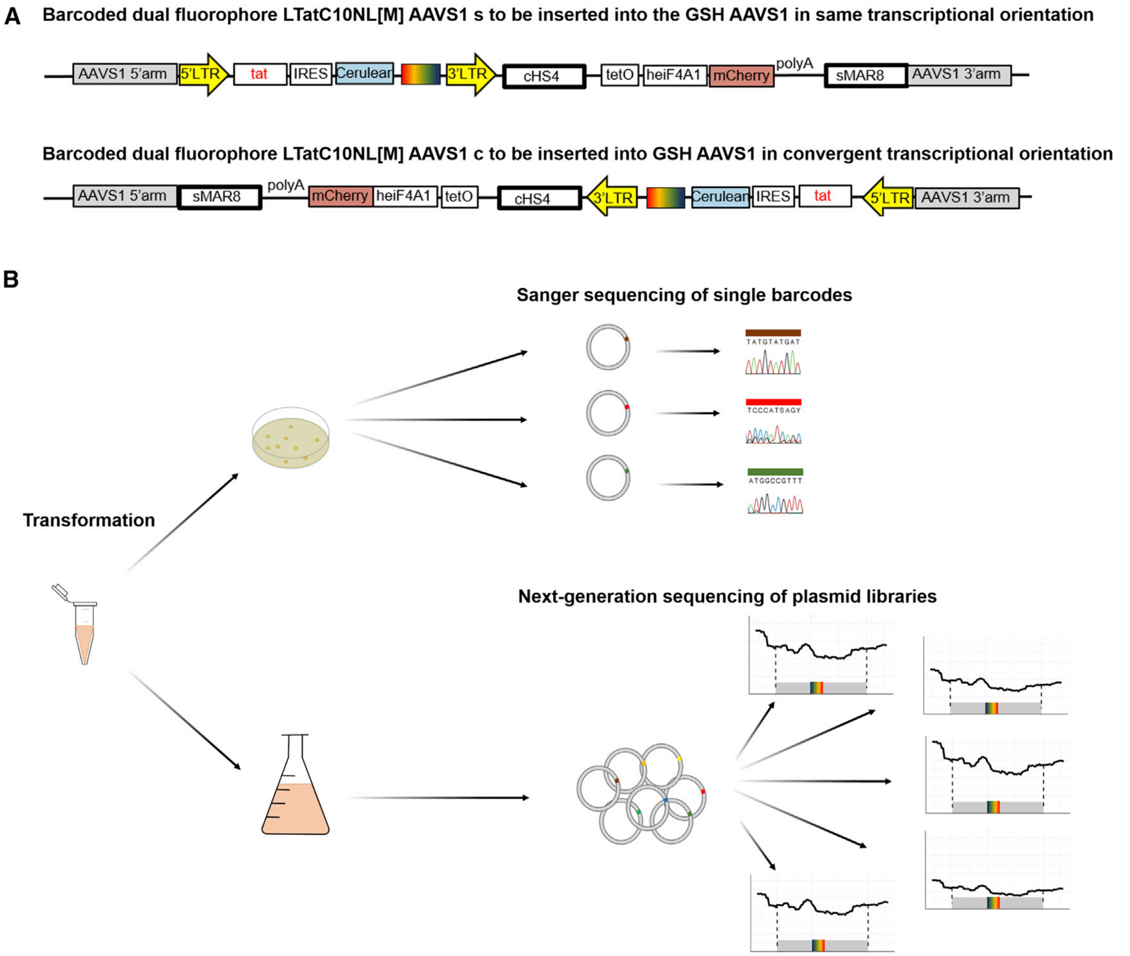

A
Barcoded dual fluorophore LTatC10NL[M] AAVS1 s to be inserted into the GSH AAVS1 in same transcriptional orientation
AAVS1 5'arm
5'LTR
tat
IRES
Cerulear
3'LTR
cHS4
tetO
heiF4A1
mCherry
polyA
sMAR8
AAVS1 3'arm
Barcoded dual fluorophore LTatC10NL[M] AAVS1 c to be inserted into GSH AAVS1 in convergent transcriptional orientation
AAVS1 5'arm
sMAR8
polyA
mCherry
heiF4A1
tetO
cHS4
3'LTR
Cerulean
IRES
tat
5'LTR
AAVS1 3'arm
B
Sanger sequencing of single barcodes
TATGTATGAT
TCCCATSAGY
ATGGCCGTTT
Transformation
Next-generation sequencing of plasmid libraries

Figure 1. Workflow to establish and validate barcoded LTatC10NL[M] AAVS1 plasmid libraries

(A) Scheme of the vectors LTatC10NL[M] AAVS1 s and LTatC10NL[M] AAVS1 c in the same (s) and convergent (c) transcriptional orientation, respectively. The barcode is marked with a rainbow-colored box. Flanking the vectors are 5′ and 3′ arms homologous to the GSH AAVS1 locus (gray). LTR, HIV-1 long terminal repeat; tat, HIV-1 transactivator of transcription; IRES, internal ribosomal entry site; cHS4, chicken hypersensitive site 4 (insulator); tetO7, tetracycline operator sequence, 7 repeats; helF4, human eukaryotic initiation factor 4A1; sMAR8, synthetic matrix attachment region 8 (insulator); LTatC10NL[M], 5′ **L**TR-**Tat**-Cerulean-barcode **10N**-3′ **L**TR-cHS4(**[**)-**m**Cherry-sMAR8(**]**). (B) Scheme of the generation and validation of barcoded LTatC10NL[M] AAVS1 s and AAVS1 c plasmid libraries. Depicted is the transformation with (1) plating of a small aliquot for assessment of quality control by Sanger sequencing of plasmids from single bacterial colonies to control barcode structure and (2) next-generation sequencing of the barcoded plasmid libraries from the bacterial suspension to estimate the number of barcodes and their complexity, which was performed five times for each library. The gray bar represents the insert containing the barcode depicted in rainbow-colored boxes.

the order of the samples. All permutations from the five samples were considered and filtered for R^2 and growth rate (Figure S3). The remaining models (Figure S3B) were used to parametrize the growth rate in the logistic models, which we used to generate barcode complexity accumulation curves to the amount of DNA we used to transfect Jurkat T cells at a later time point (Figure 2D). The accumulation curves were validated by similarity with the simulated data, allowing for high confidence in the complexity estimates for the barcoded plasmid libraries. We were able to show that the maximally possible number of 4^{10} unique barcodes would be represented at least once in approximately 1.2-μg barcoded plasmid library. To conclude, we developed a novel method to assess and verify the complexity of barcoded plasmid libraries and we could demonstrate that both barcoded plasmid libraries are highly complex.

Insertion of the barcoded transgene LTatC10NL[M] into the GSH AAVS1 in Jurkat T cells and estimated barcode occurrence in sorted cells

The barcoded LTatC10NL[M] AAVS1 s and LTatC10NL[M] AAVS1 c plasmid libraries were integrated in the GSH AAVS1, so the transgene was integrated either in the same or in the convergent orientation. Three independent experiments were performed, where one million Jurkat T cells were transfected with 2 μg of each of the barcoded plasmid libraries. We previously estimated the transfection

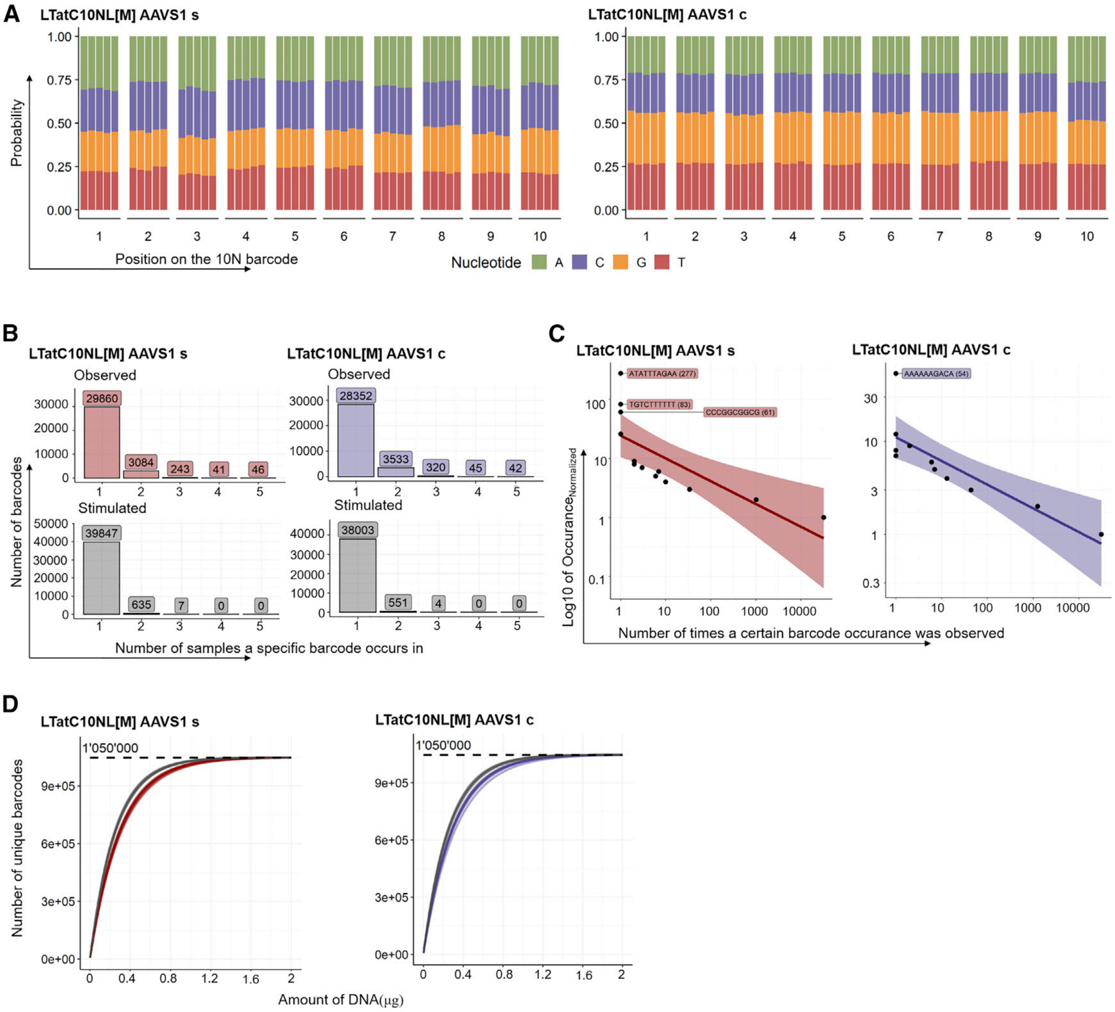


Figure 2. Barcoded plasmid libraries for LTatC10NL[M] AAVS1 s and LTatC10NL[M] AAVS1 c

(A) Frequency of nucleotides per position. The five columns at the 10 positions of the barcode represent the five-times sequenced libraries. (B) Comparison is shown of the observed overlap of barcodes between the five samples of each barcoded plasmid library with its simulated counterpart. (C) The number of times certain barcodes occurred, normalized to all five samples of each barcoded plasmid library. Expected Poisson distribution is shown for LTatC10NL[M] AAVS1 s in red line and LTatC10NL[M] AAVS1 c in blue line, and the shaded area corresponds to the 95% confidence interval. Barcodes that are observed more often than expected by chance are labeled. (D) The complexity in each of the barcoded plasmid libraries was estimated by extrapolating for complexity accumulation curves. This way, the number of barcodes available in the plasmid concentration used in nucleofection could be estimated. Simulated barcodes were analyzed identically to validate the model by similarity of observed (LTatC10NL[M] AAVS1 s, red lines and LTatC10NL[M] AAVS1 c, blue lines) and simulated data (gray lines).

efficiency to be 2%–10%.[25] On the basis of our estimations described above, we assumed that the transfected barcoded plasmids libraries contained all possible 4^{10} = 1,048,576 barcodes. To estimate the number of different barcodes in the 20,000 to 100,000 transfected cells, we compensated for the "birthday paradox,"[31] which states that in a group of 23 randomly selected people, the probability that at least two will have the same birthday is 50%, herein the same with barcodes. Thereof, we simulated drawing 20,000 and 100,000 barcodes from 4^{10} unique barcodes. Using 1,000 repeated simulations, we were able to robustly estimate the number of unique barcodes that were transfected to be 19,811 (95% CI: 19,810, 19,812) for the 2% transfection efficiency and 95,379 (95% CI: 95,375, 95,383) for the 10% transfection efficiency, respectively. After 8 days, 2,600–7,075 cells expressing Cerulean and mCherry were sorted (Figure S4). As done by Chen et al., we then calculated the percentage of a barcode occurring more than once in the Cerulean$^+$/mCherry$^+$ cells (Table 1). We concluded that barcode duplications might be present in the sorted cells, albeit only to a minor degree.

LTatC10NL[M] AAVS1 s- and LTatC10NL[M] AAVS1 c-transfected Jurkat T cells were expanded up to 14 days post-sorting and analyzed with flow cytometry showing Cerulean and mCherry expression

Table 1. Barcode occurrence in sorted cells carrying the barcoded transgenes LTatC10NL[M] AAVS1 s and LTatC10NL[M] AAVS1 c in three independent experiments

Sample	#Cells sorted	#Estimated barcode drawn	%Barcode estimated to occur more than once[a]
LTatC10NL[M] AAVS1 s, replicate 1	4,589	$\frac{4589}{19'811\ \textit{to}\ 95'379} = 0.232–0.048$	0.11–2.3
LTatC10NL[M] AAVS1 s, replicate 2	3,625	$\frac{3625}{19'811\ \textit{to}\ 95'379} = 0.183–0.038$	0.07–1.48
LTatC10NL[M] AAVS1 s, replicate 3	7,057	$\frac{7057}{19'811\ \textit{to}\ 95'379} = 0.356–0.074$	0.26–5.02
LTatC10NL[M] AAVS1 c, replicate 1	2,585	$\frac{2585}{19'811\ \textit{to}\ 95'379} = 0.13–0.027$	0.04–0.78
LTatC10NL[M] AAVS1 c, replicate 2	2,600	$\frac{2600}{19'811\ \textit{to}\ 95'379} = 0.131–0.027$	0.04–0.79
LTatC10NL[M] AAVS1 c, replicate 3	3,980	$\frac{3980}{19'811\ \textit{to}\ 95'379} = 0.042–0.201$	0.08–1.77

Depicted are the number of sorted Jurkat T cells carrying the barcoded transgenes LTatC10NL[M] AAVS1 s and LTatC10NL[M] AAVS1 c, each in triplicates. The number of estimated barcodes drawn upon sorting for the 2% and 10% transfection efficiency is calculated for each sample. Finally, the percentage of a barcode occurring more than once is calculated on this basis.

[a] As described by Chen et al. (2017),[28] these probabilities represent rare events, which can be well described using the Poisson distribution.[28] The probability X of a barcode being present in more than one of the nucleofected cells can thus be estimated with $P(X > 1)$, where X has a Poisson distribution with means of the probabilities of barcodes being drawn.

(Figure S5). We further characterized the cell populations as follows. We confirmed the insertion for both LTatC10NL[M] AAVS1 s and LTatC10NL[M] AAVS1 c in all six populations by amplifying and sequencing the junction of the insert and the AAVS1 locus[25] (Figure S6). We were also able to show that integration occurred in only one AAVS1 allele (Figure S7).

To summarize, we successfully integrated the barcoded plasmid LTatC10NL[M] AAVS1 s and LTatC10NL[M] AAVS1 c libraries into Jurkat T cells in three independent experiments.

Barcoded transgene LTatC10NL[M] shows little numerical overlap of barcodes between and within LTatC10NL[M] AAVS1 s and LTatC10NL[M] AAVS1 c

Integrated and expressed barcodes were characterized in LTatC10NL [M] AAVS1 s- and LTatC10NL[M] AAVS1 c-transfected Jurkat T cells in each of the three independent transfection experiments. After 14 days post-sorting, genomic DNA and cellular RNA were isolated, RNA was than reverse transcribed to cDNA, resulting in a total of 12 samples. Each sample was independently sequenced three times using NGS, resulting in 35/36 samples successfully sequenced (Figure S8). No NGS reads were obtained from one of the LTatC10NL[M] AAVS1 c-derived cDNA samples. The number of raw reads, trimmed and mapped reads, and total and unique barcode counts per sample for each library are shown in Table S2. We obtained 422, 624, and 574 unique barcodes from LTatC10NL[M] AAVS1 s cDNA replicates; 271, 635 and 854 unique barcodes from LTatC10NL[M] AAVS1 s DNA replicates; 445, 245, and 359 unique barcodes from LTatC10NL[M] AAVS1 c cDNA replicates; and 411, 848, and 780 unique barcodes from LTatC10NL[M] AAVS1 s DNA replicates (Figure 3). The barcode composition of LTatC10NL[M] AAVS1 s and AAVS1 c DNA and cDNA showed only a small overlap of 0–9 co-occurring unique barcodes within all three replicates, confirming the independence of the replicated experiments (Figures 3A and 3B).

To compare co-occurring unique barcodes between LTatC10NL[M] AAVS1 s and LTatC10NL[M] AAVS1 c DNA and cDNA, each of the three replicates of DNA and cDNA were combined resulting in 1,718, 1,610, 1,981 and 1,041 unique barcodes from LTatC10NL[M] AAVS1 s DNA and cDNA and from LTatC10NL[M] AAVS1 c DNA and cDNA, respectively. After merging of the replicates, 624 co-occurring unique barcodes were observed in LTatC10NL[M] AAVS1 s DNA and cDNA, and for LTatC10NL[M] AAVS1 c DNA and cDNA 539 co-occurring unique barcodes were observed(Figure 3C). When we compared co-occurring unique barcodes in independent LTatC10NL[M] AAVS1 s and LTatC10NL[M] AAVS1 c DNA and cDNA samples, we observed only 3–26 co-occurring unique barcodes (Figure 3C). This emphasizes the independence of the cell populations and reinforces the claim of our barcoded plasmid libraries being of high diversity.

In each dataset of LTatC10NL[M] AAVS1 s and LTatC10NL[M] AAVS1 c DNA and cDNA replicates, there were few dominant barcodes (Tables S3 and S4) To rule out unequal distribution of certain barcodes in the barcoded plasmid libraries, we compared the abundance of barcodes in LTatC10NL[M] AAVS1 s and LTatC10NL[M] AAVS1 c DNA and cDNA samples with the abundance of those barcodes in the barcoded plasmid libraries, and 218 and 189 barcodes co-occurred in LTatC10NL[M] AAVS1 s and LTatC10NL[M] AAVS1 c, respectively, and their corresponding plasmid libraries (Tables S3 and S4). The median GC content was 50% for the unique barcodes in the barcoded plasmid libraries for LTatC10NL[M] AAVS1 s and LTatC10NL[M] AAVS1 c (Figure S9). None of the high-abundance barcodes observed in cell samples were also found in high abundance in the barcoded plasmid libraries, indicating that during expansion some cells proliferated faster than others.

Next, we generated a heatmap to display the 15 most abundant DNA- and cDNA-derived barcodes across all samples (Figure 4). The

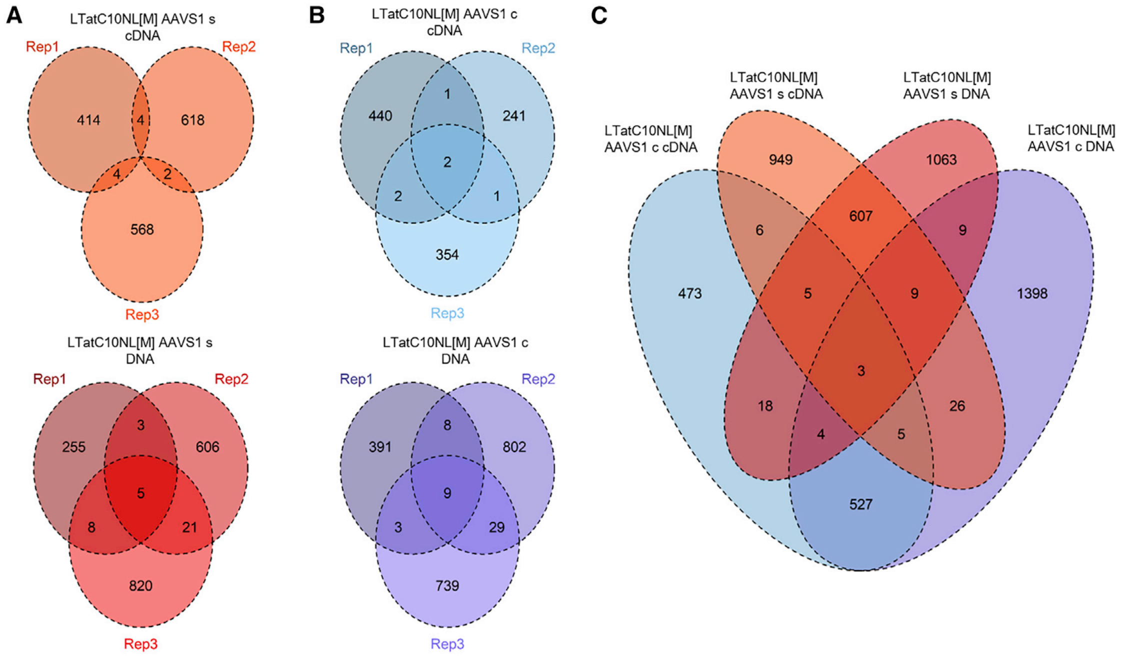


Figure 3. Overlap of unique barcodes between cDNA and DNA of LTatC10NL[M] AAVS1 s and AAVS1 c transfected Jurkat T cells

Depicted are the overlaps in unique barcodes from the LTatC10NL[M] AAVS1 s and LTatC10NL[M] AAVS1 c cDNA and DNA samples, each in three replicates from independently performed experiments, whereby each is comprised of NGS triplicates for (A) LTatC10NL[M] AAVS1 s- and (B) LTatC10NL[M] AAVS1 c-transfected Jurkat T cells, and in (C) overlap of merged three replicates shown in (A) and (B).

samples first clustered among the three replicates, and second, clustered according to their DNA or cDNA origin. Among those barcodes, minor overlap was observed between LTatC10NL[M] AAVS1 s and LTatC10NL[M] AAVS1 c samples and independent replicates.

To summarize, the DNA-derived barcodes overlapped with the cDNA-derived barcodes within LTatC10NL[M] AAVS1 s, and within LTatC10NL[M] AAVS1 c. There was only minute overlap between barcodes from LTatC10NL[M] AAVS1 s and LTatC10NL[M] AAVS1 c, indicating that there was no cross-contamination. Barcodes were heterogeneously expressed and a few highly abundant barcodes dominated in each sample.

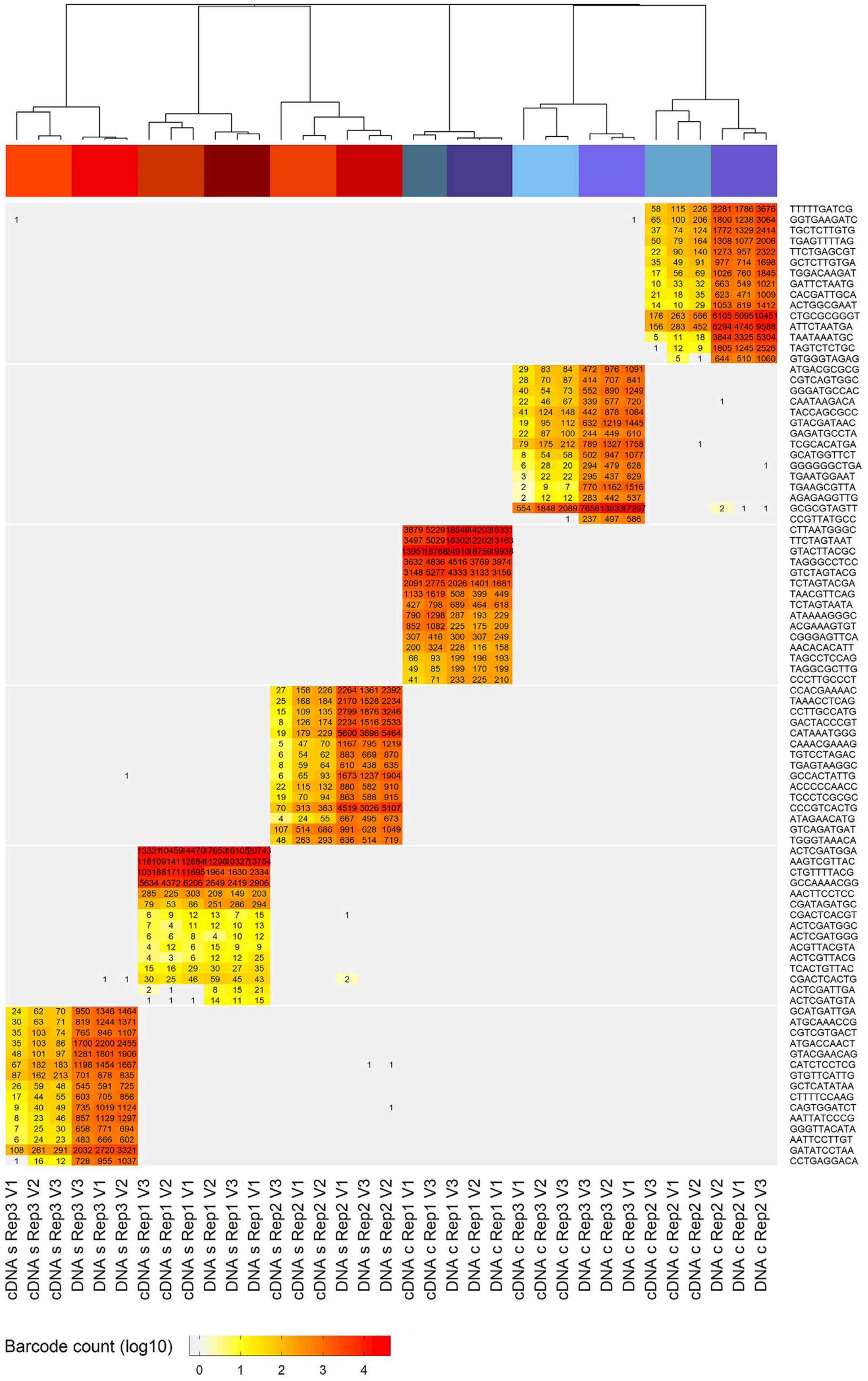
Barcoded transgene LTatC10NL[M] insertion into the GSH AAVS1 shows large transcriptional variation

To quantify transcriptional variation after insertion of the barcoded transgene LTatC10NL[M] into the GSH AAVS1, we compared the abundance of integrated barcodes, i.e., DNA samples with barcode expression in corresponding cDNA samples. For this analysis, the 607 and 526 barcodes exclusively co-occurring more than once in matched DNA and cDNA from LTatC10NL[M] AAVS1 s and LTatC10NL[M] AAVS1 c samples, respectively (Figure 3), were considered. Comparing the proportions of specific barcodes in both DNA and cDNA samples was used as normalization and had the advantage of not being affected by the number of reads generated during NGS. We could observe a similar distribution in barcode proportions between DNA- and cDNA-derived barcodes for the high-abundance barcodes (Figure S10). To qualitatively analyze variations in barcode expression, we calculated the relative fold change of barcodes by normalizing the occurrence of barcodes in the cDNA samples with the occurrence in the DNA samples in each replicate (Figure 5A). Upon combining the three replicates, the relative transcriptional fold change of barcodes integrated in both transcriptional orientations in AAVS1 ranges over two logs and was significantly higher in the LTatC10NL[M] AAVS1 c sample than in the LTatC10NL[M] AAVS1 s sample (1.29× higher, p value = 0.0023; Figure 5B). Herein, we observed high fluctuations in the expression of the barcoded transgene LTatC10NL[M] in the GSH AAVS1 in both transcriptional orientations. These results were consistent in a sensitivity analysis, where barcodes occurring only once in either DNA or cDNA samples were excluded (Figure S11).

DISCUSSION

We developed a novel model to explore stochastic fluctuations of barcoded transgene LTatC10NL[M] inserted in both orientations in the GSH AAVS1 on a DNA and RNA level in Jurkat T cells. We observed substantial variation in transcript levels of barcode expression spanning two logs, suggesting that the GSH AAVS1 is not as stable as previously reported and requires further investigation, particularly for therapeutic approaches.

The discovery of position effects on the reactivation potential of latently HIV-1-infected cells relied on the B-HIVE technology, where single-cell tracking using DNA barcodes was accomplished.[28] We



(legend on next page)

combined the latency model vector LTatCL[M] with the barcode insert as designed by Chen et al.[28] The barcoding method has the common three initial steps: (1) generation of barcoded vectors, (2) transformation, and (3) barcoded plasmid library control. Similar to other studies, we also addressed the lack of reliable established methods for quality control of barcoded plasmid libraries.[27–29,32,33] In particular, plating out a small aliquot of the transformed bacteria and counting the colonies might lead to variable results, as it depends on (1) the quality of the agar plate, (2) the person performing the experiment, (3) the plating out technique, and (4) the homogeneity of the bacterial concentration in the transformation mixture. Our approach is based on an accumulation curve using NGS data. It covers all apparent shortcomings of previous controls and allows us to accurately estimate the complexity of barcoded plasmid libraries, as well as to identify barcodes that occur more frequently than probable by chance, for potential exclusion in downstream analyses.

We inserted the barcoded transgene LTatC10NL[M] into the GSH AAVS1 in Jurkat T cells to explore stochastic fluctuation. To do this, we sorted for Cerulean$^+$/mCherry$^+$ cells representing cells that contain the entire vector cassette, as we demonstrated in our previous study, where no internal deletions were observed in Cerulean$^+$/mCherry$^+$ cells.[25] We successfully inserted the barcoded transgene LTatC10NL[M] into the GSH AAVS1 and in the correct respective orientations. We ruled out substantial off-target integration of our vector constructs because no evidence of frequent off-target integration was observed in another study using a similar method.[34] Those authors observed ~0.1% off-target integration using a dsDNA template for homology-directed repair (HDR) without Cas9 and ~1% off-target integration when Cas9 was present.

Our NGS data analysis showed that, in both cell populations, in which the construct LTatC10NL[M] was integrated into AAVS1 in the same and convergent orientation, the barcodes were dominated by a few specific sequences. Although we cannot completely exclude the possibility that the cells were already composed of unbalanced barcode proportions after transfection, it seems unlikely, considering the thorough quality control of the barcoded plasmid libraries and the estimated chance of duplicate barcode insertions is low. We are therefore inclined to believe that the observation of a few high-frequency barcodes is due to some steps and bottlenecks such as certain cells having a slightly higher fitness, thus dominating the population by expanding slightly faster after the bottleneck represented by cell sorting. Additional bottlenecks may occur during reverse transcription, where not all RNA molecules may be reverse transcribed, during PCR amplification of barcodes, where not all templates may be amplified and frequencies may be biased, and during sequencing, where library preparation may represent a bottleneck for individual fragments. We compared the barcodes extracted from the transfected LTatC10NL[M] AAVS1 s and LTatC10NL[M] AAVS1 c Jurkat T cells with those revealed during the barcoded plasmid library preparation and were able to identify a certain overlap, as expected. However, there was no correlation between high-frequency barcodes in the library and those extracted from Jurkat T cells. This validates the quality of our barcoded plasmid libraries and adds further evidence to the hypothesis that clonal expansion is the cause of the high-frequency barcodes in transfected LTatC10NL[M] AAVS1 s and LTatC10NL[M] AAVS1 c Jurkat T cells.

When the barcodes were analyzed for similarities within and between the three replicates, there were no obvious biases introduced during reverse transcription or PCR to amplify the barcode insert, as indicated by the very similar barcode profiles between replicates. The dendrogram based on the hierarchical clustering of the barcode profile shows that the replicates clustered primarily with each other, followed by clustering with their DNA or cDNA counterpart according to the same or convergent orientation. The heatmap of the barcode profile shows minor overlap between the independent samples, which was to be expected on the basis of the number of sorted barcoded cells compared with the complexity of the possible barcodes in each barcoded plasmid library.

We compared the proportion of barcodes in the cDNA samples to their proportion in the corresponding DNA samples and could observe a relative fold change of the transcription of barcodes spanning over two logs, which is a higher transcriptional variance than desired, given the insertion into a GSH site. Although barcode expression is significantly different in the LTatC10NL[M] AAVS1 s and LTatC10NL[M] AAVS1 c orientations, the difference is small compared with the range in barcode expression that we observed within the samples. Fluctuations of transgene expression was also observed in other studies; for example, Bhagwan et al.[24] observed variability in transgene expression in human induced pluripotent stem cell (hiPSC) reporter cell lines using two different promoters (pCAG and pTRE promoter). The selection of promoter might play a role in silencing of transgene expression, whose silencing was observed when using the EF-1α promoter. This effect could be overcome using a stronger CAG promoter,[35] exhibiting some insulation from methylation.[23] In our previous study, we were able to show stable transgene expression in the GSH AAVS1 in a monoclonal Jurkat T cell line over 160 days independent of transcriptional orientation, illustrated by reporter expression.[25] The fluctuation observed at the single-cell level may be due to the HIV-1 LTR promoter used in this model, which may be influenced by a position effect, as the

Figure 4. Barcode profile of LTatC10NL[M] AAVS1 s and LTatC10NL[M] AAVS1 c DNA and cDNA replicates of transfected Jurkat T cells

Heatmap of the number of observations for the 15 most abundant barcode expressions in the barcoded transgene LTatC10NL[M] AAVS1 s and AAVS1 c. The heatmap shows observed barcode counts and is color-coded by $\log_{10}$-transformed barcode counts. Barcodes depicted in rows and samples in columns. Barcode profiles were hierarchically clustered by their correlation and displayed as a dendrogram on top, visualizing similarity. All 35 samples from three independent experiments are displayed. The sample names are simplified: cDNA s, LTatC10NL[M] AAVS1 s cDNA; DNA s, LTatC10NL[M] AAVS1 s DNA; cDNA c, LTatC10NL[M] AAVS1 c cDNA;, and DNA c, LTatC10NL[M] AAVS1 c DNA.

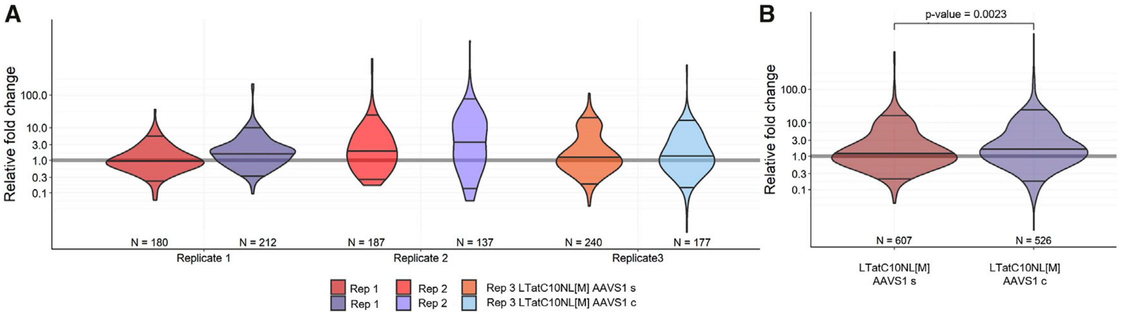


Figure 5. Quantification of the expression of barcodes integrated in the GSH AAVS1 in both transcriptional orientations

The violin plots show relative fold change on a $\log_{10}$ scale for all barcodes co-occurring in corresponding DNA and cDNA samples. (A) Relative transcriptional fold changes in barcode expression for the three replicates for DNA and cDNA. (B) Relative transcriptional fold changes of the aggregated three independent experiments (replicates) in same and convergent orientation. *t*-test was used to test for statistical significance of log transformed values (p value = 0.0023).

HIV-1 LTR promoter has been reported to be sensitive to the local chromatin environment.[36]

The benefit of using a barcoded transgene in research and clinical application is to study a cell population on a single-cell level, so that clonal dominance can be detected in a sensitive and relatively simple way. This model can be further used for research applications in cancer or infectious disease therapy, e.g., not only for the validation of GSH but also for the assessment of drug resistance mutations or reactivation of novel or already used drugs at the single-cell level. The vector can be easily adapted by, for instance, replacing the HIV-1 promoter with any other promoter of interest. In addition, to overcome transgene variability for research and clinical use, the transgene could be flanked by insulators.

Conclusion

We developed a method of constructing barcoded plasmid libraries and a novel quality control tool thereof, fulfilling the demand for appropriate quality assessment tools. It allows for a detailed and thorough insight into barcoded plasmid libraries necessary for any downstream analysis. We applied this by generating barcoded plasmid libraries based on the HIV-1-based dual-fluorophore vector LTatC10NL[M] for CRISPR-Cas9-mediated targeted insertion into the genomic safe-harbor locus AAVS1 in both transcriptional orientations in Jurkat T cells. We characterized barcode expression in the Cerulean$^+$/mCherry$^+$ population based on DNA and RNA, whereby we could show a substantial variation in transcriptional levels of barcode expression, spanning two logs. Our barcoded model allows the characterization of stochastic fluctuations of transgene insertion in any GSH.

MATERIALS AND METHODS

Generation of LTatC10NL[M] AAVS1 same (s) and LTatC10NL[M] AAVS1 convergent (c) barcode libraries

The dual-fluorophore vector LTatCL[M][25] was modified to LTatC10NL[M]; i.e., a barcode consisting of 10 random nucleotides (10N) was inserted downstream of the reporter gene Cerulean (Figure 1). We designed the insert as described in Chen et al.,[28] consisting of a human T7 promoter and an Illumina primer. Briefly, the oligonucleotides containing the 10N barcode were amplified to obtain double-stranded DNA fragments flanked by vector-overlapping sequences for subsequent infusion cloning. All oligonucleotides were purchased from Microsynth and are listed in the Table S1. The PCR reaction contained 10 ng of the barcoded DNA oligonucleotide Cerulean-10N-LTR, 0.5 μM each of forward and reverse oligonucleotide InFusion-cerulean-fw and InFusion-LTR-rc, and the CloneAmp HiFi PCR Premix (Takara Bio). The PCR cycling conditions were as follows: 35 cycles at 98°C for 10 s, 57°C for 15 s, and 72°C for 15 s. The 173-bp-long amplicon was purified using the NucleoSpin Gel and PCR Clean Up kit (Macherey-Nagel). The plasmids LTatCL[M] AAVS1 s and AAVS1 c[25] were linearized using 100 units of AsiSI (NEB)/20 μg DNA. The 10-kb-long fragment was purified using the NucleoSpin Gel and PCR Clean Up kit. In-Fusion Cloning (TaKaRa) was performed following the manufacturer's instruction in a 1:5 ratio of linearized plasmids LTatCL[M] AAVS1 s and AAVS1 c, respectively, and the barcode insert. After transformation of XL10-gold ultracompetent cells (Agilent Technologies), an aliquot was plated and incubated overnight, while the remaining bacteria were incubated in Luria-Bertani (LB) medium with 100 μg/mL ampicillin overnight at 300 rpm at 37°C. Sanger sequencing was used to confirm PCR products and cloning intermediates and was performed by Microsynth.

Analysis of the structure of the LTatC10NL[M] AAVS1 s and LTatC10NL[M] AAVS1 c barcoded libraries by next-generation sequencing

To analyze the structure and complexity of the barcoded plasmid libraries, both libraries were independently sequenced five times using the Illumina MiSeq next-generation sequencing (NGS) platform (MiSeq Reagent Kit v3, 150-cycle). The sequence reads were mapped to the reference in CLC Genomics Workbench 20 (Qiagen). The following analyses were performed.

First, barcodes were identified based on their position inherited by the mapping reference. We evaluated the proportion of integrated vectors that contained the barcode fragment compared with those without. The coverage at the 10 barcode positions relative to the coverage 15 bases up- and downstream of the entire insert was used to estimate the percentage of the integrated vectors that contained a barcode, as mapped non-barcoded inserts result in a local drop in coverage.

Second, the frequencies of the four bases at each of the 10 positions were analyzed to ensure not being biased toward specific nucleotides, and the GC content of the barcodes was calculated. A sequence logo of the barcode region was generated with the ggseqlogo package.[37]

Further analysis was restricted to complete barcodes. These sequences were controlled for their uniqueness, and the number of unique barcodes was calculated. To ensure that the results were not influenced by mapping errors, we additionally checked the local concordance of the bases up- and downstream of the barcodes with the expected reference nucleotides.

Third, we compared the number of unique barcodes obtained per NGS sample to the expected number of unique barcodes based on a simulation relying on random sampling. Thereby, pseudo-barcodes were drawn as many times as barcodes were found in each NGS sample from a pool of 4^{10} = 1,048,576 possible barcodes. We compared the overlap in barcodes between the NGS samples with the overlap of the simulated samples.

Fourth, to estimate the complexity of the barcoded plasmid libraries, we developed a new quality control method based on rarefaction, a technique used in ecology to quantify biodiversity. By randomly resampling a pool of X samples and plotting the average number of unique species, the produced rarefaction curve allows estimation of the expected number of species in the total pool,[30] wherein we considered the barcoded plasmid library as a pool and single barcodes as species.

The complexity of the barcode ensemble and the simulated data generated in the third step were estimated and compared. The cumulative number of total unique barcodes was calculated for all 120 possible permutations of the five NGS samples (beginning with the first NGS sample and then calculating the total number of unique barcodes when adding the second NGS sample and then the total number of unique barcodes when also adding the third NGS sample, etc.). Exponential models were fitted to the unique barcode counts for each of the 120 permutations of the observed as well as the simulated data to estimate the complexity growth rate. The estimates of models with an $R^2 \geq 0.975$ and the growth rate <1 were used to parametrize logistic growth models with the carrying capacity equal to 4^{10} to estimate the barcode complexity in the number of barcoded plasmids used in the subsequent nucleofection (2 μg) by averaging the complexity estimates of all logistic growth models at the corresponding number of samples necessary for this amount of DNA. The chance of duplicated insertion was estimated by taking into consideration transfection efficiency and the number of cells to be barcoded.

Fifth, the occurrence of specific barcodes versus their abundance was analyzed as described by Davidsson et al.[29] Since multiple NGS samples were available, the occurrence of barcodes was normalized, averaging their occurrence per available number of NGS samples for both observed and simulated data.

CRISPR-Cas9-mediated insertion of the barcoded transgene LTatC10NL[M] in both orientations in the GSH AAVS1 into Jurkat T cells

Prior to transfection, the two vectors LTatC10NL[M] AAVS1 s and AAVS1 c were linearized using 100 units of NsiI (NEB)/20 μg DNA. The transfection protocol using those vectors in combination with pX458_gAAVS1 expressing the guide RNA was performed as previously described.[25] The pX458_gAAVS1 contained the same gRNA sequence as the AAVS1 T2 gRNA.[17] Transfected Jurkat T cells[38] were cultured in RPMI 1640 medium supplemented with 10% fetal bovine serum (FBS) and 1% penicillin-streptomycin (10,000 U/mL penicillin, 10 mg/mL streptomycin). Eight days post-transfection, the cells were sorted for the Cerulean$^+$/mCherry$^+$ population by use of a BD FACSAria III (BD Biosciences). Cells were expanded and analyzed by flow cytometry 14 days post-nucleofection (dpn) using the LSRFortessa II (BD Biosciences), and data were analyzed using the FlowJo Software v.10.0.8. (FLOWJO, LLC).

Amplification of the barcode in cDNA and DNA

Cellular RNA and DNA were extracted 14 days post-nucleofection from 5 million transfected barcoded cells, using the All Prep DNA/RNA kit (Qiagen) and quantified by Nanodrop 1000. RNA (800 ng) was reverse transcribed with Prime Script Reverse Transcriptase (Takara) according to the manufacturer's instructions. Each amplification of the barcodes in cDNA and DNA was performed in three replicates to compensate for any potential bias during PCR. PCRs were performed using the Phusion DNA polymerase (Thermo Fisher Scientific) in 1× Phusion GC buffer with 200 μM dNTPs and 1 μM of each oligonucleotide Barcode_cDNA_IlluminaPrimer_Fw and RT-Primer_cer_rc (Table S1). The PCR cycling conditions were as follows: 98°C for 1 min; 29 cycles of 98°C for 20 s, 63°C for 30 s, 72°C for 1 min, and 72°C for 5 min. The amplicons were purified using the NucleoSpin Gel and PCR Clean Up kit and paired-end sequenced with an Illumina MiSeq using the MiSeq Reagent Kit v.3 (150-cycle).

Confirming insertion site and orientation of the barcoded transgene LTatC10NL[M] in the GSH AAVS1

The insertion site was confirmed by amplifying the junction between the inserted DNA and the intended locus, as previously described.[25]

Data analyses of barcoded Jurkat T cells

Sequence trimming and alignment were performed using CLC Genomics Workbench v.20. Trimming was based on an automatic removal of adapter sequences, removal of low-quality sequences (limit of 0.01 in quality score, which refers to base calling error probability[39]), and removal of sequences shorter than 15 and longer than 1,000 nucleotides. Read mapping was done with affine gap costs for barcode inserts to optimize the mapping of inserts (e.g., barcodes). Length and similarity fraction were set to 0.8 to reduce the number of mismatches. All statistical tests were performed in R with default settings.

The barcode analysis of the resulting bam files was performed with custom R scripts. Barcodes were identified based on their position inherited by the mapping reference. The three replicates were combined, keeping track of their origin, and the number of unique barcodes for every sample was examined. The replicates were compared among one another regarding the unique barcodes they represented and the number of occurrences thereof.

The VennDiagram[40] package was used to construct Venn diagrams visualizing overlaps among and across samples, meaning within the three replicates and among the DNA- and cDNA-derived samples, where replicates were merged prior to this analysis.

By use of the barcodes and their number of occurrences of each sample, a heatmap was generated using the 15 most abundant barcodes for DNA and cDNA origin. The expression profile was clustered by correlation to control whether the three replicates of each sample were indeed most similar to each other, and on the next level whether the orientation of the barcode clustered DNA and cDNA together. Furthermore, the independence of barcode sequences between the forward and reverse oriented samples was confirmed.

The three replicates of each sample were combined for barcode expression analysis. Since there were different numbers of barcodes available for each sample and among replicates, the barcodes were normalized by calculating the observed proportion of a specific barcode in a sample. A comparison of barcode-specific observed proportion was compared between DNA and cDNA samples in each replicate and merged afterward. Relative fold change was calculated by normalizing cDNA proportion with DNA proportion for each barcode. A sensitivity analysis was performed excluding barcodes occurring only once in both DNA and cDNA samples in order to reduce noise. Differences in fold change between transcriptional orientations was tested for significance by *t*-test.

DATA AVAILABILITY

All relevant data are within the paper and its Supporting Information files. All raw data are available on Zenodo (https://doi.org/10.5281/zenodo.6599269).

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.omtm.2022.06.003.

ACKNOWLEDGMENTS

We are very thankful to Dr. Claudia Dumrese, Dr. Sebastian Utz, and Mario Wickert (Flow Cytometry Facility, UZH, Zurich, Switzerland) for the cell-sorting service, and Gabriela Ziltener, Daniela Lang, Stefan Schmutz, and Michael Huber (Institute of Medical Virology, UZH) for next-generation sequencing. This study was funded by the Swiss National Science Foundation Grant 310030_141067 (to K.J.M.) and 310030_204404 (to K.J.M.) and from the Forschungskredit Candoc, GRANT FK-19-032 (to A.I.). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

AUTHOR CONTRIBUTIONS

A.I. and T.L. designed the study, performed the experiments, analyzed the data, and wrote the manuscript. R.D.K. supported the analysis of the data. K.J.M. invented and designed the study, supported the analysis of the data, and edited the manuscript. All authors approved the manuscript.

DECLARATION OF INTERESTS

K.J.M. has received travel grants and honoraria from Gilead Sciences, Roche Diagnostics, GlaxoSmithKline, Merck Sharp & Dohme, Bristol-Myers Squibb, ViiV and Abbott; the University of Zurich received research grants from Gilead Science, Novartis, Roche, and Merck Sharp & Dohme for studies wherein K.J.M. serves as principal investigator, and advisory board honoraria from Gilead Sciences. All other authors declare no conflict of interest.

REFERENCES

1. Cartier, N., Hacein-Bey-Abina, S., Bartholomae, C.C., Veres, G., Schmidt, M., Kutschera, I., Vidaud, M., Abel, U., Dal-Cortivo, L., Caccavelli, L., et al. (2009). Hematopoietic stem cell gene therapy with a lentiviral vector in X-linked adrenoleukodystrophy. Science *326*, 818–823.
2. Gaspar, H.B., Cooray, S., Gilmour, K.C., Parsley, K.L., Zhang, F., Adams, S., Bjorkegren, E., Bayford, J., Brown, L., Davies, E.G., et al. (2011). Hematopoietic stem cell gene therapy for adenosine deaminase-deficient severe combined immunodeficiency leads to long-term immunological recovery and metabolic correction. Sci. Transl. Med. *3*, 97ra80. https://doi.org/10.1126/scitranslmed.3002716.
3. Hacein-Bey-Abina, S., Pai, S.Y., Gaspar, H.B., Armant, M., Berry, C.C., Blanche, S., Bleesing, J., Blondeau, J., de Boer, H., Buckland, K.F., et al. (2014). A modified γ-retrovirus vector for X-linked severe combined immunodeficiency. N. Engl. J. Med. *371*, 1407–1417. https://doi.org/10.1056/nejmoa1404588.
4. Biffi, A., Montini, E., Lorioli, L., Cesani, M., Fumagalli, F., Plati, T., Baldoli, C., Martino, S., Calabria, A., Canale, S., et al. (2013). Lentiviral hematopoietic stem cell gene therapy benefits metachromatic leukodystrophy. Science *341*, 1233158.
5. Aiuti, A., Biasco, L., Scaramuzza, S., Ferrua, F., Cicalese, M.P., Baricordi, C., Dionisio, F., Calabria, A., Giannelli, S., Castiello, M.C., et al. (2013). Lentiviral hematopoietic stem cell gene therapy in patients with Wiskott-Aldrich syndrome. Science *341*, 1233151.
6. Martin, D.I.K., and Whitelaw, E. (1996). The vagaries of variegating transgenes. Bioessays *18*, 919–923. https://doi.org/10.1002/bies.950181111.
7. Kioussis, D., and Festenstein, R. (1997). Locus control regions: overcoming heterochromatin-induced gene inactivation in mammals. Curr. Opin. Genet. Dev. *7*, 614–619. https://doi.org/10.1016/s0959-437x(97)80008-1.
8. Rivella, S., and Sadelain, M. (1998). Genetic treatment of severe hemoglobinopathies: the combat against transgene variegation and transgene silencing. Semin. Hematol. *35*, 112–125.
9. Bestor, T.H. (2000). Gene silencing as a threat to the success of gene therapy. J. Clin. Invest. *105*, 409–411. https://doi.org/10.1172/jci9459.
10. Ellis, J. (2005). Silencing and variegation of gammaretrovirus and lentivirus vectors. Hum. Gene Ther. *16*, 1241–1246. https://doi.org/10.1089/hum.2005.16.ft-126.
11. Papapetrou, E.P., and Schambach, A. (2016). Gene insertion into genomic safe harbors for human gene therapy. Mol. Ther. *24*, 678–684. https://doi.org/10.1038/mt.2016.38.

12. Wu, C., and Dunbar, C.E. (2011). Stem cell gene therapy: the risks of insertional mutagenesis and approaches to minimize genotoxicity. Front. Med. *5*, 356–371. https://doi.org/10.1007/s11684-011-0159-1.

13. Klatt, D., Cheng, E., Hoffmann, D., Santilli, G., Thrasher, A.J., Brendel, C., and Schambach, A. (2020). Differential transgene silencing of myeloid-specific promoters in the AAVS1 safe harbor locus of induced pluripotent stem cell-derived myeloid cells. Hum. Gene Ther. *31*, 199–210. https://doi.org/10.1089/hum.2019.194.

14. Lombardo, A., Cesana, D., Genovese, P., Di Stefano, B., Provasi, E., Colombo, D.F., Neri, M., Magnani, Z., Cantore, A., Lo Riso, P., et al. (2011). Site-specific integration and tailoring of cassette design for sustainable gene transfer. Nat. Methods *8*, 861–869. https://doi.org/10.1038/nmeth.1674.

15. Papapetrou, E.P., Lee, G., Malani, N., Setty, M., Riviere, I., Tirunagari, L.M.S., Kadota, K., Roth, S.L., Giardina, P., Viale, A., et al. (2011). Genomic safe harbors permit high β-globin transgene expression in thalassemia induced pluripotent stem cells. Nat. Biotechnol. *29*, 73–78. https://doi.org/10.1038/nbt.1717.

16. Sadelain, M., Papapetrou, E.P., and Bushman, F.D. (2011). Safe harbours for the integration of new DNA in the human genome. Nat. Rev. Cancer *12*, 51–58. https://doi.org/10.1038/nrc3179.

17. Kotin, R.M., Linden, R.M., and Berns, K.I. (1992). Characterization of a preferred site on human chromosome 19q for integration of adeno-associated virus DNA by non-homologous recombination. EMBO J. *11*, 5071–5078. https://doi.org/10.1002/j.1460-2075.1992.tb05614.x.

18. Yada, R.C., Ostrominski, J.W., Tunc, I., Hong, S.G., Zou, J., and Dunbar, C.E. (2017). CRISPR/Cas9-Based safe-harbor gene editing in rhesus iPSCs. Curr. Protoc. Stem Cell Biol. *43*, 5A.11.1–5A.11.14. https://doi.org/10.1002/cpsc.37.

19. Castaño, J., Bueno, C., Jiménez-Delgado, S., Roca-Ho, H., Fraga, M.F., Fernandez, A.F., Nakanishi, M., Torres-Ruiz, R., Rodríguez-Perales, S., and Menéndez, P. (2017). Generation and characterization of a human iPSC cell line expressing inducible Cas9 in the "safe harbor" AAVS1 locus. Stem Cell Res. *21*, 137–140. https://doi.org/10.1016/j.scr.2017.04.011.

20. Smith, S.D., Morgan, R., Gemmell, R., Amylon, M.D., Link, M.P., Linker, C., Hecht, B.K., Warnke, R., Glader, B.E., and Hecht, F. (1988). Clinical and biologic characterization of T-cell neoplasias with rearrangements of chromosome 7 band q34. Blood *71*, 395–402. https://doi.org/10.1182/blood.v71.2.395.395.

21. Ogata, T., Kozuka, T., and Kanda, T. (2003). Identification of an insulator in AAVS1, a preferred region for integration of adeno-associated virus DNA. J. Virol. *77*, 9000–9007. https://doi.org/10.1128/jvi.77.16.9000-9007.2003.

22. Samulski, R.J., Zhu, X., Xiao, X., Brook, J.D., Housman, D.E., Epstein, N., and Hunter, L.A. (1991). Targeted integration of adeno-associated virus (AAV) into human chromosome 19. EMBO J. *10*, 3941–3950. https://doi.org/10.1002/j.1460-2075.1991.tb04964.x.

23. Ordovás, L., Boon, R., Pistoni, M., Chen, Y., Wolfs, E., Guo, W., Sambathkumar, R., Bobis-Wozowicz, S., Helsen, N., Vanhove, J., et al. (2015). Efficient recombinase-mediated cassette exchange in hPSCs to study the hepatocyte lineage reveals AAVS1 locus-mediated transgene inhibition. Stem Cell Rep. *5*, 918–931. https://doi.org/10.1016/j.stemcr.2015.09.004.

24. Bhagwan, J.R., Collins, E., Mosqueira, D., Bakar, M., Johnson, B.B., Thompson, A., Smith, J.G.W., and Denning, C. (2019). Variable expression and silencing of CRISPR-Cas9 targeted transgenes identifies the AAVS1 locus as not an entirely safe harbour. F1000Res. *8*, 1911. https://doi.org/10.12688/f1000research.19894.1.

25. Inderbitzin, A., Kok, Y.L., Jörimann, L., Kelley, A., Neumann, K., Heinzer, D., Cathomen, T., and Metzner, K.J. (2020). HIV-1 promoter is gradually silenced when integrated into BACH2 in Jurkat T-cells. PeerJ *8*, e10321. https://doi.org/10.7717/peerj.10321.

26. Kok, Y.L., Schmutz, S., Inderbitzin, A., Neumann, K., Kelley, A., Jörimann, L., Shilaih, M., Vongrad, V., Kouyos, R.D., Günthard, H.F., et al. (2018). Spontaneous reactivation of latent HIV-1 promoters is linked to the cell cycle as revealed by a genetic-insulators-containing dual-fluorescence HIV-1-based vector. Sci. Rep. *8*, 10204. https://doi.org/10.1038/s41598-018-28161-y.

27. Lebedev, M.O., Yarinich, L.A., Ivankin, A.V., and Pindyurin, A.V. (2019). Generation of barcoded plasmid libraries for massively parallel analysis of chromatin position effects. Genet. Breed. *23*, 203–211. https://doi.org/10.18699/vj19.483.

28. Chen, H.C., Martinez, J.P., Zorita, E., Meyerhans, A., and Filion, G.J. (2017). Position effects influence HIV latency reversal. Nat. Struct. Mol. Biol. *24*, 47–54. https://doi.org/10.1038/nsmb.3328.

29. Davidsson, M., Diaz-Fernandez, P., Schwich, O.D., Torroba, M., Wang, G., and Björklund, T. (2016). A novel process of viral vector barcoding and library preparation enables high-diversity library generation and recombination-free paired-end sequencing. Sci. Rep. *6*, 37563. https://doi.org/10.1038/srep37563.

30. Gotelli, N.J., and Colwell, R.K. (2001). Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. Ecol. Lett. *4*, 379–391. https://doi.org/10.1046/j.1461-0248.2001.00230.x.

31. Sheward, D.J., Murrell, B., and Williamson, C. (2012). Degenerate Primer IDs and the birthday problem. Proc. Natl. Acad. Sci. USA *109*, E1330. author reply E1331. https://doi.org/10.1073/pnas.1203613109.

32. Davidsson, M., Díaz-Fernández, P., Torroba, M., Schwich, O.D., Aldrin-Kirk, P., Quintino, L., Heuer, A., Wang, G., Lundberg, C., and Björklund, T. (2018). Molecular barcoding of viral vectors enables mapping and optimization of mRNA trans-splicing. RNA *24*, 673–687.

33. Cornils, K., Thielecke, L., Hüser, S., Forgber, M., Thomaschewski, M., Kleist, N., Hussein, K., Riecken, K., Volz, T., Gerdes, S., et al. (2014). Multiplexing clonality: combining RGB marking and genetic barcoding. Nucleic Acids Res. *42*, e56. https://doi.org/10.1093/nar/gku081.

34. Roth, T.L., Puig-Saus, C., Yu, R., Shifrut, E., Carnevale, J., Li, P.J., Hiatt, J., Saco, J., Krystofinski, P., Li, H., et al. (2018). Reprogramming human T cell function and specificity with non-viral genome targeting. Nature *559*, 405–409. https://doi.org/10.1038/s41586-018-0326-5.

35. Luo, Y., Liu, C., Cerbini, T., San, H., Lin, Y., Chen, G., Rao, M.S., and Zou, J. (2014). Stable enhanced green fluorescent protein expression after differentiation and transplantation of reporter human induced pluripotent stem cells generated by AAVS1 transcription activator-like effector nucleases. Stem Cell. Transl. Med. *3*, 821–835. https://doi.org/10.5966/sctm.2013-0212.

36. Jordan, A., Defechereux, P., and Verdin, E. (2001). The site of HIV-1 integration in the human genome determines basal transcriptional activity and response to Tat transactivation. EMBO J. *20*, 1726–1738. https://doi.org/10.1093/emboj/20.7.1726.

37. Wagih, O. (2017). ggseqlogo: a versatile R package for drawing sequence logos. Bioinformatics *33*, 3645–3647. https://doi.org/10.1093/bioinformatics/btx469.

38. Schneider, U., Schwenk, H.U., and Bornkamm, G. (1977). Characterization of EBV-genome negative "null" and "T" cell lines derived from children with acute lymphoblastic leukemia and leukemic transformed non-Hodgkin lymphoma. Int. J. Cancer *19*, 621–626. https://doi.org/10.1002/ijc.2910190505.

39. QIAGEN (2020). CLC Genomics Workbench User Manual: Chapter 24: Prepare Sequencing Data (QIAGEN).

40. Chen, H., and Boutros, P.C. (2011). VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R. BMC Bioinf. *12*, 35. https://doi.org/10.1186/1471-2105-12-35.