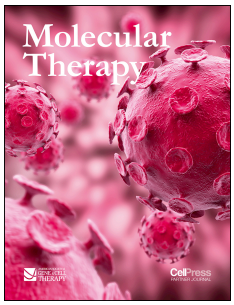
Journal Pre-proof



Efficient nuclease-directed integration of lentivirus vectors into the human ribosomal DNA locus

Diana Schenkwein, Saira Afzal, Alisa Nousiainen, Manfred Schmidt, Seppo Ylä-Herttuala

PII: S1525-0016(20)30253-7

DOI: https://doi.org/10.1016/j.ymthe.2020.05.019

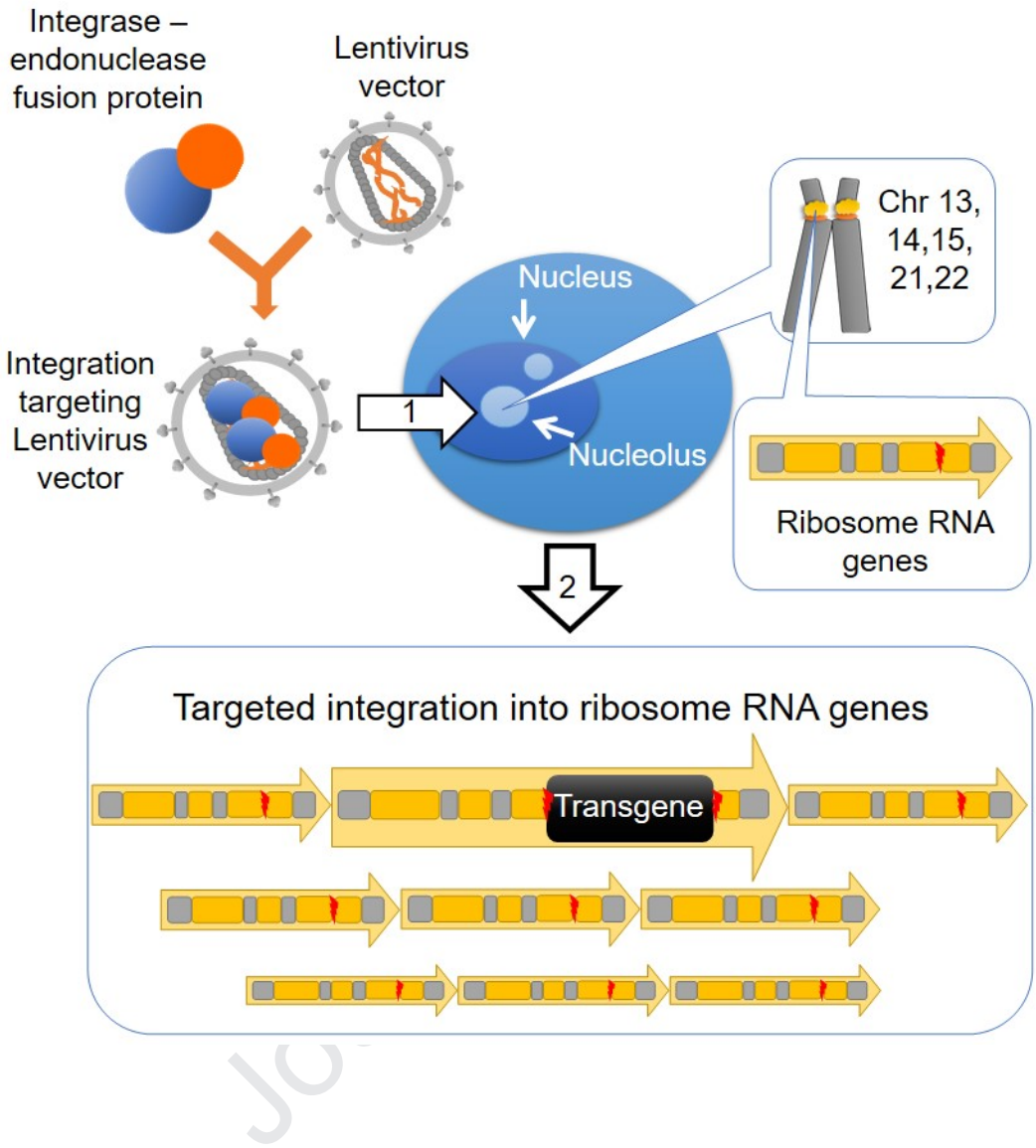
Reference: YMTHE 5171

To appear in: *Molecular Therapy*

Received Date: 21 November 2019

Accepted Date: 19 May 2020

Please cite this article as: Schenkwein D, Afzal S, Nousiainen A, Schmidt M, Ylä-Herttuala S, Efficient nuclease-directed integration of lentivirus vectors into the human ribosomal DNA locus, *Molecular Therapy* (2020), doi: https://doi.org/10.1016/j.ymthe.2020.05.019.


Integrase – endonuclease fusion protein
Lentivirus vector
Integration targeting Lentivirus vector
1
Nucleus
Nucleolus
Chr 13, 14,15, 21,22
Ribosome RNA genes
2
Targeted integration into ribosome RNA genes
Transgene

Efficient nuclease-directed integration of lentivirus vectors into the human ribosomal DNA locus

Diana Schenkwein[1§], Saira Afzal[2§], Alisa Nousiainen[1], Manfred Schmidt[2,3] and Seppo Ylä-Herttuala[1,4*]

[1] A.I. Virtanen Institute for Molecular Sciences, University of Eastern Finland, P.O. Box 1627, FIN-70211 Kuopio, Finland.

[2] Department of Translational Oncology, National Center for Tumor Diseases (NCT) and German Cancer Research Center (DKFZ), Im Neuenheimer Feld 581, 69120, Heidelberg Germany

[3] GeneWerk GmbH, Im Neuenheimer Feld 582, 69120 Heidelberg, Germany.

[4] Heart Center and Gene Therapy Unit, Kuopio University Hospital P.O. Box 1777, FIN-70211 Kuopio, Finland.

* Correspondence should be addressed to S.Y-H. (seppo.ylaherttuala@uef.fi)

[§]Joint first authorship.

Short title: Efficient integration of LVs into rDNA

KEYWORDS

Targeted integration/ Lentivirus vector/ genomic safe harbour site

Seppo Ylä-Herttuala, MD, PhD, FESC

Professor of Molecular Medicine

University of Eastern Finland, A.I. Virtanen Institute for Molecular Sciences

P.O.Box 1627, FI-70211 Kuopio , Finland

E-mail: seppo.ylaherttuala@uef.fi

Mobile phone +358 50 5924067

ABSTRACT

Lentivirus vectors (LVs) are efficient tools for gene transfer, but the nonspecific nature of transgene integration by the viral integration machinery carries an inherent risk for genotoxicity. We modified the integration machinery of LVs and harnessed the cellular DNA double strand break repair machinery to integrate transgenes into ribosomal DNA, a promising genomic safe harbor site for transgenes. LVs carrying modified I-PpoI - derived homing endonuclease proteins were characterized in detail, and we found that at least 21% of all integration sites localized to ribosomal DNA when LV transduction was coupled to target DNA cleavage. In addition to the primary sequence recognized by the endonuclease, integration was also enriched in chromatin domains topologically associated with nucleoli, that contain the targeted ribosome RNA genes. Targeting of this highly repetitive region for integration was not associated with detectable DNA deletions or negative impacts on cell health in transduced primary human T cells. The modified LVs characterized here have an overall lower risk for insertional mutagenesis than regular LVs and can thus improve the safety of gene and cellular therapy.

INTRODUCTION

Human immunodeficiency virus (HIV) 1 -based lentivirus vectors (LVs) are increasingly used in different gene therapy trials ranging from the treatment of monogenic diseases to cell therapy of cancer.[1,2] Despite being less genotoxic than the more frequently used gammaretrovirus vectors[3], LVs – like all integrating gene transfer systems – possess a risk of causing undesired genomic events that can lead to new malignancies. The genotoxicity risks of LVs are mainly related to aberrant transcriptional activation or inactivation of cellular genes and the induction of new splice variants with potentially oncogenic effects.[4]

The HIV-1 integrase protein (IN) catalyzes permanent incorporation of vector-carried transgenes into the chromatin of host cells.[5] It processes the viral long terminal repeats (LTRs), which flank the viral genome, so that a 3' GT dinucleotide is cleaved off. Cellular DNA repair enzymes finish the integration reaction by sealing remaining gaps between the provirus and genomic DNA. Mainly through IN's interaction with its cellular co-factor LEDGF/p75, lentiviruses have a strong preference to integrate within coding sequences of actively transcribed protein-encoding genes.[6,7] Although no severe adverse effects have been described to date that would result from the typical integration pattern of LVs[2], permanent transgene delivery into target cells would optimally take place in a predefined genomic region that could house transgenes with minimal risks for genotoxicity.

Ribosomal DNA (rDNA) consists of highly repetitive ribosome RNA (rRNA) genes, of which there are about 400-600 copies in each cell.[8] rRNA genes are typically organized as tandem repeats that are separated by intergenic spacer (IGS) regions (Figure 1A). Apart from the 5S rRNA that is encoded from a cluster in chromosome 1, the genes encoding for the RNA components of ribosomes reside in the short arms of the acrocentric human chromosomes 13, 14, 15, 21 and 22 that form the nucleoli.[9] Due to the wealth of rRNA genes and the isolated location of nucleolar DNA distant from protein-encoding genes with oncogenic potential, rDNA represents a promising genomic safe harbor for the integration of therapeutic transgenes.

DNA double strand breaks (DSBs) are repaired in cells mainly through two pathways, the non-homologous end joining (NHEJ) and homologous recombination (HR).[10] Small insertions or deletions (indel-mutations) frequently accompany NHEJ-driven DSB repair, but both pathways have been used successfully for genome editing and to integrate donor DNA molecules into specific sites with the aid of different nucleases.[11,12] Most currently available nuclease-based techniques, however, rely on transfection and require using at least two

separate vectors or molecules, which can reduce the efficiency of desired modifications and hampers their *in vivo* use.

We have characterized the full integration site repertoire of LVs that carry an enzymatically weakened homing endonuclease protein that was incorporated into the vectors with the aim of targeting integration to the DSBs it generates. I-PpoI recognizes a 15 bp sequence present in the 28S rRNA genes of eukaryotes (Figure 1A).[13,14] The coupling of LV-transduction with target DNA cleavage enabled an unprecedently high level of transgene integration targeting into rDNA and decreased the genotoxicity risks associated with the use of LVs for gene transfer. These vectors retain the large packaging capacity of LVs and are directly suitable for both *ex vivo* and *in vivo* gene transfer applications.

RESULTS

3rd generation LVs used for targeted integration into ribosomal DNA

In order to generate targeted DSBs into rDNA, we used an IN-I-PpoI$_{H78A}$ fusion protein that binds to and cleaves the 28S rRNA gene, but affects cellular viability less than the wild type endonuclease.[15] Third generation LVs containing the IN-I-PpoI$_{H78A}$ were produced with our previously established method that results in the incorporation of both the IN-fusion protein and the integration deficient IN (IN$_{D6V}$) molecules into vector particles (Figure 1B), which improves their titers and functionality.[16] LVs carrying the IN-I-PpoI$_{H78A}$ protein (hereafter called D+H) were characterized side-by-side with LVs carrying the enzymatically inactivated IN-I-PpoI$_{N119A}$ (D+N)[16,17] to better delineate the effects of target DNA cleavage on vector integration. Unmodified LVs (INwt) were used as a control. All vectors whose complete integrome was analyzed contained an EGFP transgene construct compatible with both LV-catalyzed and NHEJ-driven integration. The proportion of MRC-5 lung fibroblast cells positive for EGFP expression was 83-97% at day two or three post transduction when genomic DNA was extracted for IS analysis (Table S2).

IN-I-PpoI$_{H78A/N119A}$-inclusion changes the global integration pattern and genotoxicity risks of LVs

IS were analyzed separately for the non-repetitive and repetitive portions of the human genome (Hg38). The total numbers of IS retrieved for the different vector types were 20789 for LV-INwt, 7181 for LV-D+H and 2906 for LV-D+N. The proportions of IS that had multiple hits in the genome (MH-IS) of the total data was found to be significantly higher in the IN-modified LVs in comparison to the control LV (Figure 2A). The exactly mappable or unique hit (UH-) IS were used to determine the overall integration pattern for each vector. The chromosomal distribution of IS was similar between the vectors apart from deviations in seven chromosomes (Figure 2B). The distribution of IS within genes was more uniform throughout the coding region for the IN-fusion protein containing LVs than for the INwt LVs, which typically integrate less frequently in the first tenth percentile of a gene's length (Figure 2C).[18] All analyzed LVs favored integration within genes over integration in their upstream regions, but in comparison to INwt LVs, there was a small but statistically significant increase in integration within the first 5kb upstream of genes with the IN-modified LVs. The IN-fusion protein -containing LVs had fewer intragenic IS than INwt LVs (Figure 2D), and hence a smaller risk to interrupt cellular genes with important functions. A vector's tendency to integrate into or close to oncogenes is an important parameter of its safety, and HIV is known to integrate into these areas more than would be expected through chance.[19] Both IN-fusion protein -containing LVs had fewer IS within and near oncogenes in comparison to INwt-LVs (Figure 2E and Table S3). The IN-fusion protein LVs mainly integrated without IN's activity in contrast to INwt LVs, whose LTRs were most frequently processed (Figure S1).

rRNA and tRNA repeats are the most favored targets for the IN-modified LVs within the repetitive genome

The MH-IS were used to characterize the vectors' preferences to integrate within different genomic repeat elements, which were identified using RepeatMasker.[20] I-PpoI has 12 perfect recognition sites in the current genome version (Hg38), and all but two of these localize to rRNA repeat -contained sequences placed either on the acrocentric chromosome 21 or in non-acrocentric chromosomes that contain fragments of rRNA genes (Table S1). For D+H LVs, 41.9% of the vector's MH-reads were within rRNA repeats (Figure 3A). In contrast, D+N LV reads were most frequently associated with transfer RNA (tRNA) genes (17.8%), SINE/Alu-repeats and third most with rRNA repeats. tRNA genes were among the top three repeats also for the D+H LVs. INwt LVs preferred SINE/Alu (40.0%) and LINE/L1 repeats (15.5%) and had very few integrations in either rRNA or

tRNA genes. Interestingly, also signal recognition particle (srp) and other repetitive non-coding RNA (ncRNA) genes were more frequently targeted for integration by the IN-modified LVs than by the control vector (Figure 3A and Figure S2). Based on the differences between the D+H and D+N LVs it is evident that the introduction of DSBs increases vector integration into rRNA repeats.

28S rRNA gene cleavage enables highly efficient integration targeting to rDNA

In addition to nucleolus-associated rDNA, rRNA gene segments are also found in the non-nucleolar genome[21], and a fraction of the uniquely mapping IS reads localized to these sites. The compiled IS data comprising both the unique and multiple hit IS reads was therefore analyzed to determine the absolute numbers of rDNA-localized integrations. For the D+H LVs, 21.3% of all IS localized to sequences contained within an rDNA unit (Figure 3B) and the most favored locus within the rRNA gene was the 28S rRNA (Figure 3C). rDNA-localized IS comprised 2.6% and 0.08% of all IS for the vectors D+N and INwt, respectively (Figure 3B), which is well in line with our previous characterizations of these vectors.[16] Similar to D+H LVs, the majority of D+N LV proviruses clustered into 28S rRNA, but with a much lower frequency (Figure 3C).

To verify the differences between the vectors in catalyzing targeted integration, we used a ddPCR-based method that detects integrated vector genomes within a 235 bp window around the I-PpoI site in the 28S rRNA gene (Figure S3). At day nine post transduction, 20.9% of the D+H LV proviruses were estimated to reside in this locus in transduced MRC-5 cells (Figure 3B; see also Table S4). The proportion of IS reads within the same window was 9.9%. In comparison, for the LVs containing D+N and INwt the proportion of IS reads was 0.8% and 0.02%, respectively, and the ddPCR-based targeting estimates 0.2% and 0.1% (Figure 3B). Integration of the IN-modified LVs occurred more frequently in sense orientation both near the I-PpoI site (66% for D+H and 71% for D+N; Figure 3D) and within it (Figure 3E). Typical for DSB repair through NHEJ, integration into the I-PpoI site involved small indel mutations, which were observed more frequently in the D+H LV -treated than in the D+N LV -transduced cells (Figure S4).

The ddPCR result suggested that for LV D+H the actual level of integration targeting into the immediate vicinity of the I-PpoI site in the rRNA gene is at least two times higher than resolved with the IS sequencing method. Next we used vectors containing a selectable marker for zeocin resistance to test whether the 28S rRNA-insertions remained stable through conditions that require expression of the transgene. The proportion of

proviruses in and near the I-PpoI site remained similar between selected and unselected hTERT-RPE1 cells, as verified with ddPCR (Table S5). Taken together, when LV transduction is coupled with the cleavage of target DNA by a vector-carried endonuclease, stable and highly efficient targeted integration of transgenes into rDNA is achieved.

Integrase-I-PpoI fusion proteins target integration into strong hotspots that are distinct from the areas naturally preferred by HIV-derived LVs

Specific genomic loci have been identified that recur as preferential integration loci, or integration hotspots, for HIV-1 and lentivirus vectors.[22,23] Such common integration sites (CIS) were identified to see if the inclusion of the IN-I-PpoI-fusion proteins altered the natural preferences of LVs. Significant CIS containing at least three IS were characterized for their genomic coordinates and for the features they contained. In comparison to the IN-modified LVs, a larger proportion of INwt LV's unique IS were engaged with integration hotspots, but proportionally fewer IS formed the strongest CIS (Figure S5, File S1). The majority of the 15 strongest CIS (n=18 individual CIS) of the LV INwt were localized within protein-encoding genes (77.8%) (Table 1) with many of the hotspots residing in regions previously characterized as preferred integration sites for LVs and HIV-1 (Tables S6 and S7).[22–26] The median CIS positions (CIS foci) of the seven strongest hotspots of the D+H LVs (n=26) were frequently found in intergenic loci (35%), and in many cases the RefSeq-gene within the hotpot or nearest to it was a ncRNA gene (31%) (Table 1 and Figure S6A). Altogether six D+H LV CIS foci were within an rRNA repeat and five of them localized to I-PpoI cleavage sites on separate non-acrocentric chromosomes (Table 1 and File S1), verifying correct I-PpoI activity and NHEJ-driven insertion at the generated DSBs. The five strongest CIS foci (n=21 individual CIS) of the D+N LVs revealed a similar preference towards intergenic areas and ncRNA gene proximity as was seen for D+H LVs, but instead of rRNA gene repeats, the hotspots frequently associated with tRNA repeats (29%) (Table 1; Figure S6B). Altogether 9.5% of all D+N LV's unique CIS-associated IS were within tRNA repeats, whereas neither tRNA nor rRNA repeats were found in the hotspot-contained IS of the INwt LVs (n=8450) (Figure S6B). Analysis of all CIS-associated UH-IS confirmed that both IN-modified LVs had significantly more intergenic IS than the control vector (Figure 4A). INwt-LVs' CIS-associated IS localized into or near protein-encoding genes more frequently than those of D+H LVs, and the latter targeted RNA genes more often than the control vector. Genes and pseudogenes of the

large and small ribosome subunit proteins (RPL or RPS, respectively) were also frequently associated with the CIS of the D+H LVs (Table 1).

The repeat-associated IS make up at least one third of the total IS number in the IN-fusion protein LVs, and a more accurate representation of genomic features and gene types preferentially targeted for integration by these vectors could be obtained by analyzing CIS in a combined data set containing both the UH and the MH IS. In this analysis, the D+H LVs' strongest CIS was now identified in the 28S rRNA gene and it contained 19% (n=1367 IS) of all IS (Table 2 and Figure S7A). The strongest CIS of the D+N vectors also localized into the 28S rRNA gene with 2.5% of all IS. Integration targeting to the most preferred locus was again the weakest for LV INwt, as only 0.3% (n=68 IS) of the vector's IS localized to the strongest CIS (Table 2 and Figure S7A). Inclusion of the MH data into the CIS analysis enabled the detection of new repetitive gene types, such as 5S rRNA and srpRNA genes, in the integration hotspots of the IN-modified LVs (Table 2). The characteristic preferences of these LVs to integrate into tRNA and rRNA repeats and intergenic loci remained the same but became more pronounced (Table 2 and Figure S7B). Similarly, the differences between the IN-modified LVs and the control LV in targeting protein-encoding genes, RNA genes and the multiple ribosome subunit genes grew stronger (Figure 4B). Finally, a clear increase in the IS numbers per strongest CIS was observed owing to the large proportion of MH-IS forming them (Table 2). For the INwt LV the differences between the two analysis types were much subtler and mainly related to slightly higher IS numbers per identified CIS (Tables 1 and 2). Taken together, the integration hotspots of the IN-modified LVs strongly associate with repetitive RNA-encoding genes and show very little resemblance to the well-characterized hotspots near protein-encoding genes of unmodified LVs.

I-PpoI protein inclusion increases vector integration in genomic features that are enriched in nucleolus associated domains

Nucleolus associated domains (NADs) are defined chromatin domains that dynamically interact with nucleoli.[27] Enrichment of pseudogenes in NADs has been characterized in plants[28] and the ribosomal protein encoding genes are known to have multiple processed pseudogenes in the human genome. Also specific gene families and genes, such as those encoding for tRNAs and the protein constituents of the ribosomes, are enriched in NADs.[29–33] Since these gene types were frequently hit by the IN-modified LVs (Figure 3A and 4B) and identified in their

integration hotspots (Tables 1 and 2 and Figures S6 and S7), we asked whether additional similarities would exist between the identified CIS-loci and NAD-contained regions. After annotating the IS of the different LVs with pseudogenes, we found that integration in pseudogenes occurred more frequently with the IN-modified LVs than with the control LV (Figure 5A). When the pseudogene-annotations were used in place of the original Refseq gene annotations, integration was found to be more frequent also in RPL and RPS gene -derived sequences with the IN-modified LVs than with the INwt LVs (Figure 5A). In addition to these structural proteins of the ribosomes, also larger groups of genes related to ribosome biogenesis contained more integrations with the IN-modified LVs than with the control LV (Figure 5B).

Significantly enriched gene ontology (GO) terms among NAD-genes include ribosome, mitochondrion, cytosolic large/small ribosomal subunit and nucleolus.[29] A GO-analysis of the CIS-engaged genes revealed that several pathways and processes related to ribosome structure and function were enriched among the genes preferentially targeted for integration by the IN-fusion protein LVs, and that similar GO-terms were enriched as among NAD-associated genes (Figure 5C-D and File S2). Interestingly, also mitochondria-related terms were enriched for D+N LVs but not for D+H LVs. For the INwt LV no enrichment of ribosomal structure or function -related terms was observed (Figure 5E). In line with previous studies[34], the most enriched pathways and processes were instead related to cell cycle and its control as well as chromatin organization. The similarities between NAD-associated features and the gene types preferentially targeted for integration by the IN-fusion protein LVs indicates that the localization of a chromosomal region close to nucleoli is an additional determinant of the vectors' preferential integration, in addition to the primary sequence recognized by I-PpoI.

Integration targeting and cellular responses to transduction in primary human T cells

Having confirmed rDNA-targeted integration in both the slowly and finitely dividing lung fibroblast cells (MRC-5) and in the non-cancerous but immortalized retinal pigment epithelium cells (hTERT-RPE1), we asked how the IN-modified vectors would perform in the transduction of primary human T cells, which represent a relevant cell type for clinical gene and cell therapy. For this aim, T cells from two individuals were enriched, transduced with the different LVs and assayed for targeted integration and different indicators of cell health and cytotoxicity. Estimation of targeted integration at day 10 post transduction with the ddPCR-based method showed that up to 8% of the D+H LV's integration events reside in the immediate vicinity of the I-PpoI site in

the 28S rRNA gene, the mean targeting efficiencies ranging from 2.6% to 5.7% (Figure 6 A and B; Tables S8 [day 2] and S9 [day10]). With the INwt control LVs the mean targeting efficiencies were 0.0-0.1%.

The number of metabolically active live cells was determined to study if T cells transduced with the D+H -containing LVs proliferate similarly to cells transduced with the control LV. In a test using 5000 vector partices (5k vp) per cell, the number of viable cells was the highest in the INwt LV group, and no differences between the groups were observed that could be specifically addressed to the IN-content of the modified LVs (Figure S8 A and B). When using a higher vector dose of 10k vp/cell, the only test group having significantly fewer metabolically active cells in comparison to the INwt control at the last time point assayed was the D+H LV group, whose mean cell numbers were 81-85% of those of the control vector's (Figure S8 C and D).

Next it was studied whether the cleavage of rRNA genes and subsequent transgene integration would cause direct cytotoxicity or induce apoptosis that is followed by secondary necrosis. Of the three LVs tested, a statistically significant increase in the apoptosis signal in relation to untreated cells was observed only for LV D+N at day three post transduction (5k vp/cell, $p<0.05$) (Figure S9). An elevated necrosis signal was observed for INwt LVs in altogether three time points ($p<0.05$; $p<0.01$ and $p<0.001$), and for D+H LV at one time point ($p<0.05$) in comparison to non-transduced cells (Figure S10). Etoposide-treated cells were positive for apoptosis induction at day one and for necrosis at days two and three post treatment (Figures S9 and S10). Since there was no increase of necrosis in T cells that would be clearly attributable to the D+H content of the vectors, it is likely that the decrease in cell numbers we observed in the viability test results from a moderate slowdown of division and/or metabolism in LV D+H -transduced cells.

As learned from studies using the Cas-nucleases, target DNA cleavage can cause different types of mutations and rearrangements of genomic DNA, including large deletions.[35–38] rDNA represents a recombination hotspot in meiotic cells and in cancer, and hence the number of rRNA genes can vary substantially both between and within individuals.[8,39–42] To see if the number of rRNA genes would be affected by the use of D+H LVs, we quantitated the 18S rRNA gene copies in transduced T cells at day two post transduction. Consistent with previous studies[8], the mean gene copy numbers or rRNA genes varied between 478-701 per cell, and no statistically significant differences were observed between the non-transduced cells and D+H or INwt LV -transduced cells (Figure 6 C and Table S10). To address the occurrence of larger deletions potentially affecting whole acrocentric chromosome arms, we studied the copy number of the distal junction (DJ) sequence that

flanks the rRNA array at the telomeric side.[43] Similar to the rRNA genes, no statistically significant differences were observed between the three groups, and 13 to 18 copies of these sequences were detected per cell (Figure 6 D). In conclusion, transduction with the 28S rRNA gene -cleaving D+H LVs does not cause detectable variations in the rRNA gene nor in the DJ sequence copy numbers in T cells.

Cleavage of the rRNA gene and transgene integration into it can affect the transcription of both the rDNA and the provirus. To address the question of whether vectors integrated into the I-PpoI site become transcribed, we analyzed total RNA extracted from D+H and INwt LV -transduced T cells at days two and 10 post transduction with site-specific RT-ddPCR. Vector sequence -containing rRNA transcripts were detected at both time points and only in the D+H LV group, confirming that proviruses within the targeted 28S rRNA gene become transcribed (Tables S11 and S12).

DISCUSSION

In this study we show that LV integration can be directed to the rDNA of normal human cells with an unprecedently high efficiency when transduction is coupled with target site cleavage. In non-selected MRC-5 cells, the vectors carrying an endonuclease with reduced DNA cleaving activity integrated 266 times more frequently into rDNA than the control vectors, and 8.2 times more than LVs whose IN-endonuclease content can only bind the target DNA. Other researchers have attempted to direct the integration of recombinant adeno-associated virus vectors (rAAVs) to the same locus, but achieved only modest efficiencies: the increase in targeted integration was 8–13-fold in comparison to control vectors[44], and 2-3% of selected hepatocytes were estimated to have the intended integration event within the 28S rRNA gene.[45] The LVs characterized in our study promote much higher rDNA-targeting, but further comparisons with the rAAVs are challenging due to profound differences in the study designs, IS analysis methods and in the numbers of IS retrieved (n=12-176 for the rAAVs).[44,45] In addition to rAAVs, also non-viral vectors have been developed to target integration into the rDNA genomic safe harbor locus.[46,47] However, in these studies the levels of both transfection and targeted integration were low and the analysis lacked thorough examination of the potential off-target integration events.

Our primary focus was to characterize both the complete integrome and the integration targeting efficiency of two IN-modified LVs as comprehensively as possible, which was achieved through the analysis of all IS at an early time point where minimal clonal expansion of transduced cells had occurred. Analysis of LV D+H

transduced MRC-5 cells at later time points with ddPCR revealed that the efficiency of integration targeting into the 28S rRNA gene is at least two times higher than resolved through IS sequencing, reaching 21% of all proviruses. When comparing unselected and Zeocin-selected hTERT-RPE1 cells, we found that the proportion of proviruses integrated within the 28S rRNA gene remains stable in this repetitive DNA locus. Transduction tests with primary human T cells confirmed that integration within the 28S rRNA gene is increased also in this clinically relevant cell type, albeit to a lower degree than observed in the MRC-5 cells.

Subsampling and partitioning errors are known sources for variability in ddPCR, and its precision is decreased at the extremes.[48,49] Other factors that can have contributed to the observed differences between the tested cell types include inherent differences in their replication kinetics and susceptibilities to transduction with LVs, lot-to-lot variability between the produced LVs and a limited number of replicates analyzed per sample. On the other hand, with the IS sequencing method the number of unique integrations within a highly targeted locus is easily underestimated due to saturation of potential unique MuA transposition sites and read lengths that were used to differentiate individual integrations from PCR-borne replicates. Despite the differences in efficiencies that likely originated from subsampling-related issues, the ddPCR-based method clearly demonstrated that D+H LVs catalyze targeted integration in both primary and cultured cells.

Cleavage of the 28S rRNA gene, its subsequent repair and simultaneous insertion of proviruses into it could cause genomic rearrangements in this highly repetitive locus, including large deletions. We tested for this possibility and found no signs of gross deletions in the acrocentric chromosomes or in the rRNA genes after transduction with the D+H LVs. A moderate reduction in viable cell numbers was observed in LV D+H -transduced T cells at day four after transduction, but no clear indications of cytotoxicity were evident. Ribosome RNA gene transcription is halted upon DSB introduction into rDNA, which causes the formation of specific nucleolar cap structures and facilitates repair of the lesions (reviewed in[50]). The observed reduction in the numbers of metabolically active cells may hence have resulted from the decreased production of the building blocks for ribosomes, which directly affects the metabolic activity of the cell. At days two and ten post transduction, we were able to detect provirus-containing transcripts from the 28S rRNA gene, which proves that transcription of this locus and the genetic material inserted into it is recommenced after DSB repair.

By analyzing the complete integrome of the modified LVs in MRC-5 cells we found that proviruses residing outside of the targeted rDNA locus had a lower tendency to integrate within genes and oncogenes, but showed a

higher preference towards genomic features that are also enriched in NADs, chromatin domains that co-localize with rRNA gene arrays in the three-dimensional organization of the genome.[24–28] One explanation for the preferential targeting to these loci could be that nicks or DSBs occurring randomly in NAD-containing chromosomes capture a proportion of vector genomes that were tethered to nucleolar proximity by the LV-contained I-PpoI protein. For the D+N LVs the localization of genomic regions in NADs seems to be a stronger determinant of integration hotspot site selection than the distance to an I-PpoI site. The transcriptional status of transgenes inserted into NADs and further verification of this phenomenon remain to be addressed with additional techniques in the future. To our knowledge this is this first description of distinct genomic regions, that are distant from another on the linear axis of DNA but near in the three-dimensional genome, to become jointly affected when site-specific transgene integration was pursued based on primary DNA sequence recognition. This observation may have utility in the prediction of possible off-target sites also when using other nucleases for genome editing, such as the CRISPR/Cas system.

The most desired integrating vectors in gene therapy are those that can direct transgenes into genomic safe harbor sites to minimize the risks related to insertional mutagenesis. LVs have many benefits as vectors, but their integration profile may endanger normal cellular gene function. First attempts to direct LV integration to specific sites were based on IN-fusion proteins[52], and more recent approaches relied on new chromatin binding preferences assigned for the IN-tethering LEDGF proteins.[53–56] After our first report of using LVs for protein transduction without the previously necessary Vpr-protein fusions[57], many studies have described different LV- or retrovirus vector (RV) -based virus like particles, or nanoparticles, to transport desired proteins into cells often with the aim of delivering DNA editing or integration targeting enzymes.[58–67] In addition, LVs and RVs can deliver these components into cells as transgenes (reviewed in[68]) or messenger RNA.[69–71] Systems in which single vector particles contain both the donor DNA and the enzymes required for targeted integration are superior to multi-construct approaches, that may suffer from decreased efficiency if only a fraction of the intended components reach target cells. The majority of recent studies aiming for genome editing and targeted integration utilize the CRISPR/Cas-system. With the help of different technical advances and the discovery of alternative Cas-variants it has been possible to improve the specificity of targeted genome modifications (reviewed in[72]), but major concerns related to the safety[35–38] and efficacy of the CRISPR-based approaches remain, precluding their wide utility in the clinic at the moment.

In comparison to most genomic safe harbor (GSH) site candidates, rDNA is unique owing to its repetitive gene context. This feature could pose challenges to both the cells upon transgene integration, and to the stability of the transgene itself, but our results in primary human T cells did not support such concerns nor point to major adverse effects. The most important safety features of rDNA as a GSH include its isolated location from potentially oncogenic protein-encoding genes, and the high number of rRNA genes that remain intact despite transgene integration into the locus. rDNA is typically ruled by RNA polymerase I, but it is also accessible to the RNA polymerase II machinery.[73–76] We show that integration can be targeted to the rRNA gene array with an unprecedented efficiency using modified LVs that carry both the donor DNA molecules and the integration targeting enzyme within single vector particles. These LVs can deliver large transgenes, are easy to produce with minor modifications to standard protocols and are suitable for both *ex vivo* and *in vivo* gene transfer applications, hence potentially advancing the development next generation applications to treat human diseases.

MATERIALS AND METHODS

Generation of third generation lentivirus vectors.

Vesicular stomatitis virus G glycoprotein (VSV-G) pseudotyped third-generation HIV-1-based lentivirus vectors (LV) containing the IN-fusion proteins were produced as described earlier.[15,16,57,77] Briefly, monolayers of 293T cells were transfected with the production plasmids using calcium phosphate transfection. The plasmids used were pRSV-Rev (encoding for HIV-1 Rev), pCMV-VSVG (encoding for VSV-G), pLV1 (vector construct that contains a PGK promoter -driven EGFP transgene) or pLV1-ZeoR (vector construct carrying a PGK promoter -driven *Sh ble* gene), and either one or two of the packaging plasmids encoding for the wild type integrase (pMDLg/pRRE), the integration deficient integrase (pMDLg/pRRE-IN$_{D64V}$), the IN-fusion protein with DNA cleavage -disabled I-PpoI (pMDLg/pRRE-IN-I-PpoI$_{N119A}$) or the IN-fusion protein with DNA cleavage -proficient I-PpoI that carries an activity-reducing mutation (pMDLg/pRRE-IN-I-PpoI$_{H78A}$). Culture supernatants were collected 48 hr after transfection, filtered, suspended in phosphate-buffered saline (PBS) and stored at -70°C until use. Functional vector titers (transducing units [TU]/ml) were estimated through EGFP expression in transduced HeLa cells approximately 68 hr post transduction and particle titers were determined based on the level of HIV-1 p24 capsid (CA) antigen using an enzyme-linked immunosorbent assay (PerkinElmer Life and Analytical Sciences, Waltham, MA).

Cells, transductions and cell health assays

All transductions were carried out by diluting the LVs into cell culture medium immediately before use, or alternatively by pipetting undiluted LVs directly into cell culture medium. On the day after transduction, vector-containing medium was replaced with fresh medium. All cells were incubated at 37°C in a 5% CO2-containing humidified atmosphere.

For the IS sequencing experiment, human MRC-5 lung fibroblasts (ATCC® CCL-171™) were used. The cells were cultured in Dulbecco's modified Eagle's medium (DMEM; high-glucose, Sigma D6429) supplemented with 1% Penicillin–Streptomycin (Sigma, P0781), 1 % MEM Non-essential amino acids (biowest, Cat. X0557-100), 1% Sodium pyruvate (biowest Cat. L0642-100) and 10% Fetal Bovine Serum (FBS; Sigma, F7524). On the day before transduction MRC-5 cells were seeded onto 6-well plates at a density of 2x10e5 cells per well. An MOI of 4 was used for transduction with the IN-modified LVs (56k-120k vp/cell) and an MOI 1 for transduction with the INwt LV (1k vp/cell). Cells were pelleted at days two and three post transduction and stored at -70°C until used for DNA extraction and integration site analysis. To study the proportion of IS occurring near the I-PpoI site with ddPCR, MRC-5 cells were seeded as above and transduced in two separate experiments with the EGFP-LVs using 7.5K vp per cell, that equaled MOI 19 for LV INwt. Cells were collected for analysis at day 9 post-transduction.

For the study of targeted integration in unselected and phleomycin D1 selected cells, hTERT-RPE1 cells (ATCC® CRL-4000™) were used. Cells were cultivated in 1 X DMEM/F-12 (Gibco, 31330-038) supplemented with 10% FBS and 0.01 mg/ml of hygromycin B. On the day before transduction the cells were seeded onto 6-well plates at a density of 4x10e5 cells per well. Transduction was carried out with the *Sh ble* antibiotic resistance gene containing vectors (ZeoR LVs) at a concentration of 5K vp/cell. At day one post transduction, cells to undergo selection were given culture medium supplemented with Zeocin™ (Invivogen, ant-zn-05) at a final concentration of 300µg/ml and thereafter subcultivated as necessary. Cell pellets were collected for DNA extraction at days 13 and 15 post-transduction and stored at -70°C until use.

Peripheral blood mononuclear cells (PBMCs) were enriched from two leukoreduction system (LRS) chambers (Finnish Red Cross Blood Service, Helsinki, Finland) using the prefilled Leucosep™ centrifuge tubes (Greiner Bio-One, #227288). Untouched human T cells were isolated from the PBMCs by using the Pan T Cell Isolation Kit (Miltenyi Biotech, #130-096-535Y). 2.5x10e7 T cells from both donors were activated with Dynabeads™

Human T-Activator CD3/CD28 (Gibco, #11132D) according to the kit protocol. T cells were cultivated in X-Vivo™ 15 (Lonza, #BE02-060F) supplemented with 5% Human AB Serum (Biowest, #S4190) and 20 U/ml of human recombinant IL-2 (Prospec-Tany Technogene Ltd, #CYT-209-b) for 4 days before LV transductions. All transductions were done in triplicate for T cells of both donors using the ZeoR LVs at vector doses of 5k and 10k vp per cell, which equaled MOIs of 5 and 10 of LV INwt-EGFPs, respectively. Cells to be studied for targeted integration with ddPCR were transduced on 24 well plates (1,5x10e6 cells per well) and sampled for analysis at days 2 and 10 post transduction. For the cells analyzed for viability, apoptosis and necrosis, the activation beads were removed and then the cells were seeded on white 96 well plates with clear bottoms (PerkinElmer, View-Plate®-96-TC, #6005181) at densities of 6000 cells per well for the viability assay and 10 000 cells per well for the apoptosis/necrosis assay. After vector removal at day one post transduction, the cells were given fresh medium and the assay reagents according to kit protocols. Etoposide (Cayman Chemical Company, #12092) was used as a positive control for apoptosis induction and necrosis at a final concentration of 8µM. The viability of transduced cells was monitored with daily luminescence recording for four days (days 1, 2 and 4 post transduction) using the RealTime-Glo™ MT Cell Viability Assay (Promega, # G9711). Apoptosis and necrosis were examined with the RealTime-Glo™ Annexin V Apoptosis and Necrosis Assay (Promega, #JA1011) that simultaneously measures annexin V exposure and DNA release to differentiate secondary necrosis occurring during late apoptosis from necrosis caused by other cytotoxic events. Annexin V binding (luminescence) and loss of membrane integrity (fluorescence) were recorded at days 1, 2 and 3 post transduction.

Integration site extraction and EGFP expression analysis.

MRC-5 cells were transduced with an MOI of one for the control vector (LV INwt) and four for the IN-modified LVs (Table S2). Separate wells were transduced for genomic DNA extraction and for FACS-analysis of EGFP expression. Genomic DNA was extracted two or three days post transduction using the NucleoSpin Tissue kit (Macherey-Nagel, ref:740952.250) from two separate wells per vector. Vector IS were extracted with the MuA transposon -based protocol described in Brady et al, 2011[78], using BtsαI for genomic DNA digestion (NEB #R0667S) and primers and linkers listed in Supplemental Methods. Primers and oligonucleotides used in the study were ordered from Integrated DNA Technologies and the MuA transposon used was from Thermo Scientific (F-750, lot# 00383099). Digested DNA was purified before the MuA reactions using Speedbead Magnetic Carboxylate Modified Particles (GE Healthcare, Part no. 65152105050250). Each of the two individual genomic DNA extractions analyzed per vector were tagged with unique sequence identifiers in both

the linker oligo and in the primer (molecular identifier, MID) to minimize sequence carry-over between samples and to maximize the resolution of integration sites occurring near each other (Table S2). Amplification of the integration sites was carried out using Phusion Flash PCR Master Mix (Thermo Scientific, F-548) in two rounds of PCR. In the first PCR, 2 µl of the MuA reaction was used as template. The first PCR program was as follows: 98°C for 10s, 7 cycles of 98°C for 1s and 72°C for 15s, 37 cycles of 98°C for 1s, 57°C for 5s and 72°C for 15s, with a final extension at 72°C for 1 min. The amplicons from the first round of PCR were diluted 1:50 with nuclease-free water, and 1 µl of the dilution was used as template for the second round of PCR. The second PCR program was as follows: 98°C for 10s, 7 cycles of 98°C for 1s, 67°C for 5s and 72°C for 15s, 37 cycles of 98°C for 1s and 72°C for 15s, with a final extension at 72°C for 1 min. The amplicons were sequenced in Biocenter Oulu Sequencing Center with an IonTorrent PGM instrument (University of Oulu, Finland). EGFP expression was analyzed with flow cytometry from triplicate wells per vector at the day of gDNA extraction from cells fixed with 4% paraformaldehyde in PBS.

ddPCR

The primers, assays, materials and PCR programs used in the different ddPCR (Bio-Rad) reactions are listed in Supplemental Methods. DdPCR was carried out according to Bio-Rad's recommended protocol. For the study of integration in the immediate vicinity of the I-PpoI site in MRC-5 cells, genomic DNA was extracted for analysis from cells collected at day 9 post transduction using QIAGen's DNeasy Blood & Tissue Kit (ref. 69506) and digested with BsuRI (ThermoFisher, ref. ER0151) at a concentration of 1 unit/1 µg DNA. Digested genomic DNA was used as template in ddPCR to measure the copy numbers of all vector genomes, episomal vector forms, production plasmid carryover, and integration near the I-PpoI recognition site in the 28S rRNA gene in both sense and antisense orientation.

For the ddPCR analysis of targeted integration in Zeocin™ selected cells, genomic DNA was extracted from hTERT-RPE1 cells pelleted at day 13 (unselected) and 15 (selected) post-transduction and processed for ddPCR as described above. DdPCR analysis consisted of assays measuring the copy numbers of all vector genomes, episomal vector forms and vectors integrated in sense orientation near the I-PpoI recognition site in the 28S rRNA gene.

For the detection of targeted integration in primary human $CD3^+$ T cells, genomic DNA was extracted from cells pelleted at days two and 10 post transduction using the AllPrep DNA/RNA Mini Kit (Qiagen, #80204). DNA was processed and analyzed with ddPCR as described for MRC-5 cells above. DdPCR was carried out for two replicate wells of non-transduced cells, INwt transduced cells and D+H transduced cells. Each well's DNA was sampled twice for ddPCR.

Analysis of transgene transcription from the 28S rRNA locus at days two and 10 post transduction was carried out with RT-ddPCR using total RNA extracted from T cells with the AllPrep DNA/RNA Mini Kit (Qiagen, #80204) and the protocol established for the detection of targeted integration. One microgram of RNA was treated with DNase I (ThermoScientific ref. EN0521) and cDNA synthesis was carried out with RevertAid RT Reverse Transcription Kit (ThermoScientific, ref. K1691) with random hexamer primers according to the kit's protocol. Depending on the assay, 0.5-2.0µl of the RT reaction was used as template for RT-ddPCR.

The presence of deletions in the rRNA gene array and in the acrocentric chromosome arms was assayed with ddPCR using genomic DNA extracted from T cells transduced with 10k vp/cell and extracted at day 2 post transduction. Probes binding to the distal junction (DJ) region, that flanks the rRNA gene array on the telomeric side[43] , and to the 18S rRNA gene were designed and used for the quantification of the respective areas.

Bioinformatics data analysis

Integration site analysis. Single end FASTQ data files were quality filtered and trimmed by Skewer.[79] The reads were processed to check for the presence of the linker cassette (LC) sequence that was specific for each sample, and for the transposon-linker sequence. After trimming of LC sequences the set of reads was aligned with vector sequence by BLAT[80] aligner to subtract potential vector only -reads and to avoid any false positive vector reads detection. The reads were then mapped with the LV 3'LTR sequence using a minimum identity threshold of 95%. The LTR mapped part was trimmed and the rest of the read region was mapped with human genome reference hg38 with minimum identity of 95%. The reads that mapped uniquely or at multiple sites within the genome were separated in the subsequent steps. A threshold of 90% was employed between the ratio of the BLAT score for primary and secondary mapped reads so that reads with a score ratio greater than this were designated as multiple hit (MH) integration sites (IS) and others as unique hit (UH) IS. To simplify analysis of integration within rDNA, the reads mapping to Chr 21 that had exactly same primary and secondary mapping scores were preferred for their alignment positions in the region between Chr21:8433222-8446572. Exact

sequence duplicates were removed, and reads were filtered using multiple criteria in order to filter out potential duplicates of a single original integration event. Filtering involved restricting the number of non-mapping base pairs before the start of the genomic region (i.e., between LTR and the region mapping to the genome) using a threshold of 4 bp: the reads that had non-mapping base pairs less than or equal to this threshold were further processed to next steps. Next, only reads that had three or fewer base pairs of non-mapping nucleotides at their 3' end were considered. The reads were compared to one another and only those reads that had a difference in the number of deleted base pairs at their LTR ends of >=2, and whose IS and "shear sites" (transposition sites) were at least 3 bp apart from other reads were further processed. The collision sequences among samples were subtracted from each sample and the final reads were mapped against the pLV1 plasmid sequence to remove remaining artifacts. Finally, the genomic positions were annotated according to the RefSeq from UCSC[81] and the RepeatMasker rmblast web version[20] was used to annotate repeat regions. To identify integration into pseudogenes, IS were also annotated with the retro genes -table (Retroposed Genes V9, Including Pseudogenes) obtained from UCSC. Additionally, the oncogenes table (v4 May 2018) was retrieved (http://www.bushmanlab.org/links/genelists) and final set of genes obtained from clustered result files were annotated with this set. The plots shown in Figure 3 were generated for rRNA reads by creating bed and bedgraph files using bedtools[82], that were processed by in-house script and R packages (karyoploteR and regioneR).[83,84]

Analysis of the integration frequency in selected gene sets. Integration frequency in gene sets involved in the SuperPaths[85] of ribosome biogenesis in eukaryotes and rRNA processing in the nucleus and cytosol were conducted using single genes (each IS-tagged gene represented once in the gene list comparison) using the IS data sets where pseudogene-annotations were used in place of the initial RefSeq gene -annotation.

Analysis of common integration sites (integration hotspot analysis): Common integration site (CIS) analysis was performed using a graph-based framework for CIS identification[86,87] with a threshold of 50kb between individual IS. For the analysis of hotspots only CIS with a p-value of less than 0.05 and with a minimum of three IS were accepted. The CIS analysis was performed separated for the IS data sets containing only uniquely mappable IS (UH-IS data set) and for the complete IS data sets (UH and MH IS data). The features in the median CIS-positions in Tables 1 and 2 were annotated using the RepeatMasker, RefSeq-gene and RetrogenesV9 tracks of the UCSC Genome Browser.

Gene ontology analysis of the CIS-associated IS: Analysis of the most overrepresented pathways and processes among genes present in the CIS-engaged IS was performed using Metascape[88] (http://metascape.org/gp/index.html#/main/step1) that uses the following ontology sources: KEGG Pathway, GO Biological Processes, Reactome Gene Sets, Canonical Pathways and CORUM. In the analysis all genes in the genome are used as the enrichment background and terms with a p-value < 0.01, a minimum count of 3, and an enrichment factor > 1.5 are collected and grouped into clusters based on their membership similarities. Each cluster is represented with the most statistically significant term within that cluster. The analyzed gene lists contained all genes (both hit genes and nearest genes) from the identified CIS using the complete IS data (UH and MH IS).

Comparison of "recurrent integration gene" (RIG) loci with the CIS foci of INwt LVs: The genomic coordinates from RIG and "Hotter zone" (HZ) loci listed by Marini and others[22] were converted to the current genome version (Dec. 2013 (GRGh38/hg38)) assembly using the "LiftOver" tool from the University of California Santa Cruz (UCSC) Genome Browser Database.[89] The average positions of the RIGs/ HZs and the INwt LV CIS were compared, and the RIGs and CIS foci that fell within a 100kb distance from one another were listed in Table S7.

Statistics

Statistical differences in the integration preferences between LV groups were calculated using two-sided Fisher's Exact test and with two-sided Chi-square test. Statistical comparisons between groups in the viability and necrosis assays were done with Repeated Measures analysis of variance (ANOVA) followed by the Bonferroni post-test to compare replicate means by row to the control. In the apoptosis assay each time point was analyzed separately with one-way ANOVA followed by Dunnett's multiple comparison test. The differences in copy numbers of 18S and DJ sequences were analyzed with one-way ANOVA by comparing the vector-groups' values to the same donor's NTD control with Dunnett's Multiple Comparison Test. All statistical analysis was done with GraphPad Prism version 5.03 for Windows, GraphPad Software, San Diego California USA, www.graphpad.com.

Data availability

The final IS datasets generated and analyzed in this study are available upon a reasonable request.

ACKNOWLEDGEMENTS

This work was supported by the Finnish Academy Centre of Excellence [307402], The European research Council [GA670951] and by the Eemil Aaltonen Foundation (to D.S.). This work also got support from the National Virus Vector Laboratory/ A.I.Virtanen Institute, University of Eastern Finland, Kuopio and from the Kuopio Center for Gene and Cell Therapy (KCT). Anssi Kailaanmäki, Elina Koli, Annu Luostarinen and Tanja Kaartinen are acknowledged for their help with T cell extraction and culture -related methods.

AUTHOR CONTRIBUTION

D.S. conceived the study, designed the experiments, conducted cell culture experiments, performed bioinformatics analysis, analysed the data and wrote the manuscript. S.A. designed bioinformatics analysis strategy, performed the data analyses and contributed to writing the analysis method section. A.N. conducted cell culture experiments, performed the integration site extractions, designed and executed the ddPCR assays and performed the analysis, analysed the data and participated in writing the manuscript. M.S. and S.Y-H. supervised and financed the study and edited the manuscript.

CONFLICT OF INTEREST STATEMENT

M.S. is co-founder and CEO of GeneWerk GmbH, Heidelberg, Germany.

REFERENCES

1. Cavazzana, M, Bushman, FD, Miccio, A, André-Schmutz, I and Six, E (2019). Gene therapy targeting haematopoietic stem cells for inherited diseases: progress and challenges. *Nat. Rev. Drug Discov.* **18**: 447–462.

2. Milone, MC and O'Doherty, U (2018). Clinical use of lentiviral vectors. *Leukemia* **32**: 1529–1541.

3. Montini, E, Cesana, D, Schmidt, M, Sanvito, F, Ponzoni, M, Bartholomae, C, *et al.* (2006).

Hematopoietic stem cell gene transfer in a tumor-prone mouse model uncovers low genotoxicity of lentiviral vector integration. *Nat. Biotechnol.* **24**: 687–696.

4. Cavazza, A, Moiani, A and Mavilio, F (2013). Mechanisms of retroviral integration and mutagenesis. *Hum. Gene Ther.* **24**: 119–31.

5. Craigie, R and Bushman, FD (2012). HIV DNA Integration. *Cold Spring Harb. Perspect. Med.* **2**: a006890.

6. Schröder, ARW, Shinn, P, Chen, H, Berry, C, Ecker, JR and Bushman, F (2002). HIV-1 integration in the human genome favors active genes and local hotspots. *Cell* **110**: 521–9.

7. Ciuffi, A, Llano, M, Poeschla, E, Hoffmann, C, Leipzig, J, Shinn, P, *et al.* (2005). A role for LEDGF/p75 in targeting HIV DNA integration. *Nat. Med.* **11**: 1287–9.

8. Stults, DM, Killen, MW, Pierce, HH and Pierce, AJ (2008). Genomic architecture and inheritance of human ribosomal RNA gene clusters. *Genome Res.* **18**: 13–8.

9. Schöfer, C and Weipoltshammer, K (2018). Nucleolus and chromatin. *Histochem. Cell Biol.* **150**: 209–225.

10. Scully, R, Panday, A, Elango, R and Willis, NA (2019). DNA double-strand break repair-pathway choice in somatic mammalian cells. *Nat. Rev. Mol. Cell Biol.* **20**: 698–714.

11. Yamamoto, Y and Gerbi, SA (2018). Making ends meet: targeted integration of DNA fragments by genome editing. *Chromosoma* **127**: 405–420.

12. Urnov, FD (2018). Ctrl-Alt-inDel: genome editing to reprogram a cell in the clinic. *Curr. Opin. Genet. Dev.* **52**: 48–56.

13. Muscarella, DE and Vogt, VM (1989). A mobile group I intron in the nuclear rDNA of Physarum polycephalum. *Cell* **56**: 443–54.

14. Ellison, EL and Vogt, VM (1993). Interaction of the intron-encoded mobility endonuclease I-PpoI with its target site. *Mol. Cell. Biol.* **13**: 7531–9.

15. Turkki, V, Schenkwein, D, Timonen, O, Husso, T, Lesch, HP and Ylä-Herttuala, S (2014). Lentiviral

protein transduction with genome-modifying HIV-1 integrase-I-PpoI fusion proteins: Studies on specificity and cytotoxicity. *Biomed Res. Int.* **2014**: 1–11.

16. Schenkwein, D, Turkki, V, Ahlroth, MK, Timonen, O, Airenne, KJ and Ylä-Herttuala, S (2013). rDNA-directed integration by an HIV-1 integrase--I-PpoI fusion protein. *Nucleic Acids Res.* **41**: e61.

17. Mannino, SJ, Jenkins, CL and Raines, RT (1999). Chemical mechanism of DNA cleavage by the homing endonuclease I-PpoI. *Biochemistry* **38**: 16178–86.

18. Mitchell, RS, Beitzel, BF, Schroder, ARW, Shinn, P, Chen, H, Berry, CC, *et al.* (2004). Retroviral DNA integration: ASLV, HIV, and MLV show distinct target site preferences. *PLoS Biol.* **2**: E234.

19. Brady, T, Agosto, LM, Malani, N, Berry, CC, O'Doherty, U and Bushman, F (2009). HIV integration site distributions in resting and activated CD4+ T cells infected in culture. *AIDS* **23**: 1461–71.

20. Smit, AFA, Hubley, R and Green, P. RepeatMasker Open-4.0: http://www.repeatmasker.org.

21. Robicheau, BM, Susko, E, Harrigan, AM and Snyder, M (2017). Ribosomal RNA genes contribute to the formation of pseudogenes and junk DNA in the human genome. *Genome Biol. Evol.* **9**: 380–397.

22. Marini, B, Kertesz-Farkas, A, Ali, H, Lucic, B, Lisek, K, Manganaro, L, *et al.* (2015). Nuclear architecture dictates HIV-1 integration site selection. *Nature* **521**: 227–231.

23. Biffi, A, Bartolomae, CC, Cesana, D, Cartier, N, Aubourg, P, Ranzani, M, *et al.* (2011). Lentiviral vector common integration sites in preclinical models and a clinical trial reflect a benign integration bias and not oncogenic selection. *Blood* **117**: 5332–9.

24. Aiuti, A, Biasco, L, Scaramuzza, S, Ferrua, F, Cicalese, MP, Baricordi, C, *et al.* (2013). Lentiviral hematopoietic stem cell gene therapy in patients with Wiskott-Aldrich syndrome. *Science* **341**: 1233151.

25. Biffi, A, Montini, E, Lorioli, L, Cesani, M, Fumagalli, F, Plati, T, *et al.* (2013). Lentiviral hematopoietic stem cell gene therapy benefits metachromatic leukodystrophy. *Science* **341**: 1233158.

26. Cartier, N, Hacein-Bey-Abina, S, Bartholomae, CC, Veres, G, Schmidt, M, Kutschera, I, *et al.* (2009). Hematopoietic stem cell gene therapy with a lentiviral vector in X-linked adrenoleukodystrophy. *Science* **326**: 818–823.

27. Németh, A and Längst, G (2011). Genome organization in and around the nucleolus. *Trends Genet.* **27**: 149–56.

28. Pontvianne, F, Carpentier, M-C, Durut, N, Pavlištová, V, Jaške, K, Schořová, Š, *et al.* (2016). Identification of Nucleolus-Associated Chromatin Domains Reveals a Role for the Nucleolus in 3D Organization of the A. thaliana Genome. *Cell Rep.* **16**: 1574–1587.

29. Yu, S and Lemos, B (2018). The long-range interaction map of ribosomal DNA arrays. *PLoS Genet.* **14**: e1007258.

30. Németh, A, Conesa, A, Santoyo-Lopez, J, Medina, I, Montaner, D, Péterfia, B, *et al.* (2010). Initial genomics of the human nucleolus. *PLoS Genet.* **6**: e1000889.

31. Dillinger, S, Straub, T and Nemeth, A (2017). Nucleolus association of chromosomal domains is largely maintained in cellular senescence despite massive nuclear reorganisation. *PLoS One* **12**: e0178821.

32. van Koningsbruggen, S, Gierlinski, M, Schofield, P, Martin, D, Barton, GJ, Ariyurek, Y, *et al.* (2010). High-resolution whole-genome sequencing reveals that specific chromatin domains from most human chromosomes associate with nucleoli. *Mol. Biol. Cell* **21**: 3735–48.

33. Yu, S and Lemos, B (2016). A Portrait of Ribosomal DNA Contacts with Hi-C Reveals 5S and 45S rDNA Anchoring Points in the Folded Human Genome. *Genome Biol. Evol.* **8**: 3545–3558.

34. Wang, GP, Ciuffi, A, Leipzig, J, Berry, CC and Bushman, FD (2007). HIV integration site selection: analysis by massively parallel pyrosequencing reveals association with epigenetic modifications. *Genome Res.* **17**: 1186–94.

35. Kosicki, M, Tomberg, K and Bradley, A (2018). Repair of double-strand breaks induced by CRISPR–Cas9 leads to large deletions and complex rearrangements. *Nat. Biotechnol.* **36**: 765–771.

36. Cullot, G, Boutin, J, Toutain, J, Prat, F, Pennamen, P, Rooryck, C, *et al.* (2019). CRISPR-Cas9 genome editing induces megabase-scale chromosomal truncations. *Nat. Commun.* **10**: 1136.

37. Simeonov, DR, Brandt, AJ, Chan, AY, Cortez, JT, Li, Z, Woo, JM, *et al.* (2019). A large CRISPR-induced bystander mutation causes immune dysregulation. *Commun. Biol.* **2**: 70.

38. Xu, S, Kim, J, Tang, Q, Chen, Q, Liu, J, Xu, Y, *et al.* (2020). CAS9 is a genome mutator by directly disrupting DNA-PK dependent DNA repair pathway. *Protein Cell* **11**: 352–365.

39. Stults, DM, Killen, MW, Williamson, EP, Hourigan, JS, Vargas, HD, Arnold, SM, *et al.* (2009). Human rRNA gene clusters are recombinational hotspots in cancer. *Cancer Res.* **69**: 9096–104.

40. Gibbons, JG, Branco, AT, Godinho, S a, Yu, S and Lemos, B (2015). Concerted copy number variation balances ribosomal DNA dosage in human and mouse genomes. *Proc. Natl. Acad. Sci. U. S. A.* **112**: 2485–2490.

41. Xu, B, Li, H, Perry, JM, Singh, VP, Unruh, J, Yu, Z, *et al.* (2017). Ribosomal DNA copy number loss and sequence variation in cancer. In: Eng, C (ed.). *PLoS Genet.* **13**: e1006771.

42. Killen, MW, Stults, DM, Adachi, N, Hanakahi, L and Pierce, AJ (2009). Loss of Bloom syndrome protein destabilizes human gene cluster architecture. *Hum. Mol. Genet.* **18**: 3417–28.

43. Floutsakou, I, Agrawal, S, Nguyen, TT, Seoighe, C, Ganley, ARD and McStay, B (2013). The shared genomic architecture of human nucleolar organizer regions. *Genome Res.* **23**: 2003–2012.

44. Lisowski, L, Lau, A, Wang, Z, Zhang, Y, Zhang, F, Grompe, M, *et al.* (2012). Ribosomal DNA Integrating rAAV-rDNA Vectors Allow for Stable Transgene Expression. *Mol. Ther.* **20**: 1912–23.

45. Wang, Z, Lisowski, L, Finegold, MJ, Nakai, H, Kay, M a and Grompe, M (2012). AAV Vectors Containing rDNA Homology Display Increased Chromosomal Integration and Transgene Persistence. *Mol. Ther.* **20**: 1902–1911.

46. Wang, Y, Zhao, J, Duan, N, Liu, W, Zhang, Y, Zhou, M, *et al.* (2018). Paired CRISPR/Cas9 Nickases Mediate Efficient Site-Specific Integration of F9 into rDNA Locus of Mouse ESCs. *Int. J. Mol. Sci.* **19**: 3035.

47. Liu, B, Chen, F, Wu, Y, Wang, X, Feng, M, Li, Z, *et al.* (2017). Enhanced tumor growth inhibition by mesenchymal stem cells derived from iPSCs with targeted integration of interleukin24 into rDNA loci. *Oncotarget* **8**: 40791–40803.

48. Basu, AS (2017). Digital Assays Part I: Partitioning Statistics and Digital PCR. *SLAS Technol.* **22**: 369–

386.

49. Quan, PL, Sauzade, M and Brouzes, E (2018). DPCR: A technology review. *Sensors (Switzerland)* **18**: 1271.

50. Larsen, DH and Stucki, M (2016). Nucleolar responses to DNA double-strand breaks. *Nucleic Acids Res.* **44**: 538–544.

51. Diesch, J, Bywater, MJ, Sanij, E, Cameron, DP, Schierding, W, Brajanovski, N, *et al.* (2019). Changes in long-range rDNA-genomic interactions associate with altered RNA polymerase II gene programs during malignant transformation. *Commun. Biol.* **2**: 39.

52. Bushman, FD (1994). Tethering human immunodeficiency virus 1 integrase to a DNA site directs integration to nearby sequences. *Proc. Natl. Acad. Sci. U. S. A.* **91**: 9233–7.

53. Meehan, AM, Saenz, DT, Morrison, JH, Garcia-Rivera, J a, Peretz, M, Llano, M, *et al.* (2009). LEDGF/p75 proteins with alternative chromatin tethers are functional HIV-1 cofactors. *PLoS Pathog.* **5**: e1000522.

54. Silvers, RM, Smith, JA, Schowalter, M, Litwin, S, Liang, Z, Geary, K, *et al.* (2010). Modification of integration site preferences of an HIV-1-based vector by expression of a novel synthetic protein. *Hum. Gene Ther.* **21**: 337–49.

55. Gijsbers, R, Ronen, K, Vets, S, Malani, N, De Rijck, J, McNeely, M, *et al.* (2010). LEDGF hybrids efficiently retarget lentiviral integration into heterochromatin. *Mol. Ther.* **18**: 552–560.

56. Ferris, AL, Wu, X, Hughes, CM, Stewart, C, Smith, SJ, Milne, TA, *et al.* (2010). Lens epithelium-derived growth factor fusion proteins redirect HIV-1 DNA integration. *Proc. Natl. Acad. Sci. U. S. A.* **107**: 3135–40.

57. Schenkwein, D, Turkki, V, Kärkkäinen, H-R, Airenne, K and Ylä-Herttuala, S (2010). Production of HIV-1 integrase fusion protein-carrying lentiviral vectors for gene therapy and protein transduction. *Hum. Gene Ther.* **21**: 589–602.

58. Voelkel, C, Galla, M, Maetzig, T, Warlich, E, Kuehle, J, Zychlinski, D, *et al.* (2010). Protein

transduction from retroviral Gag precursors. *Proc. Natl. Acad. Sci. U. S. A.* **107**: 7805–10.

59. He, C, Gouble, A, Bourdel, A, Manchev, V, Poirot, L, Paques, F, *et al.* (2014). Lentiviral protein delivery of meganucleases in human cells mediates gene targeting and alleviates toxicity. *Gene Ther.* **21**: 759–766.

60. Uhlig, KM, Schülke, S, Scheuplein, VAM, Malczyk, AH, Reusch, J, Kugelmann, S, *et al.* (2015). Lentiviral Protein Transfer Vectors Are an Efficient Vaccine Platform and Induce a Strong Antigen-Specific Cytotoxic T Cell Response. *J. Virol.* **89**: 9044–9060.

61. Choi, JG, Dang, Y, Abraham, S, Ma, H, Zhang, J, Guo, H, *et al.* (2016). Lentivirus pre-packed with Cas9 protein for safer gene editing. *Gene Ther.* **23**: 627–633.

62. Prel, A, Caval, V, Gayon, R, Ravassard, P, Duthoit, C, Payen, E, *et al.* (2015). Highly efficient in vitro and in vivo delivery of functional RNAs using new versatile MS2-chimeric retrovirus-like particles. *Mol. Ther. - Methods Clin. Dev.* **2**: 15039.

63. Cai, Y, Bak, RO, Krogh, LB, Staunstrup, NH, Moldt, B, Corydon, TJ, *et al.* (2014). DNA transposition by protein transduction of the piggyBac transposase from lentiviral Gag precursors. *Nucleic Acids Res.* **42**: e28.

64. Cai, Y, Bak, RO and Mikkelsen, JG (2014). Targeted genome editing by lentiviral protein transduction of zinc-finger and TAL-effector nucleases. *Elife* **3**: e01911.

65. Skipper, KA, Nielsen, MG, Andersen, S, Ryø, LB, Bak, RO and Mikkelsen, JG (2018). Time-Restricted PiggyBac DNA Transposition by Transposase Protein Delivery Using Lentivirus-Derived Nanoparticles. *Mol. Ther. - Nucleic Acids* **11**: 253–262.

66. Lyu, P, Javidi-Parsijani, P, Atala, A and Lu, B (2019). Delivering Cas9/sgRNA ribonucleoprotein (RNP) by lentiviral capsid-based bionanoparticles for efficient 'hit-and-run' genome editing. *Nucleic Acids Res.* **47**: e99.

67. Mangeot, PE, Risson, V, Fusil, F, Marnef, A, Laurent, E, Blin, J, *et al.* (2019). Genome editing in primary cells and in vivo using viral-derived Nanoblades loaded with Cas9-sgRNA ribonucleoproteins. *Nat. Commun.* **10**: 45.

68. Chen, X and Gonçalves, MAFV (2016). Engineered viruses as genome editing devices. *Mol. Ther.* **24**: 447–457.

69. Mock, U, Riecken, K, Berdien, B, Qasim, W, Chan, E, Cathomen, T, *et al.* (2014). Novel lentiviral vectors with mutated reverse transcriptase for mRNA delivery of TALE nucleases. *Sci. Rep.* **4**: 1–8.

70. Knopp, Y, Geis, FK, Heckl, D, Horn, S, Neumann, T, Kuehle, J, *et al.* (2018). Transient Retrovirus-Based CRISPR/Cas9 All-in-One Particles for Efficient, Targeted Gene Knockout. *Mol. Ther. - Nucleic Acids* **13**: 256–274.

71. Lu, B, Javidi-Parsijani, P, Makani, V, Mehraein-Ghomi, F, Sarhan, WM, Sun, D, *et al.* (2019). Delivering SaCas9 mRNA by lentivirus-like bionanoparticles for transient expression and efficient genome editing. *Nucleic Acids Res.* **47**: e44.

72. Broeders, M, Herrero-Hernandez, P, Ernst, MPT, van der Ploeg, AT and Pijnappel, WWMP (2020). Sharpening the Molecular Scissors: Advances in Gene-Editing Technology. *iScience* **23**: 100789.

73. Kuroki-Kami, A, Nichuguti, N, Yatabe, H, Mizuno, S, Kawamura, S and Fujiwara, H (2019). Targeted gene knockin in zebrafish using the 28S rDNA-specific non-LTR-retrotransposon R2Ol. *Mob. DNA* **10**: 23.

74. Johansen, SD, Haugen, P and Nielsen, H (2007). Expression of protein-coding genes embedded in ribosomal DNA. *Biol. Chem.* **388**: 679–686.

75. Bierhoff, H, Schmitz, K, Maass, F, Ye, J and Grummt, I (2010). Noncoding transcripts in sense and antisense orientation regulate the epigenetic state of ribosomal RNA genes. *Cold Spring Harb. Symp. Quant. Biol.* **75**: 357–364.

76. Bierhoff, H, Dammert, MA, Brocks, D, Dambacher, S, Schotta, G and Grummt, I (2014). Quiescence-Induced LncRNAs Trigger H4K20 Trimethylation and Transcriptional Silencing. *Mol. Cell* **54**: 675–682.

77. Follenzi, A and Naldini, L (2002). Generation of HIV-1 derived lentiviral vectors. *Methods Enzymol.* **346**: 454–65.

78. Brady, T, Roth, SL, Malani, N, Wang, GP, Berry, CC, Leboulch, P, *et al.* (2011). A method to sequence

and quantify DNA integration for monitoring outcome in gene therapy. *Nucleic Acids Res.* **39**: e72.

79. Jiang, H, Lei, R, Ding, SW and Zhu, S (2014). Skewer: A fast and accurate adapter trimmer for next-generation sequencing paired-end reads. *BMC Bioinformatics* **15**: 182.

80. Kent, WJ (2002). BLAT---The BLAST-Like Alignment Tool. *Genome Res.* **12**: 656–664.

81. Karolchik, D (2004). The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.* **32**: 493D – 496.

82. Quinlan, AR and Hall, IM (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841–842.

83. Gel, B and Serra, E (2017). KaryoploteR: An R/Bioconductor package to plot customizable genomes displaying arbitrary data. In: Hancock, J (ed.). *Bioinformatics* **33**: 3088–3090.

84. Gel, B, Díez-Villanueva, A, Serra, E, Buschbeck, M, Peinado, MA and Malinverni, R (2016). RegioneR: An R/Bioconductor package for the association analysis of genomic regions based on permutation tests. *Bioinformatics* **32**: 289–291.

85. Belinky, F, Nativ, N, Stelzer, G, Zimmerman, S, Stein, TI, Safran, M, *et al.* (2015). PathCards: Multi-source consolidation of human biological pathways. *Database* **2015**: 6.

86. Fronza, R, Vasciaveo, A, Benso, A and Schmidt, M (2016). A Graph Based Framework to Model Virus Integration Sites. *Comput. Struct. Biotechnol. J.* **14**: 69–77.

87. Vasciaveo, A, Velevska, I, Politano, G, Savino, A, Schmidt, M and Fronza, R (2016). Common integration sites of published datasets identified using a graph-based framework. *Comput. Struct. Biotechnol. J.* **14**: 87–90.

88. Zhou, Y, Zhou, B, Pache, L, Chang, M, Khodabakhshi, AH, Tanaseichuk, O, *et al.* (2019). Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat. Commun.* **10**: 1523.

89. Hinrichs, AS, Karolchik, D, Baertsch, R, Barber, GP, Bejerano, G, Clawson, H, *et al.* (2006). The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res.* **34**: D590-8.

FIGURES LEGENDS

Figure 1: rDNA and the LVs generated in this study to direct integration into the I-PpoI site. A) An illustration of an acrocentric chromosome (top), the repeating rDNA units (yellow arrows) that contain the rRNA genes and the IGS (middle), and one rRNA gene (bottom). Each rRNA gene unit encodes a 45S pre-rRNA which serves as the precursor for the 18S, 5.8S and 28S rRNAs of mature ribosomes. The I-PpoI site within the 28S rRNA gene is highlighted with a red box. In the current genome version hg38 there are three I-PpoI sites on chromosome 21 that are annotated with a 28S rRNA gene (Table S1). B) Illustration of the different IN molecule -containing LVs studied in this work, with an enlargement of one IN-fusion protein -containing LV particle. rDNA: ribosomal DNA; rRNA: ribosomal RNA; ETS: external transcribed spacer; ITS: internal transcribed spacer; LV: lentivirus vector; IN: integrase; IN_{D64V}: integration deficient IN.

Figure 2: Effects of IN-I-$PpoI_{H78A/N119A}$ fusion protein inclusion on the integration characteristics of LVs. A) Composition of the integration site data and numbers of unique IS (UH) and multiple hit IS (MH) retrieved for the different vectors. B) Chromosomal distribution of integration sites. Chromosome numbers are shown on the X-axis. C) Distribution of integration sites with respect to upstream (US) regions of genes, the gene length (% of within gene) and downstream (DS) of genes. D) A more detailed illustration of IS distribution within the uniquely mapping (UH; blue) and repetitive (MH; orange) portions of the genome. E) Integration frequency within oncogenes. A list comprising 2579 human cancer genes (http://www.bushmanlab.org/links/genelists) was used for the comparison. The statistical differences between the IN-modified LVs and the control LV are shown above the bars ($p<0.0001$ for both). Statistical differences between LVs were calculated using two-sided

Fisher's Exact test (D+H LVs vs. D+N LVs) or with two-sided Chi-square test (INwt LV compared to D+H or D+N LVs). ***p<0.001; **p<0.01; *p<0.05. In C) the black asterisks denote differences between the control vector INwt LV and the IN-modified LVs, and grey asterisks denote differences between the D+H and D+N LVs. Intrag.: intragenic IS; Interg.: intergenic IS; INwt: wild type integrase; D+H: IN_{D64V} and IN-I-$PpoI_{H78A}$ -containing LVs; D+N: IN_{D64V} and IN-I-$PpoI_{N119A}$ -containing LVs.

Figure 3: Characterization of vector integration within the repetitive genome and rDNA. A) Integration frequency into different repeat types within the repetitive genome. B) Total efficiency of integration targeting into an rDNA unit (including the rRNA coding region and the IGS) and within a 235bp window around the I-PpoI site. For the ddPCR-based quantification of I-PpoI site -directed integration the mean (with SEM) of six measurements is shown. C-E: Coverage plots where read coverage on the positive strand (+ve; scale on the right Y-axis) is shown with a darker shade and on the negative strand (-ve; scale on the left Y-axis) with a lighter shade for each LV type. C) A large-scale view of IS read localization within the Chr21 locus containing annotated rRNA genes (window size: 50 kb). D: A close-up view of IS distribution within the 28S rRNA gene (window size: 1,6kb). E: Illustration of the reads mapping within and near the I-PpoI site (shown with purple fonts). Window size: 300bp. *Repeatmasker-identified repeats without manual correction and annotation of additional rRNA gene unit features. **Repeatmasker-identified repeats ¤: Integration frequency within an area extending 203bp upstream and 32bp downstream of the cleaved I-PpoI site (see Figure S3 for details).

Figure 4: Characterization of CIS-associated IS. A) All unique IS associated with CIS were analyzed for their occurrence in intergenic loci, pseudogenes, ncRNA genes ("RNA genes") and protein-encoding genes. The proportions of IS within each feature are shown as a percentage of all CIS-associated UH-IS. The numbers of CIS-contained IS are: 8450 for LV INwt; 333 for LV D+H and 81 for LV D+N. B) Characterization of the proportion of IS localizing to protein-encoding genes, pseudogenes, ncRNA genes and ribosomal protein -encoding genes (RPL and RPS genes) of all CIS-associated IS (UH-MH-CIS). The numbers of all CIS-associated IS are 2506 for LV D+H; 498 for LV D+N and 10367 for LV INwt. The differences between the vectors were analyzed with two-sided Fisher's exact test (D+H LVs vs. D+N LVs) or with two-sided Chi-square test (INwt LV compared to D+H or D+N LVs). ***p<0.001; *p<0.05. In B the asterisks are shown only for INwt LV, whose difference to each IN-modified LVs was similar. Ribosomal prot.: genes encoding for the protein constituents of mature ribosomes.

Figure 5: Characterization of preferential LV integration in specific gene sets and gene ontology terms. A) Integration frequency within pseudogenes and ribosomal protein genes, or pseudogenes derived of them. B) Integration frequency in gene sets involved in ribosome biogenesis (Ribosome biogenesis in eukaryotes SuperPath[85]) and ribosome RNA processing (rRNA processing in the nucleus and cytosol SuperPath[85]). C-E: Enrichment heatmaps of the most overrepresented pathways and processes among genes present in the CIS-engaged integration sites, colored by p-values. Heatmap in C: for D+H LVs; in D: for D+N LVs and in E: for INwt LVs. RPL/RPS genes: large subunit ribosomal proteins/small subunit ribosomal proteins, respectively, or pseudogenes derived of these genes. In A and B the differences between the data sets were calculated with two-sided Chi-square tests. ***$p<0.001$; *$p<0.05$.

Figure 6: Quantification of targeted integration in the 28S rRNA gene and detection of potential deletions in the rRNA gene and in the short arms of the acrocentric chromosomes. The proportion of vectors integrated near the I-PpoI site in the 28S rRNA gene was quantitated with ddPCR (A and B). The vector dose used (5k and 10k vp/cell) is shown in parenthesis after the LV abbreviation. The values of the two analyzed wells per vector and vp-dose combinations are shown (mean with SEM from duplicate measurements per sample; see also Table S9) with the results from T cells extracted from Donor 1 shown in A) and T cells from Donor 2 in B). The copy number of the 18S rRNA gene (C) and the DJ region (D) were quantitated from T cells transduced with 10k vp/cell at day 2 post transduction. The same sample replicates were used as in A and B. These four measurements per vector group (Table S10) are shown with their mean and SEM. The differences in copy numbers were analyzed with one-way ANOVA by comparing the vector-groups' values to the same donor's NTD control with Dunnett's Multiple Comparison Test. NTD: non-transduced cells; DJ: distal junction sequence; p.td: post transduction; rRNA: ribosomal RNA.

ABBREVIATIONS

IN: integrase; rDNA: ribosomal DNA; rRNA: ribosomal RNA; IS. integration site; NHEJ, non-homologous end joining; DSB: DNA double strand break; EGFP: enhanced green fluorescent protein; UH-IS: unique hit-IS; MH-IS: multiple hit -IS; ETS: external transcribed spacer; ITS: internal transcribed spacer; IGS, intergenic spacer; LV: lentivirus vector; LTR: long terminal repeat; IN_{D64V}: integration deficient IN; Intrag.; intragenic IS; Interg.: intergenic IS; INwt: wild type integrase; D+H: IN_{D64V} and IN-I-$PpoI_{H78A}$ -containing LVs; D+N: IN_{D64V} and IN-

I-PpoI$_{N119A}$ -containing LVs; ns: not significant; CIS: common integration site; NAD: nucleolus associated domain; tRNA: transfer RNA; ncRNA, non-coding RNA; DJ: distal junction sequence.

Table 1: Characterization of the strongest integration hotspots among the uniquely mappable IS.

	Rank	IS #	Median location	Gene[a]	Repeat[a,b]	Nearest RefSeq gene	Dimension (kB)
INwt (UH)	1	67	chr16:1633220	CRAMP1	SINE/Alu		524
	2	53	chr8:144306704	HSF1	LINE/L1		475
	3	52	chr16:2080539	TSC2	SINE/Alu		334
	4	44	chr11:66094636	PACS1	LINE/L1		465
	5	35	chr11:65566836	intergenic	na	SSSCA1-AS1	235
	6	33	chr16:688665	WDR24	na		368
	7	31	chr1:1334252	TAS1R3	na		184
	8	28	chr19:1199664	intergenic	na	STK11	223
	9	27	chr6:30681690	PPP1R18	SINE/Alu		317
	10	25	chr17:81593484	NPLOC4	DNA/hAT-Charlie		163
	11	22	Chr17:82147186	CCDC57	simple		279
	12	21	Chr9:128599563	SPTAN1	SINE/Alu		311
	13	19	Chr12:49150673	intergenic	SINE/Alu	TUBA1B	247
	13	19	Chr19:49842535	PTOV1-AS1	SINE/Alu		157
	14	18	Chr6:31687953	ABHD16A	na		182
	14	18	Chr10:112589294	VTI1A	LTR/ERV-MaLR		174
	15	17	Chr11:65218552	SLC22A20P	na		166
	15	17	Chr17:81880539	intergenic	LINE/L1	ALYREF	84
D+H (UH)	1	12	chr6:27631516	intergenic	(tRNA)	LINC01012	37
	2	11	chr6:28658243	intergenic	tRNA	LINC00533	86
	3	10	chr5:140711372	VTRNA1-1	na		8
	4	9	chr2:38482053	LOC101929596 (RPLP0P6)	na		1
	4	9	chr3:182901763	ATP11B	na		0
	4	9	chr20:30512867	intergenic	rRNA (LSU)	MLLT10P1	1
	5	6	chr2:131102011	intergenic	na		69
	5	6	chr2:132279863	intergenic	rRNA (LSU)	ANKRD30BL	0
	6	5	chr11:65611215	MAP3K11	na		55
	6	5	chr17:81897445	ANAPC11	na		52
	6	5	chr20:44466866	intergenic (RPL37AP1)	na	LINC01620 /C20orf62	0
	7	4	chr1:8866735	ENO1	na		17
	7	4	chr1:174904258	RABGAP1L	SINE/Alu		48
	7	4	chr2:3577177	RPS7	SINE/Alu		19
	7	4	chr2:27050883	intergenic	(tRNA)	AGBL5-AS1	30
	7	4	chr4:145884509	ZNF827	na		47
	7	4	chr5:122352156	SNCAIP	na		37
	7	4	chr6:153282725	intergenic (RPL27AP6)	na	RGS17	32
	7	4	chr10:125738308	EDRF1	na		0
	7	4	chr11:77886544	INTS4/AAMDC	rRNA (LSU)		15
	7	4	chr12:56175248	SMARCC2	SINE/Alu		22
	7	4	chr16:685472	WDR24	na		29
	7	4	chr19:1131901	SBNO2	na		36
	7	4	chr19:12894097	GCDH (RPS6P25)	na		36
	7	4	chr21:8415028	intergenic	simple (45S rRNA)[c]	MIR6724-1	39
	7	4	chrX:135542502	INTS6L	SINE/Alu		0
D+N (UH)	1	10	chr6:27631467	intergenic	tRNA	LINC01012	167
	2	7	chr8:144456689	CYHR1	na		114
	3	5	chr11:66348159	LOC102724064	tRNA		7
	3	5	chr12:56190397	intergenic	tRNA	SMARCC2	0
	3	5	chr19:3982952	EEF2	na		6
	4	4	chr5:140711372	VTRNA1-1	na		8
	5	3	chr1:951876	NOC2L	na		6
	5	3	chr1:145157237	intergenic	tRNA	LOC103091866	0
	5	3	chr1:156312177	CCT3	na		8
	5	3	chr2:27050871	intergenic	tRNA (SINE/Alu)	AGBL5-AS1	15
	5	3	chr5:178204539	HNRNPAB	na		38
	5	3	chr5:181236966	RACK1	na		51
	5	3	chr7:5634480	RNF216	na		39
	5	3	chr8:144311250	HSF1	na		5
	5	3	chr9:127972911	FAM102A	na		44
	5	3	chr9:136375334	intergenic	na	SNAPC4	8
	5	3	chr16:1817574	HAGH	na		18
	5	3	chr16:1960749	NDUFB10	SINE/MIR		15
	5	3	chr16:67887498	NRN1L	SINE/Alu		8
	5	3	chr17:8221619	LINC00324	tRNA		6
	5	3	chr20:63678092	RTEL1	na		4

Table 2: Characterization of the strongest integration hotspots among all IS.

	Rank	IS #	Median location	Gene[a]	Repeat[a,b]	Nearest RefSeq gene	Dimension (kB)	MH%
INwt (UH+MH)	1	68	chr16:1163	CRAMP1L	na		524	1.5

		9939						
2	61	chr16:20837500	TSC2	na		334	148.	
3	57	chr8:144321862	DGAT1	na		475	70.	
4	51	chr11:66093177	PACS1	LINE/L1		465	137.	
5	38	chr11:65553655	LTBP3	na		258	79.	
6	34	chr1:133648 83	DVL1	na		184	88.	
7	33	ch	WD	na		36	0.	

		chr16:68866565	R24			8	0	
8	31	chr19:26670953	centromeric	Satellite/centr.	LINC00662	544	100.0	
9	30	chr6:30665510	DHX16	na		317	100.0	
10	28	chr19:11996964	intergenic	na	STK11	223	0.0	
11	26	chr17:81592393	NPLOC4	na		163	3.8	
12	25	chr17:8	CCDC57	na		371	120.0	

			21427221				
	12	25	chr22:50382044	PPP6R2	LINE/L1	279	280.
	13	23	chr9:122860611 5	SPTAN1	SINE/Alu	311	87.
	14	22	chr12:49147302	intergenic	SINE/Alu	TUBA1A 247	136.
	15	21	chr19:49849663	PTOV1-AS1	na	157	95.
D+H (UH+MH)	1	1367	chr21:8444	RNA28SN1	rRNA (28S)	MIR6724-4 135	997.

		914					
2	130	chr14:49862760	RN7SL2	srpRNA/7SLRNA	9		100.0
3	53	chr14:49586605	RN7SL1	srpRNA/7SLRNA	0		100.0
4	37	chr20:30512867	intergenic	rRNA (28S)	MLLT10P1	0	75.7
5	34	chr2:132279864	intergenic	rRNA (28S)	ANKRD30BL	0	82.4
6	33	chr11:77886544	INTS4/AAMDC	rRNA (28S)		30	87.9

7	29	chr6:27631414	intergenic	(tRNA)	LINC01012	163	48.3
7	29	chr17:81872235	intergenic	tRNA	MIR4521	111	93.1
8	26	chr1:228646038	RHOU/DUSP5P1/RNA5S17	5S rRNA		3	100.0
9	23	chr19:50128415	SNAR-A11	na		11	100.0
10	21	chr6:125780026	intergenic	tRNA	NCOA7	43	90.5

11	19	chr1:237603123	RYR2	rRNA (28S)	0	0	84.2
12	15	chrX:109054236	intergenic	rRNA (28S)	MIR6087	0	100.0
13	14	chr16:31915722	intergenic	tRNA	ORF1	83	85.7
13	14	chr21:17454798	intergenic	tRNA	LINC01549	0	100.0
13	14	chr22:32039474	intergenic (RPS17	na	SLC5A1	0	78.6

				P16)					
	14	13	chr8:69690270	SLCO5A1	rRNA (28S)		0	76.9	
	15	12	chr6:28863698	intergenic	tRNA	LINC01623	192	66.7	
D+N (UH+MH)	1	73	chr21:84449044	intergenic	rRNA (28S)	MIR6724-4	12	100.0	
	2	47	chr17:81258819	intergenic	(tRNA)	HES7	108	91.5	
	3	29	chr14:49862666	RN7SL2	srpRNA/7SLRNA		0	100.0	
	4	22	chr	int	SIN	LIN	167	54.	

		6:27618534	ergenic	E/Alu (tRNA)	C01012		5
5	15	chr5:181207830	intergenic	tRNA	TRIM7	68	80.0
5	15	chr16:31915501	intergenic	tRNA	OR1F1	16	93.3
6	11	chr6:125780305	intergenic	tRNA	NCOA7	0	100.0
6	11	chr14:49586625	RN7SL1	srpRNA/7SLRNA		0	90.9
6	11	chr19:	SNAR-A	na		16	100.0

		50128411	11				
7	9	chr19:46811031	intergenic	SINE/Alu	SNAR-E	20	77.8
8	8	chr1:145288441	intergenic	tRNA	NBPF20	0	100.0
8	8	chr1:161425134	intergenic	LINE/L1 (tRNA)	CFAP126	85	87.5
8	8	chr17:82494740	intergenic	tRNA	NARF	0	100.0
8	8	chr19:102	RNU6-2	snRNA/U6		59	87.5

		1625						
97		chr1:228646036	RHOU/DUSP5P1	(5S rRNA)		2	100.0	
97		chr5:140711373	VTRNA1-1	na		15	42.9	
97		chr8:144456689	CYHR1	na		14	0.0	
97		chr12:56190397	intergenic	tRNA	SMARCC2	0	28.6	
97		chr14:58239889	intergenic	tRNA	ACTR10	0	100.0	

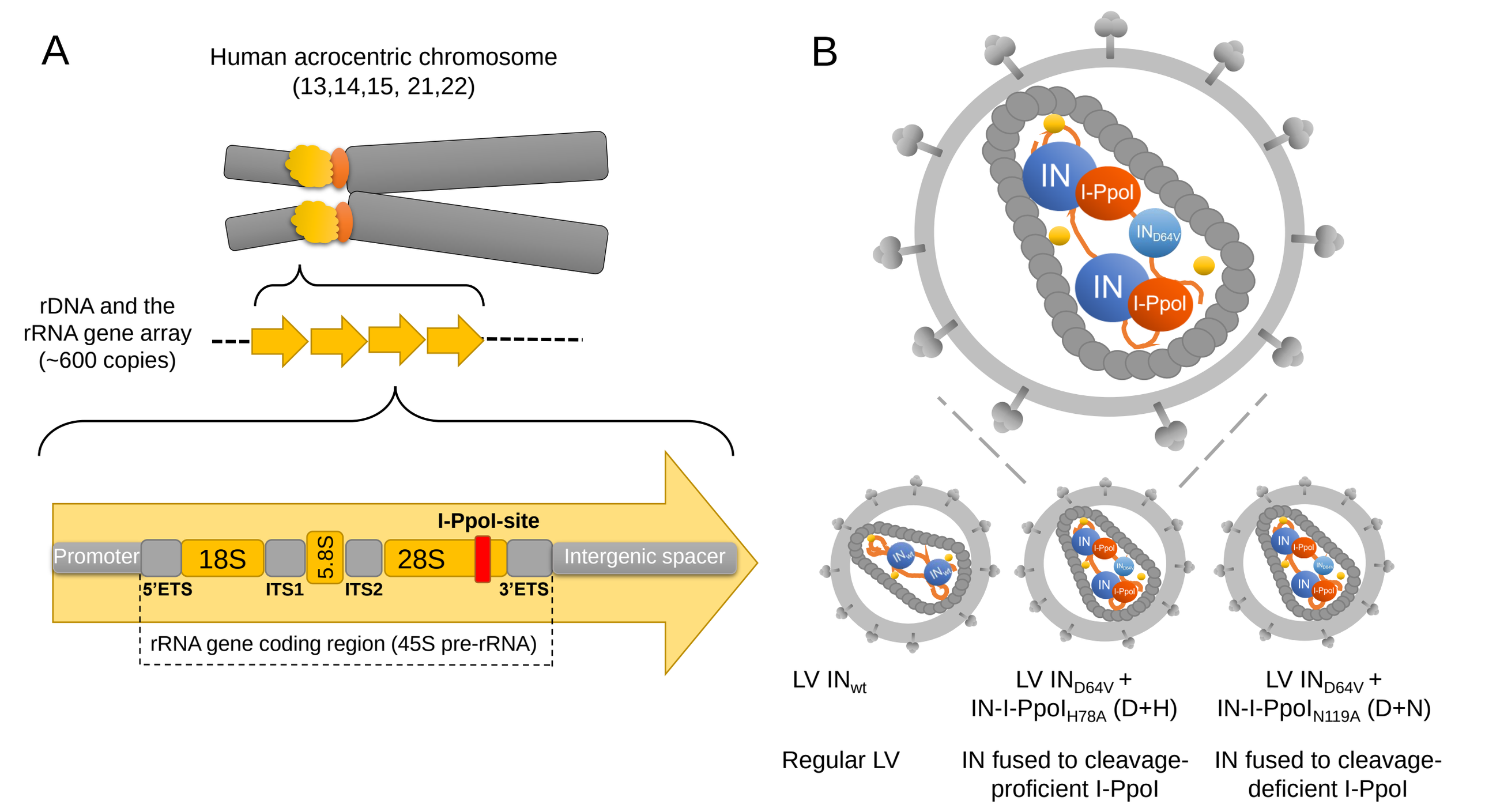
		4						
9	7	chr15:45201222	intergenic	tRNA	SHF	0	1	100.0
9	7	chr19:13835944	intergenic	tRNA	NDUFS7	4	1	85.7
9	7	chr21:17454808	intergenic	tRNA	LINC01549	0	1	100.0
10	6	chr9:133020150	intergenic(EEF1A1P5)	na	SNORD141A	1	1	100.0
10	6	chr11:6634	LOC102724240	tRNA		7	1	16.7

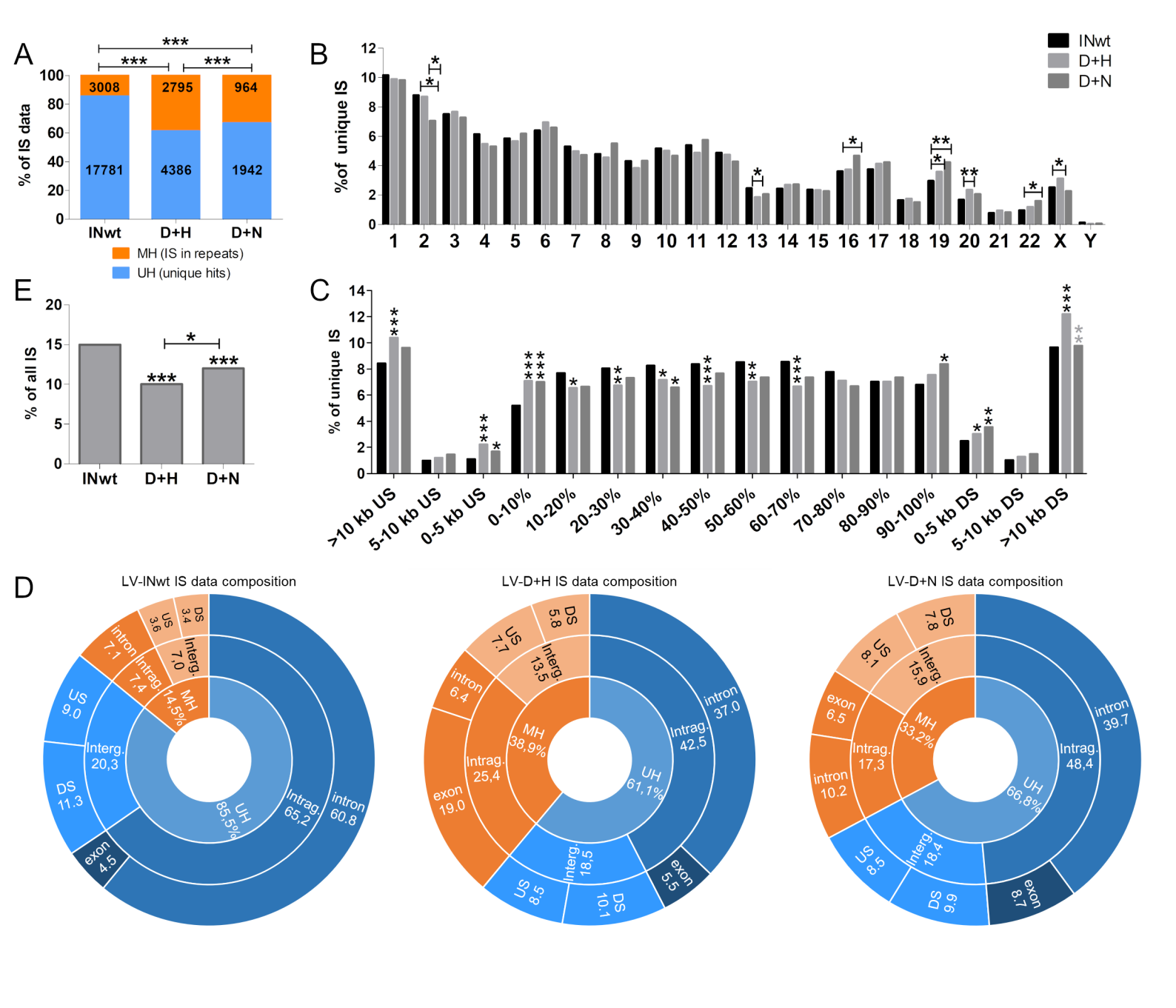
			81555	64					
10	6	chr16:68742482	CDH1	5SrRNA			0	100.0	
10	6	chr19:47241132	intergenic	tRNA	DPP9		0	100.0	

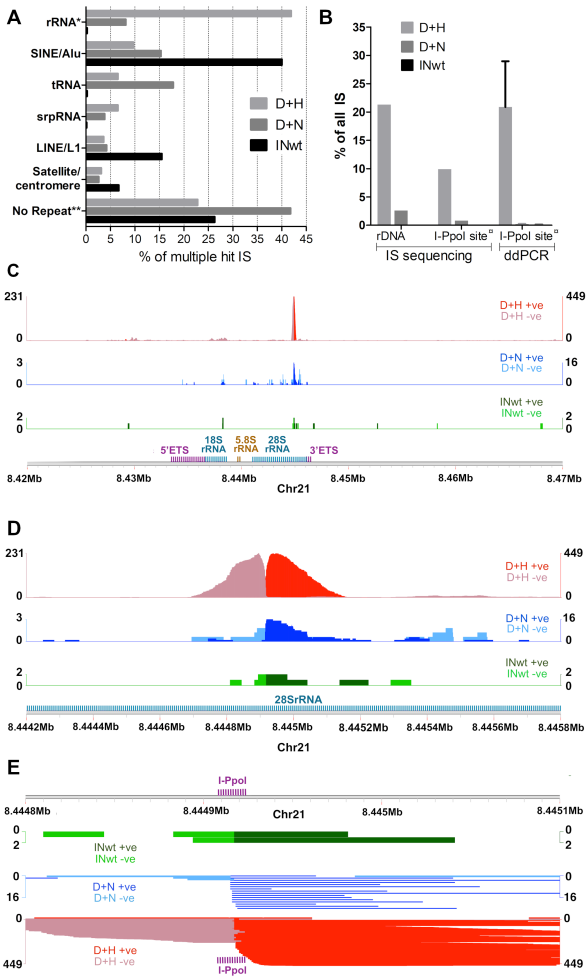
[a]Gene and repeat family in CIS median locus. [b]Repeat is shown in parenthesis if it is found in > 50% of the reads, but not in the exact CIS median locus. [c]ISs are placed into the IGS (UCSC genome browser Hg38). MH%: Fraction of the Multiple hit-IS of all CIS-forming IS. UH: unique hits

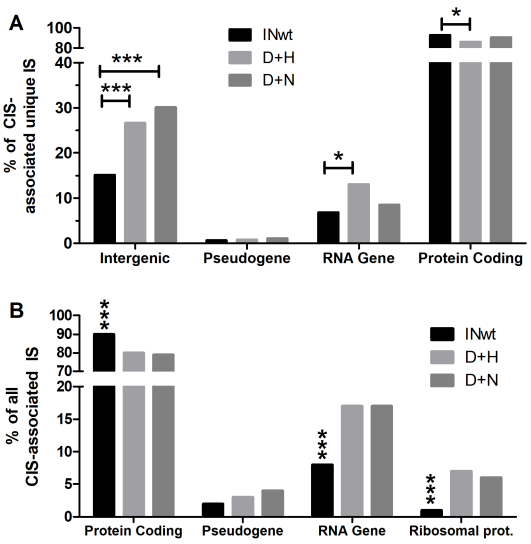
eTOC synopsis:

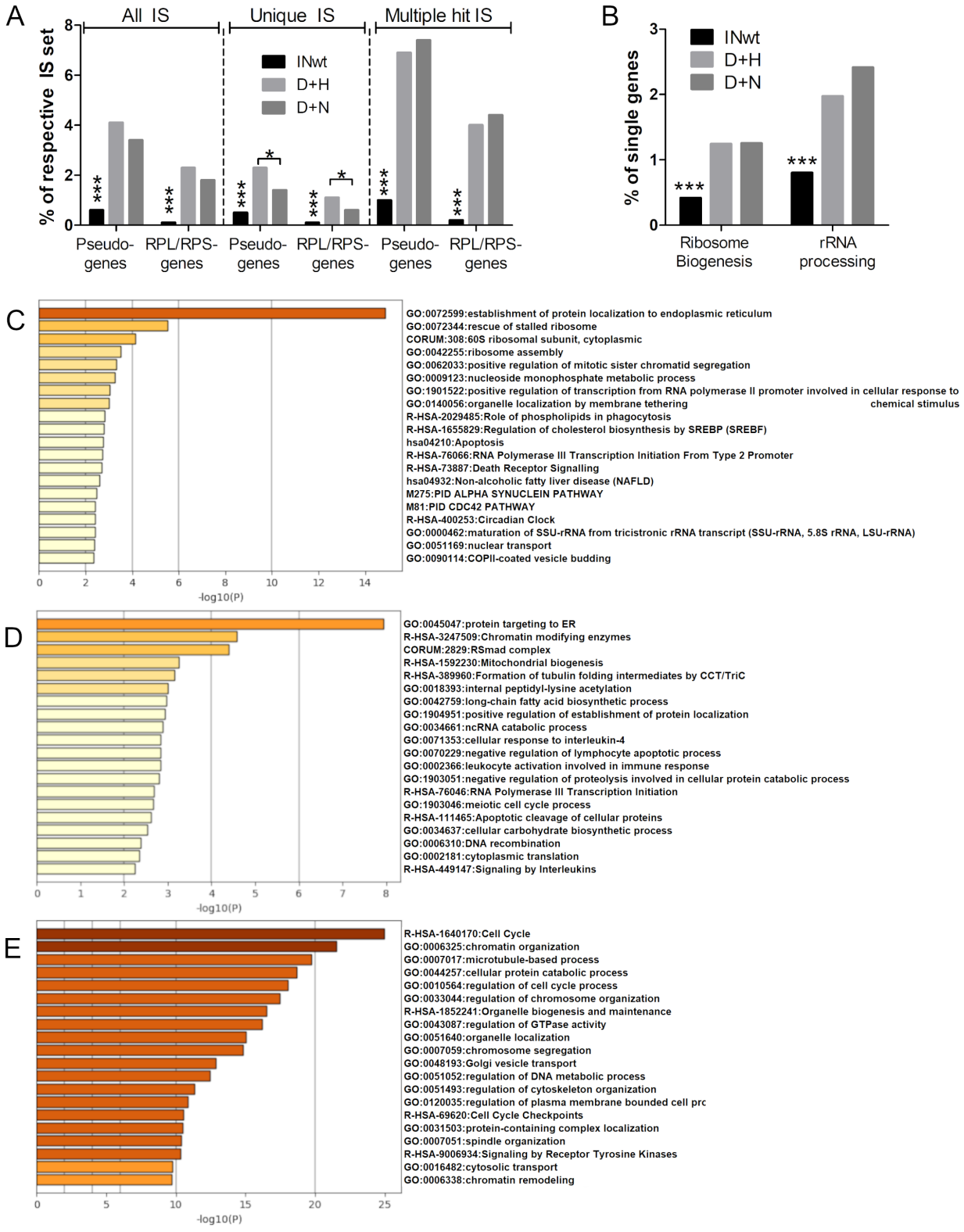
Random integration of therapeutic genes can cause undesired side-effects. This study shows that lentivirus vector integration can be efficiently targeted to ribosomal DNA with vectors that carry an endonuclease and the transgene. rDNA cleavage and targeted integration were well tolerated by primary human T cells and the transgene became transcribed.


A
Human acrocentric chromosome
(13,14,15, 21,22)
rDNA and the
rRNA gene array
(~600 copies)
I-PpoI-site
Promoter
18S
5.8S
28S
Intergenic spacer
5'ETS
ITS1
ITS2
3'ETS
rRNA gene coding region (45S pre-rRNA)
B
IN
I-PpoI
IN$_{D64V}$
IN$_{wt}$
LV IN$_{wt}$
Regular LV
LV IN$_{D64V}$ +
IN-I-PpoI$_{H78A}$ (D+H)
IN fused to cleavage-
proficient I-PpoI
LV IN$_{D64V}$ +
IN-I-PpoI$_{N119A}$ (D+N)
IN fused to cleavage-
deficient I-PpoI


A
% of IS data
3008
2795
964
17781
4386
1942
INwt
D+H
D+N
MH (IS in repeats)
UH (unique hits)
B
%of unique IS
INwt
D+H
D+N
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y
C
% of unique IS
>10 kb US
5-10 kb US
0-5 kb US
0-10%
10-20%
20-30%
30-40%
40-50%
50-60%
60-70%
70-80%
80-90%
90-100%
0-5 kb DS
5-10 kb DS
>10 kb DS
E
% of all IS
INwt
D+H
D+N
D
LV-INwt IS data composition
LV-D+H IS data composition
LV-D+N IS data composition
UH 85,5%
MH 14,5%
Intrag. 65,2
Interg. 20,3
intron 60.8
exon 4.5
DS 11.3
US 9.0
intron 7.1
Intrag. 7,4
US 3.6
DS 3.4
Interg. 7.0
UH 61,1%
MH 38,9%
Intrag. 42,5
Intrag. 25,4
Interg. 13,5
intron 37.0
exon 5.5
DS 10.1
US 8.5
Interg. 18,5
exon 19.0
intron 6.4
US 7.7
DS 5.8
UH 66,8%
MH 33,2%
Intrag. 48,4
Intrag. 17,3
Interg. 15,9
intron 39.7
exon 8.7
DS 9.9
Interg. 18.4
US 8.5
intron 10.2
exon 6.5
US 8.1
DS 7.8


A
rRNA*
SINE/Alu
tRNA
srpRNA
LINE/L1
Satellite/ centromere
No Repeat**
D+H
D+N
INwt
0 5 10 15 20 25 30 35 40 45
% of multiple hit IS
B
D+H
D+N
INwt
% of all IS
0 5 10 15 20 25 30 35
rDNA
I-PpoI site°
I-PpoI site°
IS sequencing
ddPCR
C
231
0
D+H +ve
D+H -ve
449
0
3
0
D+N +ve
D+N -ve
16
0
2
0
INwt +ve
INwt -ve
2
0
18S rRNA
5.8S rRNA
28S rRNA
5'ETS
3'ETS
8.42Mb 8.43Mb 8.44Mb 8.45Mb 8.46Mb 8.47Mb
Chr21
D
231
0
D+H +ve
D+H -ve
449
0
3
0
D+N +ve
D+N -ve
16
0
2
0
INwt +ve
INwt -ve
2
0
28SrRNA
8.4442Mb 8.4444Mb 8.4446Mb 8.4448Mb 8.445Mb 8.4452Mb 8.4454Mb 8.4456Mb 8.4458Mb
Chr21
E
I-PpoI
8.4448Mb 8.4449Mb Chr21 8.445Mb 8.4451Mb
0
2
INwt +ve
INwt -ve
0
2
0
16
D+N +ve
D+N -ve
0
16
0
449
D+H +ve
D+H -ve
I-PpoI
0
449




A
All IS
Unique IS
Multiple hit IS
INwt
D+H
D+N
% of respective IS set
Pseudo-genes
RPL/RPS-genes
B
% of single genes
Ribosome Biogenesis
rRNA processing
C
GO:0072599:establishment of protein localization to endoplasmic reticulum
GO:0072344:rescue of stalled ribosome
CORUM:308:60S ribosomal subunit, cytoplasmic
GO:0042255:ribosome assembly
GO:0062033:positive regulation of mitotic sister chromatid segregation
GO:0009123:nucleoside monophosphate metabolic process
GO:1901522:positive regulation of transcription from RNA polymerase II promoter involved in cellular response to
GO:0140056:organelle localization by membrane tethering
chemical stimulus
R-HSA-2029485:Role of phospholipids in phagocytosis
R-HSA-1655829:Regulation of cholesterol biosynthesis by SREBP (SREBF)
hsa04210:Apoptosis
R-HSA-76066:RNA Polymerase III Transcription Initiation From Type 2 Promoter
R-HSA-73887:Death Receptor Signalling
hsa04932:Non-alcoholic fatty liver disease (NAFLD)
M275:PID ALPHA SYNUCLEIN PATHWAY
M81:PID CDC42 PATHWAY
R-HSA-400253:Circadian Clock
GO:0000462:maturation of SSU-rRNA from tricistronic rRNA transcript (SSU-rRNA, 5.8S rRNA, LSU-rRNA)
GO:0051169:nuclear transport
GO:0090114:COPII-coated vesicle budding
-log10(P)
D
GO:0045047:protein targeting to ER
R-HSA-3247509:Chromatin modifying enzymes
CORUM:2829:RSmad complex
R-HSA-1592230:Mitochondrial biogenesis
R-HSA-389960:Formation of tubulin folding intermediates by CCT/TriC
GO:0018393:internal peptidyl-lysine acetylation
GO:0042759:long-chain fatty acid biosynthetic process
GO:1904951:positive regulation of establishment of protein localization
GO:0034661:ncRNA catabolic process
GO:0071353:cellular response to interleukin-4
GO:0070229:negative regulation of lymphocyte apoptotic process
GO:0002366:leukocyte activation involved in immune response
GO:1903051:negative regulation of proteolysis involved in cellular protein catabolic process
R-HSA-76046:RNA Polymerase III Transcription Initiation
GO:1903046:meiotic cell cycle process
R-HSA-111465:Apoptotic cleavage of cellular proteins
GO:0034637:cellular carbohydrate biosynthetic process
GO:0006310:DNA recombination
GO:0002181:cytoplasmic translation
R-HSA-449147:Signaling by Interleukins
-log10(P)
E
R-HSA-1640170:Cell Cycle
GO:0006325:chromatin organization
GO:0007017:microtubule-based process
GO:0044257:cellular protein catabolic process
GO:0010564:regulation of cell cycle process
GO:0033044:regulation of chromosome organization
R-HSA-1852241:Organelle biogenesis and maintenance
GO:0043087:regulation of GTPase activity
GO:0051640:organelle localization
GO:0007059:chromosome segregation
GO:0048193:Golgi vesicle transport
GO:0051052:regulation of DNA metabolic process
GO:0051493:regulation of cytoskeleton organization
GO:0120035:regulation of plasma membrane bounded cell prc
R-HSA-69620:Cell Cycle Checkpoints
GO:0031503:protein-containing complex localization
GO:0007051:spindle organization
R-HSA-9006934:Signaling by Receptor Tyrosine Kinases
GO:0016482:cytosolic transport
GO:0006338:chromatin remodeling
-log10(P)