5-8-2014

Lentivector integration sites in ependymal cells from a model of metachromatic leukodystrophy: Non-b dna as a new factor influencing integration

Robert G. McAllister
Schulich School of Medicine & Dentistry

Jiahui Liu
Western University

Matthew W. Woods
Schulich School of Medicine & Dentistry

Sean K. Tom
Schulich School of Medicine & Dentistry

Anthony Rupar
Western University, tony.rupar@lhsc.on.ca

See next page for additional authors

Citation: Molecular Therapy—Nucleic Acids (2014) 3, e187; doi:10.1038/mtna.2014.39

www.nature.com/mtna

Lentivector Integration Sites in Ependymal Cells From a Model of Metachromatic Leukodystrophy: Non-B DNA as a New Factor Influencing Integration

Robert G McAllister[1], Jiahui Liu[2], Matthew W Woods[1] , Sean K Tom[1] ,C Anthony Rupar[2,3,4,5] and Stephen D Barr[1]

The blood–brain barrier controls the passage of molecules from the blood into the central nervous system (CNS) and is a major challenge for treatment of neurological diseases. Metachromatic leukodystrophy is a neurodegenerative lysosomal storage disease caused by loss of arylsulfatase A (ARSA) activity. Gene therapy via intraventricular injection of a lentiviral vector is a potential approach to rapidly and permanently deliver therapeutic levels of ARSA to the CNS. We present the distribution of integration sites of a lentiviral vector encoding human *ARSA* (LV-ARSA) in murine brain choroid plexus and ependymal cells, administered via a single intracranial injection into the CNS. LV-ARSA did not exhibit a strong preference for integration in or near actively transcribed genes, but exhibited a strong preference for integration in or near satellite DNA. We identified several genomic hotspots for LV-ARSA integration and identified a consensus target site sequence characterized by two G-quadruplex-forming motifs flanking the integration site. In addition, our analysis identified several other non-B DNA motifs as new factors that potentially influence lentivirus integration, including human immunodeficiency virus type-1 in human cells. Together, our data demonstrate a clinically favorable integration site profile in the murine brain and identify non-B DNA as a potential new host factor that influences lentiviral integration in murine and human cells.
Molecular Therapy—Nucleic Acids (2014) **3**, e187; doi:10.1038/mtna.2014.39; published online 26 August 2014

Introduction

The blood–brain barrier controls the passage of molecules from the blood into the central nervous system (CNS) and is a major challenge for the treatment of neurological diseases. Metachromatic leukodystrophy (MLD) is an autosomal recessively inherited neurodegenerative lysosomal storage disease caused by the loss of arylsulfatase A (ARSA) activity. ARSA is required to catalyze the first step in the degradation pathway of galactosyl-3-sulfate ceramide (sulfatide), a major sphingolipid of myelin. The loss of ARSA activity results in the accumulation of sulfatide in glial cells and neurons, followed by severe demyelination and neurodegeneration.[1,2]

Visceral manifestations of some lysosomal storage disease may be treated by intravenous enzyme replacement therapy, but the blood–brain barrier prevents access to the CNS in MLD and other lysosomal storage diseases with neurological involvement. The efficacy of brain gene therapy to correct ARSA deficiency has been demonstrated in mice and nonhuman primates using direct injection of serotype 5 recombinant adeno-associated vector in the brain.[3,4] However, the requirement of multiple injections in different regions of the brain and low transduction efficiencies warrant the development of improved vector delivery approaches. Recently, restoration of ARSA deficiency has been achieved in mice by lentiviral transduction of hematopoietic stem cells (HSCs) *ex vivo*, followed by re-infusion of the engineered HSCs.[5–7] A similar HSC gene therapy approach was shown to prevent progression to neurodegenerative disease in three presymptomatic patients who were predicted to experience early-onset MLD.[8] The most common presentation of MLD is the rapidly progressive late infantile disease. Most patients are significantly symptomatic by the time of diagnosis and the relatively slow replacement of brain microglia from bone-marrow-derived cells may pose a significant challenge in halting the rapid progression of the disease.

In mammals, the choroid plexus is a structure in the ventricles of the brain consisting of modified ependymal cells. The ependymal cell lining (ependyma) is a single-layered, cuboidal-columnar, ciliated epithelium that normally lines the cerebral ventricles and central canal of the spinal cord. The ependymocyte is a fully differentiated cell that remains in a position adjacent to the cerebrospinal fluid (CSF). The coordinated beating of the ependymal cilia creates a current of CSF along the walls of the lateral ventricle that optimizes the dispersion of neural messengers in the CSF.[9–12] The CSF also provides access to several different regions of the brain via the Virchow–Robbins spaces. Lentiviral transduction of ependymal cells is therefore an attractive approach to rapidly and permanently deliver therapeutic levels of proteins or enzymes to a broad area of the CNS and is less invasive than multiple injections throughout the brain.[13]

Among the advantages of using lentiviral vectors for delivery of therapeutic proteins are their high transduction efficiencies, permanence, and ability to transduce fully differentiated cells. The potential for adverse events such as insertional activation of oncogenes or the disruption of gene coding sequences necessitates analysis of lentiviral vector integration sites in

[1]Department of Microbiology and Immunology, Schulich School of Medicine and Dentistry, Center for Human Immunology, Western University, London, Ontario, Canada; [2]Department of Biochemistry, Western University, London, Ontario, Canada; [3]Department of Pathology and Laboratory Medicine, Western University, London, Ontario, Canada; [4]Department of Pediatrics, Western University, London, Ontario, Canada; [5]Children's Health Research Institute, Western University, London, Ontario, Canada. Correspondence: Stephen D. Barr, Department of Microbiology and Immunology, Schulich School of Medicine and Dentistry, Center for Human Immunology, Western University, London, Ontario, N6A5C1, Canada. E-mail: stephen.barr@uwo.ca
Received 10 February 2014; accepted 7 July 2014; published online 26 August 2014. doi:10.1038/mtna.2014.39

the genome of cells targeted by clinical lentiviral gene therapy vectors.[14,15] Here we present an analysis of the distribution of integration sites of a lentiviral vector encoding the human *ARSA* gene (LV-ARSA) in murine brain ependymal cells, administered via a single intracranial injection into the right lateral ventricle of *ARSA*–/– and wild-type (WT) mice.

Results

Intraventricular administration of LV-ARSA

LV-ARSA was injected into the right lateral ventricle of 4 ARSA–/– and 4 WT C57Bl6 4-month-old mice of mixed genders (**Figure 1a**,**b**). Seven days after administration, the brain tissue surrounding the lateral ventricle containing the choroid plexus was dissected from each mouse (**Figure 1b**). In a separate experiment to visualize transduced cells, *ARSA* was replaced with *green fluorescent protein* (*GFP*) to generate LV-GFP (**Figure 1a**). LV-GFP was injected into the right lateral ventricle, and immunofluorescence microscopy analysis of tissue sections showed that LV-GFP transduced cells in the choroid plexus and the ependymal cells that lined the ventricle. No substantial GFP fluorescence was observed in cells of the surrounding brain tissue, suggesting that the vector did not pass through or between the ependymal cell layer (**Figure 1c**). The distribution of GFP positive cells obtained 7 days after injection did not differ 7 months after injection (data not shown).

To characterize vector integration sites, genomic DNA was isolated from the dissected tissue and subjected to restriction enzyme digestions and linker-mediated amplification using primers as previously described (see Materials and Methods section and ref. 16). Following amplification, the purified DNA samples were processed using the Nextera XT DNA Sample Preparation kit, which uses an engineered transposome to simultaneously fragment and tag the input DNA with unique adapter and index ("barcodes") sequences on both ends of the DNA. A limited-cycle PCR reaction was performed to amplify the insert DNA, which was then sequenced using Illumina MiSeq using 2×150 base pair chemistry. To help control for potential fragment recovery bias introduced from restriction enzyme digestion, a matched set of random control sites was generated taking into account the distances between experimental sites and the nearest restriction enzyme sequence used in the digestion, as previously described.[16] All samples were independently prepared and sequenced twice by different people with 13 months between preparations.

Integration is enriched in genes and satellite DNA

From two independent sequencing datasets, we obtained a total of 77 million reads across all 8 mice and 434 unique sites (**Supplementary Table S1**). To analyze integration events in particular genomic features, unique integration sites from all mice were pooled. **Figure 2a** shows a graphical representation of the frequency of unique integration events with respect to several genomic features (see also **Supplementary Table S2**). For each genomic feature, favoring or disfavoring of integration into the feature compared with random is represented by green or magenta coloring, respectively. In all, 43.6% of the LV-ARSA integration sites were found in the protein coding region of UCSC and RefSeq genes. These frequencies were slightly enriched compared with the matched random

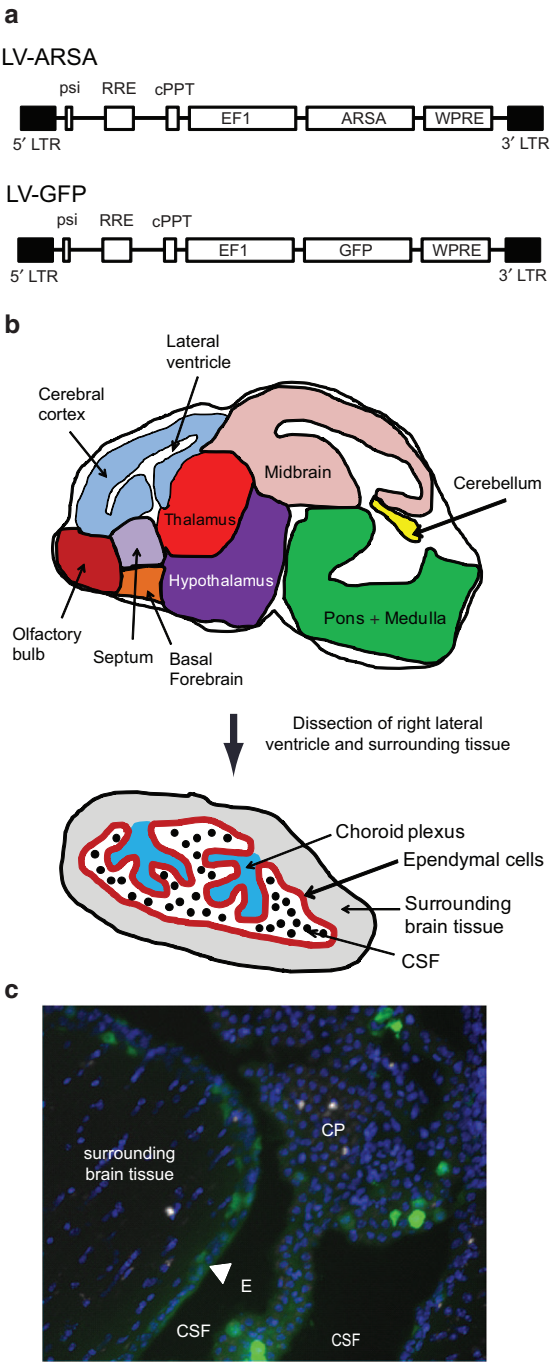


Figure 1 Lentiviral vector encoding human arylsulfatase A (LV-ARSA) vector and experimental design. (**a**) Schematic showing the various features of the LV-ARSA and LV-GFP vectors. (**b**) Diagram of the location of stereotactic intracerebral administration of the LV-ARSA vector in the murine brain and isolation of the ependyma of the choroid plexus. (**c**) Fluorescent microscopic image of an optical slice taken through the lateral ventricle of the murine brain 7 months after administration of LV-GFP. Green, green fluorescent protein (GFP); blue, Hoechst nuclear stain. V, ventricle; E (white triangles), ependyma; CSF, cerebral spinal fluid space; CP, choroid plexus.

controls by 1.17- and 1.20-fold, respectively ($P < 0.008$, χ^2 test). The single most prevalent feature of mammalian genomes is their repetitive sequences. Approximately 39%

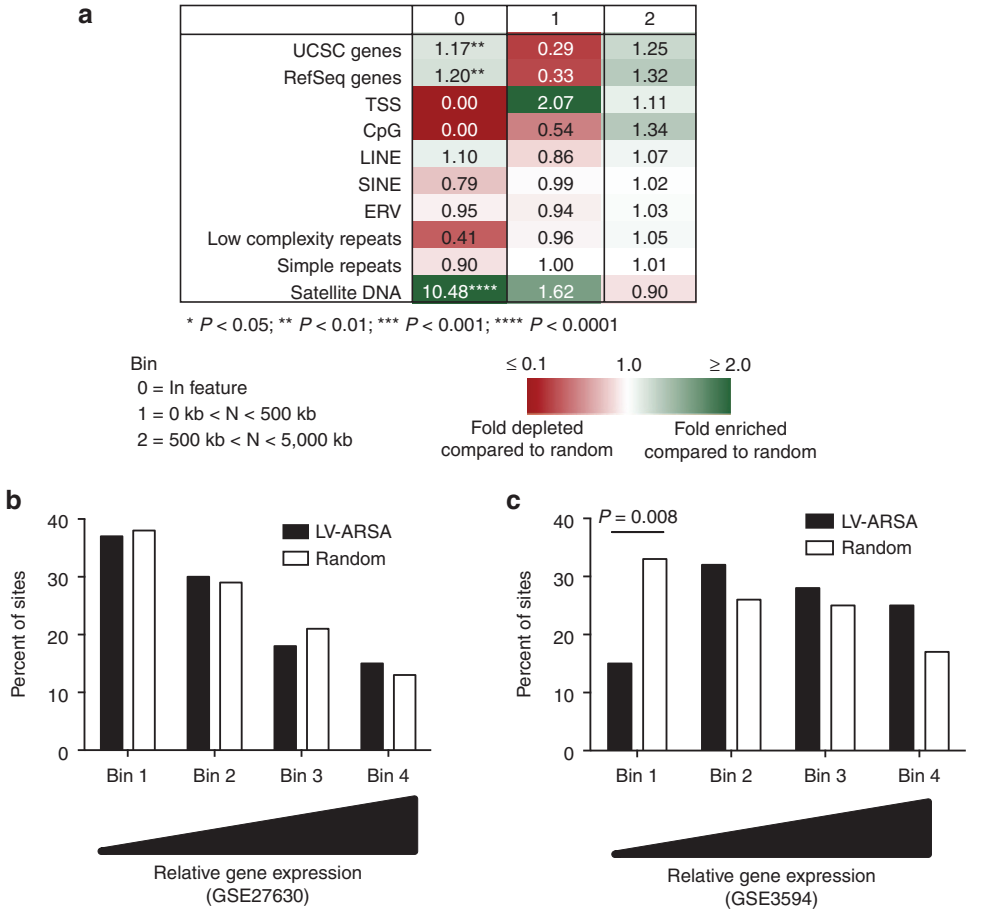


Figure 2 Distribution of unique integration sites in genomic features. Heat map illustrating the distribution of unique integration sites in genomic features. Numbers represent the fold-change in frequency of lentiviral vector encoding human arylsulfatase A (LV-ARSA) integration sites compared with random distribution. Favoring or disfavoring of integration in features is highlighted in various shades of green or magenta (respectively). Darker shades represent higher fold-changes in the frequency of integration. Significant differences are denoted by asterisks (*P < 0.05; **P < 0.01; ***P < 0.001; ****P < 0.0001) (χ^2 test). (**b and c**) Analysis of expression levels of genes targeted for integration by LV-ARSA. Gene expression levels in choroid plexus tissue from wild-type mice were obtained from published datasets GSE27630[25] (b) and GSE3594[24] (c). Roughly 45,000 and 12,000 genes were assayed (respectively) and distributed into four equal bins by relative expression levels. The bin with the lowest average expression is at the left and the highest expression is at the right. Genes used as integration targets by LV-ARSA were distributed into their corresponding bins based on their expression levels and summed.

of the murine genome consists of interspersed repeats (e.g., LINEs, SINEs, LTR elements, and satellite DNA), compared with ~46% in the human genome. Analysis of LV-ARSA integrations in interspersed repeats revealed that 0.69% of integrants were located in satellite DNA compared with 0.07% for the matched random control sites (10.5-fold increase) (P < 0.0001, χ^2 test) (**Figure 2a**). In contrast, integration sites were not significantly enriched in LINEs, SINEs, endogenous retroviral elements (ERVs), CpG islands, or low-complexity repeats.

For comparison, LV-ARSA infection of human 293T and HeLa cells showed that LV-ARSA favored integration into genes, LINEs, SINEs, ERVs, low-complexity repeats, and simple repeats (**Supplementary Tables S3 and S4**). Integration into satellite DNA was favored in both 293T and HeLa cells; however, significance was only achieved with HeLa cells (P < 0.0001, χ^2 test).

LV-ARSA does not favor integration in oncogenes or highly transcribed genes

Insertional oncogenesis by γ-retroviral gene therapy vectors has occurred in previous human gene therapy trials, contributing to the development of leukemia.[14,17–19] A major contributing factor to this adverse event is the preferential integration of γ-retroviral vectors at promoters, transcription start sites, and near genes controlling cell growth and proliferation.[20–22] Lentiviral vectors have not been previously found to favor such regions and are less likely to contribute to insertional oncogenesis. Consistent with previous findings with lentiviral vectors, LV-ARSA disfavored integration in transcription start sites (**Figure 2a** and **Supplementary Table S2**). Furthermore, none of the LV-ARSA sites were located in, or within 50 kb of, known murine cancer-related genes (the "Cancer Gene List" from http://www.bushmanlab.org/links/genelists).[23]

To determine if there was an influence of gene expression on LV-ARSA integration into genes, we examined two published gene expression profiling datasets of murine choroid plexus tissue (GSE27630 and GSE3594).[24,25] Genes were categorized into four different expression bins ranging from low- to high-level gene expression. Genes targeted for integration by LV-ARSA were distributed into the same bins based on their expression levels (**Figure 2b,c**). The distribution of genes targeted by LV-ARSA did not differ significantly from random distribution in either dataset, except for bin 1 (low-level expression) of the GSE3594 dataset where LV-ARSA sites were depleted compared with random (P=0.008, Fisher's exact test). Together, these data indicate that high-level gene expression did not influence LV-ARSA integration site targeting in murine ependymal cells. In contrast, integration of LV-ARSA in human 293T cells showed preference for highly expressed genes (**Supplementary Figure S1**), consistent with previous observations with HIV-1 integration in human cells.[26]

Regional hotspots for LV-ARSA integration

Regional genomic "hotspots" for retroviral vector integration have been reported in *in vitro* and *in vivo* datasets, including in HIV-1-infected individuals.[21,27–29] To determine if clustering of integration sites was evident in the dataset, the distribution of lengths of DNA segments between integration sites was compared with the distribution expected with random integration (**Figure 3**). Significantly more short intersegment distances were observed with the LV-ARSA sites compared with the random control sites, indicative of clustering (P < 0.0001, Fisher's exact test). Further inspection of the integration sites revealed that 66 genomic positions hosted two or more independent integration events, representing 42.9% of all integration sites (**Supplementary Table S5**). Thirty-seven of these 66 genomic regions shared the same integration site in two or more mice. Nineteen regions hosted 3 or more integration sites within 20 bases of each other. The genomic region (chr10: 9832769-9832909) located in an intron of the *stxbp5* gene was highly favored for integration. Eleven of the 434 (2.5%) independent integration sites were located in this region within 140 bases of each other. Closer inspection of the target DNA sequence for these 11 integration sites revealed that all 11 sites were located in or within 100 bases of a direct repeat sequence located at position chr10:9832707-9832930.

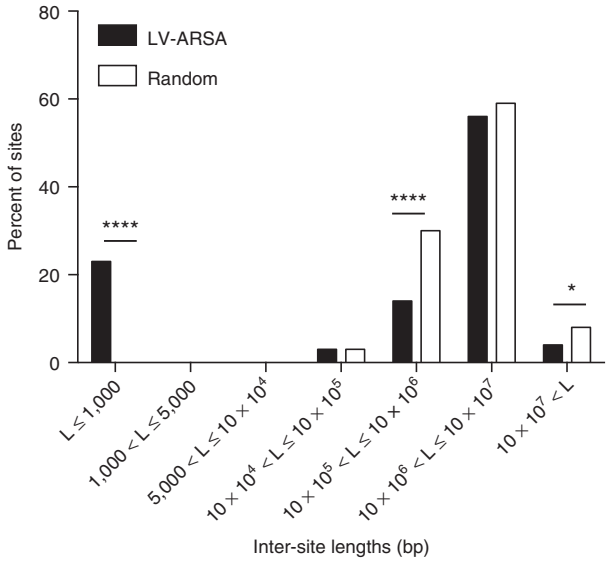


Figure 3 Analysis of integration site spacing on the murine genome. Integration site clustering was assessed by comparing the spacing between lentiviral vector encoding human arylsulfatase A (LV-ARSA) integration sites to the spacing between the same number of random sites. The lengths (L) of bases between integration sites were calculated and distributed into seven intersite length "bins," with the shortest intersite lengths to the left and the longest to the right. Significant differences are denoted by asterisks (****P < 0.0001, *P < 0.05; χ^2 test).

For comparison, analysis of 6,294 LV-ARSA integration sites in both 293T and HeLa cells showed that significantly more short intersegment distances were observed with the LV-ARSA sites compared with the random control sites, indicative of clustering (P < 0.0001, χ^2 test) (**Supplementary Figure S2**). From these integration sites, we randomly selected 434 sites to compare the number of regions hosting multiple integration sites with those obtained in the mouse dataset. In contrast with the mouse dataset, no identical integration sites were detected in the human dataset and only 4 integration sites were located within 20 bases of each other (**Supplementary Table S6**).

LV-ARSA integration target site selection is strongly influenced by flanking G-quadruplex-forming (G-quad) motifs

To investigate if the primary sequence at the integration target sites influenced target site selection, we extracted sequences flanking the LV-ARSA integration sites for further analysis. Forty bases upstream and downstream of each integration site (between −1 and +1) or the random control sites were aligned and the frequencies of A, C, G, and T at each position surrounding the integration sites were calculated. These values were compared with the expected value based on random control sites (**Figure 4** and **Supplementary Table S7**). As expected, the base composition at each position surrounding the matched random control sites varied little from the expected values[30] (**Supplementary Table S7**). The difference in frequency of each nucleotide at each position up to 40 bases upstream and downstream of the LV-ARSA integration sites was calculated (**Figure 4**). Inspection of the bases surrounding the LV-ARSA integration sites using QGRS Mapper (http://bioinformatics.ramapo.edu/QGRS/analyze.php) revealed two G-quad motifs flanking the integration site at positions −26 to −3 and +22 to +35.[31] Of note, positions +1 to +5 comprise the duplicated target site sequence after integration of lentiviruses.

LV-ARSA integration sites are enriched in or near non-B DNA-forming motifs

G-quad structures are stable DNA secondary structures that can form from motifs containing tracts of tandem guanines. These guanines hydrogen bond in a planar arrangement, forming stacks connected by single-stranded DNA loops.[32]

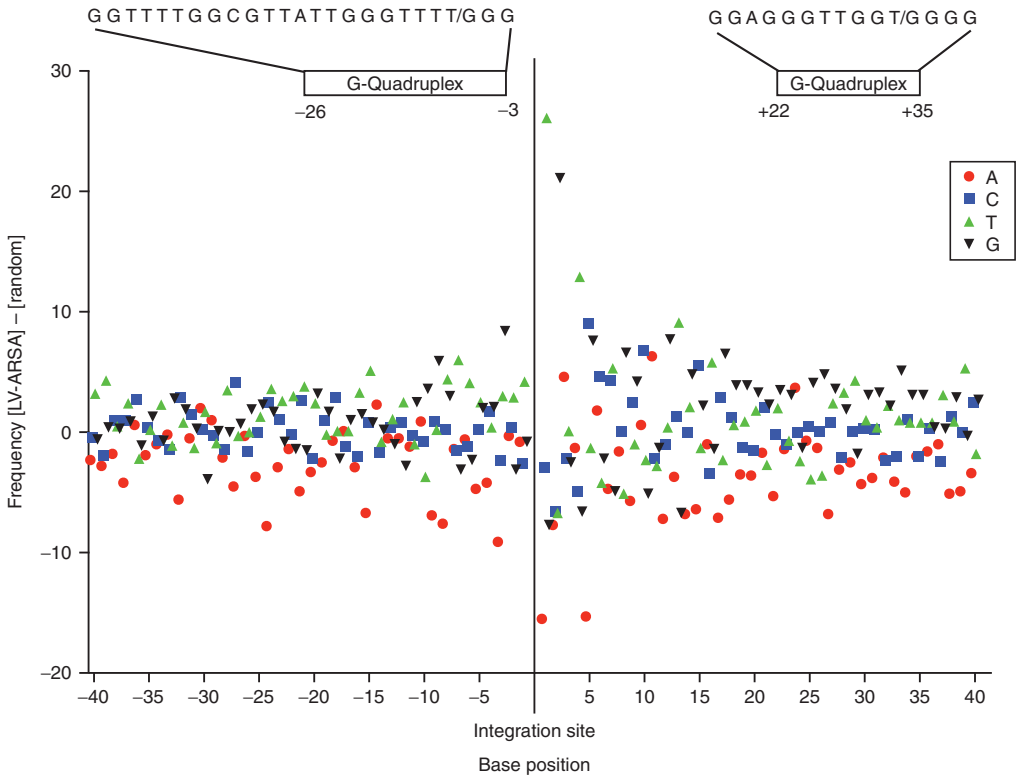


Figure 4 Consensus sequence at lentiviral integration sites. Base compositions of the top DNA strand at each position surrounding the integration sites were calculated. Integration occurs between positions −1 and +1 on the top strand. Base frequencies of nucleotides A (red circles), C (blue squares), T (green triangles), and G (black triangles) located 40 bases upstream and downstream of the integration site are plotted. The sequence and location of the flanking G-quadruplex-forming motifs are shown above the graph.

G-quad structures are only one of many (>10) non-B DNA-forming DNA secondary structures.[33] The Database for Integrated Annotations and Analysis of non-B DNA Forming Motifs (Non-B DB) provides the most complete list of alternative DNA structure predictions using the latest genome assemblies.[34,35] We used the Non-B DB to determine if LV-ARSA integration sites are enriched in or near G-quad motifs compared with the matched random control sites. LV-ARSA integration sites near G-quad motifs were enriched 1.28-fold compared with the matched random control sites ($P < 0.0001$, χ^2 test) (**Figure 5a** and **Supplementary Table S8**). We then asked if LV-ARSA integration sites are enriched in or near other non-B DNA-forming motifs. Analysis of integration site distributions revealed significant enrichment in short tandem repeats (1.16-fold; $P < 0.0001$, χ^2 test) and mirror repeats (1.56-fold; $P < 0.0001$, χ^2 test). Integration was also enriched in triplex motifs (1.36-fold), Z-DNA motifs (1.81-fold), direct repeats (1.14-fold), slipped motifs (1.77-fold), cruciform motifs (3.90-fold), and A-phased motifs (1.13-fold), although this enrichment did not reach statistical significance. These data show that LV-ARSA integration sites are highly enriched in and near non-B DNA-forming motifs.

We then asked if non-B DNA-forming motifs were also favored in a previous study of integration site analyses in murine neural tissue (brain striatum and eye), SC1 cells, and HC1 cells.[36] Integration was enriched in or near several non-B DNA motifs, most notably Z-DNA motifs, direct repeats, inverted repeats, mirror repeats, slipped motifs, and cruciform motifs (**Figure 5b-e** and **Supplementary Table S8**). Together, these data suggest that lentiviral integration is enriched in or near non-B DNA motifs in murine cells.

HIV-1 integration sites are enriched in non-B DNA-forming motifs in human cells

Next, we analyzed a variety of HIV-1 integration site datasets in diverse human cell types 293T,[37] Jurkat,[37,38] macrophage,[16] HeLa,[20,39] IMR90 (refs. 40,41), PBMC,[41] HOS,[37] H9 (ref. 20), and SupT1 (refs. 29,42) to determine if integration sites were enriched in non-B DNA-forming motifs (**Supplementary Table S9**). Integration in or near multiple non-B DNA-forming motifs was highly enriched in each of the datasets (**Figure 6** and **Supplementary Table S10**). Integration in or near several non-B DNA motifs was strongly favored among the different datasets and preference for specific types of non-B DNA-forming motifs varied for each cell type, possibly indicating cell-type-specific effects.

Discussion

Here we have presented an analysis of the distribution of integration sites of a lentiviral gene therapy vector (LV-ARSA) in murine brain choroid plexus and ependymal cells. LV-ARSA was stereotactically delivered into the right lateral ventricle of ARSA−/− and wild-type C57Bl6 mice and demonstrated a clinically favorable integration site profile. LV-ARSA showed

a

Ratio of integration sites to control sites

LV-ARSA	0	1	2	3	4
Triplex motifs	1.36	1.21	0.95	1.01	1.02
KV Z-DNA motifs	0.00	0.90	0.77**	1.08**	0.99
Z-DNA motifs	1.81	1.01	1.05	0.95	1.44
Short Tandem repeats	1.16****	0.97	1.03	0.00	1.60
Direct repeats	1.14	1.10	0.94	0.18	1.60
Mirror repeats	1.56*	1.00	0.96	0.00	1.60
G-quadruplex-forming repeats	0.73	1.28*	1.01	0.75*	1.58
Slipped motifs	1.77	1.18	0.97	0.83	1.58
Cruciform motifs	3.90	1.39	1.01	0.91*	2.12****
A Phased motifs	1.13	1.05	0.92	1.13	1.38
Inverted repeats	1.00	1.01	0.95	0.00	1.60

b

Ratio of integration sites to control sites

Bartholomae Brain	0	1	2	3	4
Triplex motifs	0.00	0.95	1.04	0.94	1.08
KV Z-DNA motifs	0.00	0.99	0.97	1.16	1.08
Z-DNA motifs	0.00	1.08	1.01	1.22	1.06
Short Tandem repeats	0.13*	1.08	1.11	1.11	0.97
Direct repeats	0.14*	1.21*	1.27*	0.99	0.75
Mirror repeats	0.39	1.15	1.15	0.95	0.91
G-quadruplex-forming repeats	0.55	0.88	1.00	0.97	1.17
Slipped motifs	0.00	0.95	1.08	1.07	0.98
Cruciform motifs	0.00	1.14	0.79	1.04	0.99
A Phased motifs	0.89	1.18	0.89	0.91	1.00
Inverted repeats	0.76	1.01	1.04	1.19	0.90

c

Ratio of integration sites to control sites

Bartholomae Eye	0	1	2	3	4
Triplex motifs	0.37	0.98	0.86	1.04	1.00
KV Z-DNA motifs	0.00	0.99	1.07	1.00	1.08
Z-DNA motifs	0.00	1.32****	0.95	0.97	0.83
Short Tandem repeats	0.35*	1.06	1.08	1.08	1.04
Direct repeats	0.22**	1.01	0.95	1.23**	1.00
Mirror repeats	0.36**	1.01	1.02	1.07	1.04
G-quadruplex-forming repeats	0.00	0.82*	1.02	1.11	1.04
Slipped motifs	0.20	0.99	1.31***	0.87	0.91
Cruciform motifs	1.47	1.09	1.03	1.09	0.90
A Phased motifs	0.47	1.05	1.04	0.96	1.01
Inverted repeats	1.16	1.15*	1.13	0.97	0.84

d

Ratio of integration sites to control sites

Bartholomae SC1	0	1	2	3	4
Triplex motifs	0.00	0.76**	1.03	0.73**	0.85
KV Z-DNA motifs	0.00	1.04	0.79	0.85	0.82
Z-DNA motifs	0.00	0.91	0.92	1.14	0.98
Short Tandem repeats	0.76	1.01	1.14	0.81	0.81
Direct repeats	0.90	1.05	0.92	0.94	0.77
Mirror repeats	1.03	1.08	0.89	0.93	0.63**
G-quadruplex-forming repeats	0.00	0.95	0.80	0.91	0.97
Slipped motifs	0.31	0.81*	0.94	1.24*	0.85
Cruciform motifs	2.28	0.91	0.80	1.05	0.98
A Phased motifs	0.00	0.91	1.19	0.94	0.83
Inverted repeats	1.18	1.08	0.91	0.79*	0.77

e

Ratio of integration sites to control sites

Bartholomae HC1	0	1	2	3	4
Triplex motifs	0.00	1.17	0.96	0.71*	0.91
KV Z-DNA motifs	0.00	0.95	1.37**	0.72*	0.78
Z-DNA motifs	0.00	0.90	1.15	0.92	1.14
Short Tandem repeats	0.52	1.11	1.06	0.78	0.86
Direct repeats	0.40	0.95	0.90	0.88	1.25
Mirror repeats	0.57	0.88	1.52****	0.89	0.64**
G-quadruplex-forming repeats	0.53	1.13	0.74*	0.85	0.83
Slipped motifs	0.37	1.04	0.78	1.05	0.81
Cruciform motifs	2.71	1.22*	0.77	0.95	0.89
A Phased motifs	0.86	0.87	1.20	0.84	0.89
Inverted repeats	0.74	0.89	1.07	1.25*	0.70*

* $P \leq 0.05$; ** $P \leq 0.01$; *** $P \leq 0.001$; **** $P \leq 0.0001$

≤ 0.1 1.0 ≥ 2.0

Fold depleted compared to random — Fold enriched compared to random

Figure 5 Distribution of unique integration sites in non-B DNA-forming motifs. Heat map illustrating the distribution of unique integration sites in and near non-B DNA-forming motifs. The ratios of integration sites to matched random control sites are shown for (**a**) lentiviral vector encoding human arylsulfatase A (LV-ARSA) in murine ependymal cells, (**b**) murine brain striatum (Bartholomae dataset), (**c**) murine eye tissue (Bartholomae dataset), (**d**) *ex vivo*-transduced fibroblasts (SC1) (Bartholomae dataset), and (**e**) bone-marrow-derived hematopoietic progenitor cells (HC1) (Bartholomae dataset). Darker shades represent higher fold-changes in the ratio of integration sites to matched random control sites. Bin 0 = number of integration sites (*N*) located within features; Bin 1 = 0 < $N \leq 0.5$ kb; Bin 2 = 0.5 kb < $N \leq 5$ kb; Bin 3 = 5 kb < $N \leq$ 50 kb; Bin 4= 50 kb < $N \leq 500$ kb. Significant differences are denoted by asterisks (*$P < 0.05$; **$P < 0.01$; ***$P < 0.001$; ****$P < 0.0001$) (χ^2 test).

no bias for integration in known murine oncogenes or near transcription start sites of genes, and did not show a bias for integration in transcriptionally active genes. LV-ARSA did show a strong bias for integration in large arrays of tandemly repeating, noncoding DNA such as satellite DNA. Integration site hotspots were identified and target site selection was strongly influenced by flanking G-quad sequence motifs. In addition, we demonstrated that integration in or near several non-B DNA-forming motifs is highly favored by LV-ARSA and HIV-1 in diverse cell types.

Previous analyses of integration site distribution of HIV-1 and other lentiviral vectors revealed that transcriptionally active genes are preferred targets for integration in human cells.[26] Many of these datasets were obtained from *in vitro*- and *ex vivo*-transduced cells. In the present *in vivo* study, analysis of lentiviral vector integration site distribution in murine brain ependymal cells revealed little bias for integration in transcriptionally active genes. This result is consistent with a different lentiviral vector integration site dataset in murine eye and brain striatum tissues.[36] In that study, low expression levels of *Psip1*/LEDGF/p75 were proposed as a potential mechanism for reduced integration into genes in murine eye and brain striatum tissues. In choroid plexus tissue, *Psip1/LEDGF/p75* expression level is moderate to high and therefore not likely a major factor for reduced integration into active genes in choroid plexus tissue (data not shown). The lentivector itself is not likely a factor since LV-ARSA favored integration into active genes

Ratio of integration sites to control sites

293T	0	1	2	3	4
Triplex motifs	0.75	1.07	1.20***	1.05	1.06
KV Z-DNA motifs	3.42	1.15****	1.22****	1.03	0.86**
Z-DNA motifs	1.37	1.11***	1.02	1.08	1.06
Short Tandem repeats	0.93	1.12***	1.16***	1.02	0.96
Direct repeats	0.98	1.09**	1.12**	0.99	1.04
Mirror repeats	0.89	1.07*	1.01	1.08*	1.07
G-quadruplex-forming repeats	0.71	1.22****	1.08	0.93	0.94
Slipped motifs	0.53	1.14****	1.03	1.07	1.04
Cruciform motifs	0.79	1.04	1.03	0.99	1.21****
A Phased motifs	0.69	0.93*	1.06	1.02	1.23****
Inverted repeats	1.10	1.07	1.07	1.04	1.09

PBMCs	0	1	2	3	4
	2.88	1.06	1.29**	1.06	0.88
	0.00	1.09	1.02	1.19*	1.00
	0.00	0.98	1.03	1.11	1.18
	1.30	1.12	1.04	1.11	0.95
	1.34	1.16*	1.12	1.00	0.99
	0.86	1.10	1.06	1.03	1.11
	0.55	1.19**	1.02	0.93	1.05
	3.01*	1.04	1.15	1.12	0.87
	1.50	1.04	1.10	0.92	1.15
	2.12	1.04	1.12	1.05	1.02
	0.78	1.05	0.95	1.32***	0.98

Jurkat	0	1	2	3	4
Triplex motifs	0.56	1.17****	1.07	0.99	0.97
KV Z-DNA motifs	0.00	1.31****	1.15**	0.94	0.81**
Z-DNA motifs	1.02	1.07	1.14*	1.00	1.07
Short Tandem repeats	0.98	1.14**	1.05	1.07	0.97
Direct repeats	0.94	1.14**	1.00	1.10*	1.02
Mirror repeats	1.04	1.07	1.10	1.01	1.07
G-quadruplex-forming repeats	0.86	1.35****	1.04	0.93	0.84*
Slipped motifs	0.59	1.15***	1.06	0.97	1.18**
Cruciform motifs	1.18	0.90*	0.99	1.14**	1.23***
A Phased motifs	0.21	0.88**	0.92	1.03	1.29****
Inverted repeats	1.19	0.99	1.24****	0.99	1.03

HOS	0	1	2	3	4
	0.00	1.11	1.08	1.10	0.99
	0.00*	1.24**	1.11	1.07	0.84
	0.00	0.94	1.15	1.16	1.05
	0.73	1.23**	1.04	1.00	1.06
	1.03	1.30***	1.05	0.80	1.04
	1.11	1.22**	0.90	1.16	0.95
	2.51	1.27***	1.11	0.88	0.92
	0.00	1.19*	0.93	1.17	0.97
	0.00	0.99	0.97	1.04	1.24
	2.44	0.84	0.98	1.13	1.09
	1.07	0.94	1.10	0.99	1.33*

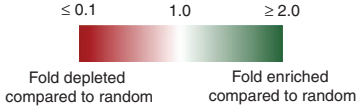
Macrophage	0	1	2	3	4
Triplex motifs	0.93	1.04	0.99	1.06	1.23**
KV Z-DNA motifs	0.00	1.16**	1.19**	0.91	1.00
Z-DNA motifs	1.13	1.01	1.10	1.01	1.23**
Short Tandem repeats	1.39	1.10	1.06	1.08	1.03
Direct repeats	1.21	1.08	1.23**	0.91	1.11
Mirror repeats	0.70	1.04	1.18*	1.11	1.08
G-quadruplex-forming repeats	1.06	1.08	0.90	1.05	1.28
Slipped motifs	0.00	1.06	1.09	1.10	1.08
Cruciform motifs	0.00	1.12*	0.97	0.98	1.11
A Phased motifs	1.03	1.07	0.99	0.98	1.17*
Inverted repeats	0.88	1.21***	1.01	1.01	1.03

H9	0	1	2	3	4
	0.00	1.04	1.01	1.08	1.26
	0.00**	1.30*	1.36*	0.83	0.88
	0.00	0.86	0.76	1.37*	1.44*
	1.67	0.86	1.23	1.19	1.06
	0.94	1.26*	0.78	1.22	1.13
	0.76	1.09	1.16	1.17	0.79
	1.92	1.16	0.98	1.01	1.09
	1.77	1.43**	0.72	0.97	1.15
	0.00	1.12	0.95	1.12	0.93
	0.00	1.34**	0.90	0.96	0.88
	1.44	1.05	1.47**	0.96	0.79

HeLa	0	1	2	3	4
Triplex motifs	0.96	1.18***	1.20**	0.96	0.92
KV Z-DNA motifs	0.00	1.34****	1.09	0.99	
Z-DNA motifs	1.16	1.17**	0.97	1.05	1.01
Short Tandem repeats	0.71	1.20***	1.09	1.08	0.92
Direct repeats	1.24	1.19***	1.14	1.02	0.90
Mirror repeats	1.53*	1.20***	1.06	1.03	0.85
G-quadruplex-forming repeats	0.73	1.41****	0.99	0.90	0.80
Slipped motifs	0.67	1.20***	1.13	1.02	0.97
Cruciform motifs	2.00	0.98	1.06	1.01	1.15
A Phased motifs	1.06	0.91	0.96	1.04	1.22*
Inverted repeats	1.15	1.01	1.12	0.99	1.19*

SupT1	0	1	2	3	4
	2.65	1.09	1.13	1.21*	0.97
	0.00	1.31****	1.21*	0.97	0.71**
	0.00	1.17**	1.08	1.00	0.91
	1.31	1.16*	1.03	1.23**	0.84
	1.23	1.23***	1.17	0.94	0.96
	0.60	1.07	1.07	1.14	1.09
	1.51	1.38****	1.03	0.91	0.83
	0.93	1.20**	1.24**	1.04	0.73**
	2.77	0.95	0.86	0.91	1.50****
	1.47	0.88	0.93	1.00	1.25*
	1.02	0.96	1.02	1.08	1.35***

IMR90	0	1	2	3	4
Triplex motifs	1.17	1.15*	1.04	1.17*	0.77*
KV Z-DNA motifs	0.00	1.15*	1.16*	0.86	1.17
Z-DNA motifs	1.42	1.01	1.06	1.03	1.13
Short Tandem repeats	1.74*	1.07	1.01	1.11	1.03
Direct repeats	0.87	1.16**	0.98	1.06	1.00
Mirror repeats	1.23	1.16*	1.02	1.14	0.82
G-quadruplex-forming repeats	2.23	1.06	0.94	1.09	1.16
Slipped motifs	2.05	1.22***	0.87	1.07	1.10
Cruciform motifs	2.45	0.99	1.14	0.98	1.11
A Phased motifs	0.86	0.98	0.96	1.11	1.13
Inverted repeats	1.34	1.12	0.94	1.02	1.01



* $P \leq 0.05$; ** $P \leq 0.01$; *** $P \leq 0.001$; **** $P \leq 0.0001$

Figure 6 Distribution of unique HIV-1 integration sites in non-B DNA-forming motifs. Heat map illustrating the distribution of unique HIV-1 integration sites in and near non-B DNA-forming motifs. The ratios of HIV-1 integration sites to matched random control sites are shown for the following previously published datasets: 293T, PBMCs, Jurkat, HOS, macrophage, H9, HeLa, SupT1, and IMR90. Darker shades represent higher fold-changes in the ratio of integration sites to matched random control sites. Bin 0 = number of integration sites (*N*) located within features; Bin 1= 0 < *N* ≤ 0.5 kb; Bin 2 = 0.5 kb < *N* ≤ 5 kb; Bin 3 = 5 kb < *N* ≤ 50 kb; Bin 4 = 50 kb < *N* ≤ 500 kb. Significant differences are denoted by asterisks (**P* < 0.05; ***P* < 0.01; ****P* < 0.001; *****P* < 0.0001) (χ^2 test).

under similar conditions in human cells, as previously established in the field.

LV-ARSA exhibited a strong bias for integration in tandem repeats (e.g., satellite DNA) in murine ependymal cells. Preferred integration into satellite DNA has been previously described for microinjected tandem repeated DNA, latent proviral HIV-1, and murine eye tissue.[36,38,43] In contrast with integration in murine brain ependymal cells, lentiviral vector

integration in murine brain striatum was not enriched in satellite DNA.[36] The reasons for this biased integration in satellite DNA in different regions of the brain are unknown. It is possible that the size of the datasets, vector differences, differential expression of host proteins in different regions of the brain, or other unknown factors influence integration site placement.[44–46]

Earlier *in vitro* integration site assays have shown that the primary sequence immediately flanking (~5–10 bases) the integration-target DNA has minor influences on site selection, with only weak target site consensus motifs identified.[42,47–51] In the present *in vivo* study, we looked at up to 40 bases flanking integration sites and identified flanking G-quad sequence motifs as a potential factor that influences site selection by flanking integration sites. Many DNA repeat sequences can exist in at least two distinct conformations. Typically, DNA sequences adopt a right-handed B-form with Watson–Crick base pairing. However, at least 10 non-B DNA conformations exist. These conformations have contorted bond angles or unpaired nucleotides compared with the orthodox right-handed B-form. Non-B DNA conformations are in higher energy states and are believed to be facilitated at specific sequence motifs by the free energy generated from negative supercoiling, which can arise from processes such as transcription, protein binding, and other factors.[52–54] Several non-B DNA motifs preferentially act as the recipient of genetic information during gene conversion events. For example, repeating units of the TG dinucleotide is characteristic of the non-B DNA-forming Z-DNA motif.[55] Substrates containing Z-DNA motifs preferentially act as the recipient of genetic information during gene conversion events.[56] Moreover, Z-DNA motifs have also been shown to stimulate homologous recombination up to 20-fold in human cells. Another example is G-quad motifs, which can promote recombination and possibly influence genomic stability and cellular processes such as transcription (reviewed in ref. 32). It is possible that lentiviral integration complexes are recruited to genomic regions that favor genetic recombination such as non-B DNA motifs via direct interactions with the motif itself or via non-B DNA motif-binding proteins.

LV-ARSA integration sites were also enriched in or near a variety of other non-B DNA-forming motifs with some integration site bias towards the type of non-B DNA-forming motif. For example, LV-ARSA exhibited a strong bias for integration into all of the non-B DNA-forming sequences except for G-quad repeats and inverted repeats, where integration was instead enriched near G-quad repeats, rather than within them. Integration site bias for certain non-B DNA-forming motifs was also observed in other published datasets with HIV-1 in various cell types, indicating that cell-type-specific factors associated with non-B DNA-forming motifs, or the genomic structures they produce, contribute to this bias. Specific host proteins that bind to specific non-B DNA-forming motifs have been identified; therefore it is conceivable that such factors could recruit preintegration complexes and/or influence integration in or near these motifs. Notably, Psip1/LEDGF/p75, which is a co-activator of general transcription and influences the location of HIV-1 integration, was recently shown to selectively bind negatively supercoiled DNA over unconstrained DNA.[37,57] It will be interesting to learn if non-B DNA plays a role in attracting Psip1/LEDGF/p75, which then recruits HIV-1 preintegration complexes. Although several of the datasets showed that integration was disfavored directly within certain non-B DNA-forming motifs, integration was significantly enriched near these motifs. Host proteins/complexes that bind non-B DNA-forming motifs may recruit lentiviral preintegration complexes, but block integration within these motifs via steric hindrance. As a result, integration near these motifs may be favored instead.

Consistent with enrichment of integration sites in or near non-B DNA-forming motifs, we observed several integration site hotspots including a particularly strong integration hotspot in a direct repeat element on chromosome 10. It is unclear why this region was targeted multiple times in one mouse and only a few times in another mouse. Several additional integration sites were detected in this region but were omitted from our analysis since they fell within 10 bases of each other due to potential errors in sequencing typically seen using this sequencing platform. Therefore, it is possible that the number of sites in this region is underrepresented and requires additional characterization. Since DNA repeats can vary in abundance and length in vertebrate genomes, it is also possible that the underlying repeats of non-B DNA-forming elements are polymorphic among the mice and that a polymorphism in this direct repeat in one mouse enhanced LV-ARSA integration events in this region.[34,35,53,58] Clustering of lentiviral integration sites has been observed previously in human datasets (reviewed in ref. 26). We also showed that LV-ARSA integration sites clustered in both mouse and human cells, suggesting that clustering is not species specific. In addition, we identified LV-ARSA clustering of integration sites in the human genes CREB-binding-protein (*CREBBP*) and DNA (cytosine-5-)methyltransferase-1 (*DNMT1*), which were previously identified as integration hotspots in human cells.[20,29,41] These data also further support the fact that LV-ARSA integration clustering is not an artifact of the method used. Interestingly, the integration sites in *CREBBP* and *DNMT* are all located adjacent to non-B DNA motifs.

Analyses of published integration site datasets thus far have shown that no single factor dictates lentiviral integration site targeting/placement and that multiple factors are likely involved. For example, target DNA structure, curvature, flexibility, rigidity in solution, and distortion within the nucleosome core all influence the frequency and placement of integration.[49,59] Host proteins such as Psip1/LEDGF/p75 that can tether HIV-1 preintegration complexes to target DNA also influence integration site placement.[37,60] Conversely, DNA-binding proteins can create regions that are refractory to integration via steric hindrance.[60–63] In conclusion, our data demonstrate a clinically favorable integration site profile of a lentiviral vector encoding the *ARSA* transgene in the murine brain, and identify non-B DNA-forming motifs as a potential new host genomic feature that influences the distribution of lentiviral integration sites.

Materials and methods

Lentivector construction. Codon-optimized cDNA of the human *arylsulfatase A* (*ARSA*) gene, or the *green fluorescent protein* (*GFP*) gene as a control, were independently

cloned into a self-inactivating third-generation HIV-based vector downstream of the human EF1-alpha promoter to generate the LV-ARSA or LV-GFP vectors, respectively. LV-ARSA and LV-GFP were generated using the ViraPower™ Lentiviral Packaging Mix (Life Technologies). 293FT cells at 80% confluency were transfected with pLV-ARSA (or pLV-GFP), pLP1, pLP2, and pLP-VSV-G transfected with TurboFectin8.0 transfection reagent (Origene). Cell supernatant was harvested after 36 hours of transfection, clarified by centrifugation at 500 × *g* for 10 minutes, and filtered (0.45 µm). Virus-containing medium was concentrated at 31,000 × *g* for 5 hours.

Stereotactical brain injection and isolation of genomic DNA from ependymal cells. Animal experiments were performed in compliance with the guidelines set by the Canadian Council for Animal Care and the policies and procedures approved by the University of Western Ontario Council on Animal Care (protocol 2007-044-12). Four-month-old C57Bl6 WT (two males and one female) and C57Bl6 ARSA–/– (two males and two females) were used in this study. Mice were anesthetized with ketamine and xylazine and placed in a Kopf stereotaxic apparatus. Sterile surgical procedures were followed to expose the injection site located at AP –0.2, ML 1.0, DV 3.0 according to mouse brain stereotaxic coordinates (The Mouse Brain in Stereotactic Coordinates, Second Edition, George Paxinos and Keith B. J. Franklin, Academic Press, 2001). A burr hole was placed followed by an injection of 10 µl of LV-ARSA vector (2.5×109 TU/ml) over 20 minutes. The needle was left in place for 5 minutes and then slowly withdrawn. The number of LV-ARSA transducing units was approximated by titering LV-EF1a-GFP vector on HT1080 cells (generated in parallel with LV-ARSA). Seven days postinjection, the region surrounding the lateral ventricle containing the choroid plexus and ependymal cells from each murine was removed by dissection. Subsequently, genomic DNA was isolated from the ependyma using the DNeasy Blood & Tissue Kit according to the manufacturer's instructions (Qiagen).

Isolation of integration sites. Genomic DNA from the dissected tissue from each mouse was restriction enzyme digested using MseI and NarI and amplified by ligation-mediated PCR, as previously described step by step.[64] After gel purification of the PCR products, the purified DNA samples were processed using the Nextera XT DNA Sample Preparation kit, which uses an engineered transposome to simultaneously fragment and tag input DNA with unique adapter and index ("barcodes") sequences on both ends of the DNA. DNA from each murine was barcoded. A limited-cycle PCR reaction was performed to amplify the insert DNA, which was then sequenced using Illumina MiSeq using 2×150 bp chemistry at the London Regional Genomics Centre (Robarts Research Institute, Western University, Canada).

Bioinformatic analyses. Integration site sequences were judged to be of acceptable quality if (1) the match to the genome began within 3 bp of the 5′-CA-3′ terminus of the viral DNA, (2) the match proximal to the LTR end showed an identity of at least 98%, (3) the match yielded a unique best hit using default parameters in the client-server BLAT ranking, as previously performed,[16] and (4) the integration site did not fall within 10 bases or less of another integration site (due to potential sequencing errors known to arise with this sequencing platform). Bowtie2 was used for aligning sequence reads to the genome and BedTools was used for computing distances between the sites and genomic features.[65,66] An integration site was scored as present in a transcription unit if it was mapped in DNA between the base pairs encoding the 5′ and 3′ ends of the transcribed region as specified in the various gene catalog annotations (http://www.genome.ucsc.edu/cgi-bin/hgGateway). Matched random control integration sites (28,800 in total) were generated by matching each experimentally determined site with 50 random sites *in silico* that were constructed to be the same number of bases from the restriction site as was the experimental site, as previously described.[16] Mapping of integration sites to non-B DNA motifs was done using the Non-B DB for species Murine 37.1 or Human 37.1 (http://nonb.abcc.ncifcrf.gov/apps/site/default).[34,35] Identification of all common clones between animals was not validated and could have resulted from cross-contamination during DNA processing. As a result, it is possible that the number of unique sites reported in this study is artificially inflated.

Supplementary material

Figure S1. Analysis of expression levels of genes in human cells targeted for integration by lentiviral vector encoding human arylsulfatase A (LV-ARSA).
Figure S2. Analysis of integration site spacing on the murine genome. Integration site clustering was assessed by comparing the spacing between lentiviral vector encoding human arylsulfatase A (LV-ARSA) integration sites to the spacing between the same number of random sites.
Table S1. Summary of integration site distribution in murine ependymal cells.
Table S2. Summary of LV-ARSA integration site distribution in various murine genomic features.
Table S3. Summary of pooled LV-ARSA and LV-GFP integration site distribution in various human genomic features.
Table S4. List of pooled LV-ARSA and LV-GFP integration sites in human 293T and HeLa cells.
Table S5. Murine genomic regions hosting two or more independent integration sites.
Table S6. Human genomic regions in 293T and HeLa cells hosting multiple independent integration events out of 434 randomly selected sites.
Table S7. Frequncy of each nucleotide up to 40 bases upstream and downstream of the LV-ARSA integration sites or matched random sites.
Table S8. Distribution of integration sites in non-b DNA-forming motifs in murine cells.
Table S9. Summary of the human integration site datasets used in this study.
Table S10. Distribution of integration sites in non-B DNA-forming motifs in various human cells.

Acknowledgments. This work was supported in part by Bethany's Hope Foundation to C.A.R.; an Ontario HIV Treatment Network salary award to S.D.B; a Canadian Institutes

of Health Research (CIHR) grant (MOP317646) to S.D.B.; and a CIHR Doctoral Research Award in partnership with the Canadian Association for HIV Research (CAHR). We thank Robert Hegele, John Robinson, and David Carter (Robarts Research Institute, Western University) for helpful discussions and sequencing support.

1. Peng, L and Suzuki, K (1987). Ultrastructural study of neurons in metachromatic leukodystrophy. *Clin Neuropathol* **6**: 224–230.
2. Sevin, C, Aubourg, P and Cartier, N (2007). Enzyme, cell and gene-based therapies for metachromatic leukodystrophy. *J Inherit Metab Dis* **30**: 175–183.
3. Sevin, C, Benraiss, A, Van Dam, D, Bonnin, D, Nagels, G, Verot, L *et al.* (2006). Intracerebral adeno-associated virus-mediated gene transfer in rapidly progressive forms of metachromatic leukodystrophy. *Hum Mol Genet* **15**: 53–64.
4. Sevin, C, Verot, L, Benraiss, A, Van Dam, D, Bonnin, D, Nagels, G *et al.* (2007). Partial cure of established disease in an animal model of metachromatic leukodystrophy after intracerebral adeno-associated virus-mediated gene transfer. *Gene Ther* **14**: 405–414.
5. Naldini, L, Blömer, U, Gallay, P, Ory, D, Mulligan, R, Gage, FH *et al.* (1996). *In vivo* gene delivery and stable transduction of nondividing cells by a lentiviral vector. *Science* **272**: 263–267.
6. Biffi, A, De Palma, M, Quattrini, A, Del Carro, U, Amadio, S, Visigalli, I *et al.* (2004). Correction of metachromatic leukodystrophy in the mouse model by transplantation of genetically modified hematopoietic stem cells. *J Clin Invest* **113**: 1118–1129.
7. Biffi, A, Capotondo, A, Fasano, S, del Carro, U, Marchesini, S, Azuma, H *et al.* (2006). Gene therapy of metachromatic leukodystrophy reverses neurological damage and deficits in mice. *J Clin Invest* **116**: 3070–3082.
8. 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.
9. Kobayashi, Y, Watanabe, M, Okada, Y, Sawa, H, Takai, H, Nakanishi, M *et al.* (2002). Hydrocephalus, situs inversus, chronic sinusitis, and male infertility in DNA polymerase lambda-deficient mice: possible implication for the pathogenesis of immotile cilia syndrome. *Mol Cell Biol* **22**: 2769–2776.
10. Taulman, PD, Haycraft, CJ, Balkovetz, DF and Yoder, BK (2001). Polaris, a protein involved in left-right axis patterning, localizes to basal bodies and cilia. *Mol Biol Cell* **12**: 589–599.
11. Brody, SL, Yan, XH, Wuerffel, MK, Song, SK and Shapiro, SD (2000). Ciliogenesis and left-right axis defects in forkhead factor HFH-4-null mice. *Am J Respir Cell Mol Biol* **23**: 45–51.
12. Roth, Y, Kimhi, Y, Edery, H, Aharonson, E and Priel, Z (1985). Ciliary motility in brain ventricular system and trachea of hamsters. *Brain Res* **330**: 291–297.
13. Sands, MS and Haskins, ME (2008). CNS-directed gene therapy for lysosomal storage diseases. *Acta Paediatr Suppl* **97**: 22–27.
14. Hacein-Bey-Abina, S, Von Kalle, C, Schmidt, M, McCormack, MP, Wulffraat, N, Leboulch, P *et al.* (2003). LMO2-associated clonal T cell proliferation in two patients after gene therapy for SCID-X1. *Science* **302**: 415–419.
15. Hacein-Bey-Abina, S, von Kalle, C, Schmidt, M, Le Deist, F, Wulffraat, N, McIntyre, E *et al.* (2003). A serious adverse event after successful gene therapy for X-linked severe combined immunodeficiency. *N Engl J Med* **348**: 255–256.
16. Barr, SD, Ciuffi, A, Leipzig, J, Shinn, P, Ecker, JR and Bushman, FD (2006). HIV integration site selection: targeting in macrophages and the effects of different routes of viral entry. *Mol Ther* **14**: 218–225.
17. Hacein-Bey-Abina, S, Garrigue, A, Wang, GP, Soulier, J, Lim, A, Morillon, E *et al.* (2008). Insertional oncogenesis in 4 patients after retrovirus-mediated gene therapy of SCID-X1. *J Clin Invest* **118**: 3132–3142.
18. Wang, GP, Garrigue, A, Ciuffi, A, Ronen, K, Leipzig, J, Berry, C *et al.* (2008). DNA bar coding and pyrosequencing to analyze adverse events in therapeutic gene transfer. *Nucleic Acids Res* **36**: e49.
19. Stein, S, Ott, MG, Schultze-Strasser, S, Jauch, A, Burwinkel, B, Kinner, A *et al.* (2010). Genomic instability and myelodysplasia with monosomy 7 consequent to EVI1 activation after gene therapy for chronic granulomatous disease. *Nat Med* **16**: 198–204.
20. Wu, X, Li, Y, Crise, B and Burgess, SM (2003). Transcription start regions in the human genome are favored targets for MLV integration. *Science* **300**: 1749–1751.
21. Cattoglio, C, Facchini, G, Sartori, D, Antonelli, A, Miccio, A, Cassani, B *et al.* (2007). Hot spots of retroviral integration in human CD34+ hematopoietic cells. *Blood* **110**: 1770–1778.
22. Wang, GP, Berry, CC, Malani, N, Leboulch, P, Fischer, A, Hacein-Bey-Abina, S *et al.* (2010). Dynamics of gene-modified progenitor cells analyzed by tracking retroviral integration sites in a human SCID-X1 gene therapy trial. *Blood* **115**: 4356–4366.
23. Ronen, K, Negre, O, Roth, S, Colomb, C, Malani, N, Denaro, M *et al.* (2011). Distribution of lentiviral vector integration sites in mice following therapeutic gene transfer to treat β-thalassemia. *Mol Ther* **19**: 1273–1286.
24. Zapala, MA, Hovatta, I, Ellison, JA, Wodicka, L, Del Rio, JA, Tennant, R *et al.* (2005). Adult mouse brain gene expression patterns bear an embryologic imprint. *Proc Natl Acad Sci USA* **102**: 10357–10362.
25. Johansson, PA, Irmler, M, Acampora, D, Beckers, J, Simeone, A and Götz, M (2013). The transcription factor Otx2 regulates choroid plexus development and function. *Development* **140**: 1055–1066.
26. Bushman, F, Lewinski, M, Ciuffi, A, Barr, S, Leipzig, J, Hannenhalli, S *et al.* (2005). Genome-wide analysis of retroviral DNA integration. *Nat Rev Microbiol* **3**: 848–858.
27. Ambrosi, A, Glad, IK, Pellin, D, Cattoglio, C, Mavilio, F, Di Serio, C *et al.* (2011). Estimated comparative integration hotspots identify different behaviors of retroviral gene transfer vectors. *PLoS Comput Biol* **7**: e1002292.
28. Ikeda, T, Shibata, J, Yoshimura, K, Koito, A and Matsushita, S (2007). Recurrent HIV-1 integration at the BACH2 locus in resting CD4+ T cell populations during effective highly active antiretroviral therapy. *J Infect Dis* **195**: 716–725.
29. Schröder, AR, 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**: 521–529.
30. Romiguier, J, Ranwez, V, Douzery, EJ and Galtier, N (2010). Contrasting GC-content dynamics across 33 mammalian genomes: relationship with life-history traits and chromosome sizes. *Genome Res* **20**: 1001–1009.
31. Kikin, O, D'Antonio, L and Bagga, PS (2006). QGRS Mapper: a web-based server for predicting G-quadruplexes in nucleotide sequences. *Nucleic Acids Res* **34**: W676–W682.
32. Bochman, ML, Paeschke, K and Zakian, VA (2012). DNA secondary structures: stability and function of G-quadruplex structures. *Nat Rev Genet* **13**: 770–780.
33. Svozil, D, Kalina, J, Omelka, M and Schneider, B (2008). DNA conformations and their sequence preferences. *Nucleic Acids Res* **36**: 3690–3706.
34. Cer, RZ, Bruce, KH, Mudunuri, US, Yi, M, Volfovsky, N, Luke, BT *et al.* (2011). Non-B DB: a database of predicted non-B DNA-forming motifs in mammalian genomes. *Nucleic Acids Res* **39**: D383–D391.
35. Cer, RZ, Donohue, DE, Mudunuri, US, Temiz, NA, Loss, MA, Starner, NJ *et al.* (2013). Non-B DB v2.0: a database of predicted non-B DNA-forming motifs and its associated tools. *Nucleic Acids Res* **41**: D94–D100.
36. Bartholomae, CC, Arens, A, Balaggan, KS, Yáñez-Muñoz, RJ, Montini, E, Howe, SJ *et al.* (2011). Lentiviral vector integration profiles differ in rodent postmitotic tissues. *Mol Ther* **19**: 703–710.
37. 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**: 1287–1289.
38. Lewinski, MK, Bisgrove, D, Shinn, P, Chen, H, Hoffmann, C, Hannenhalli, S *et al.* (2005). Genome-wide analysis of chromosomal features repressing human immunodeficiency virus transcription. *J Virol* **79**: 6610–6619.
39. Lewinski, MK, Yamashita, M, Emerman, M, Ciuffi, A, Marshall, H, Crawford, G *et al.* (2006). Retroviral DNA integration: viral and cellular determinants of target-site selection. *PLoS Pathog* **2**: e60.
40. Ciuffi, A, Mitchell, RS, Hoffmann, C, Leipzig, J, Shinn, P, Ecker, JR *et al.* (2006). Integration site selection by HIV-based vectors in dividing and growth-arrested IMR-90 lung fibroblasts. *Mol Ther* **13**: 366–373.
41. Mitchell, RS, Beitzel, BF, Schroder, AR, 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.
42. Carteau, S, Hoffmann, C and Bushman, F (1998). Chromosome structure and human immunodeficiency virus type 1 cDNA integration: centromeric alphoid repeats are a disfavored target. *J Virol* **72**: 4005–4014.
43. Allen, MJ, Jeffreys, AJ, Surani, MA, Barton, S, Norris, ML and Collick, A (1994). Tandemly repeated transgenes of the human minisatellite MS32 (D1S8), with novel mouse gamma satellite integration. *Nucleic Acids Res* **22**: 2976–2981.
44. Hawrylycz, MJ, Lein, ES, Guillozet-Bongaarts, AL, Shen, EH, Ng, L, Miller, JA *et al.* (2012). An anatomically comprehensive atlas of the adult human brain transcriptome. *Nature* **489**: 391–399.
45. Kang, HJ, Kawasawa, YI, Cheng, F, Zhu, Y, Xu, X, Li, M *et al.* (2011). Spatio-temporal transcriptome of the human brain. *Nature* **478**: 483–489.
46. Lein, ES, Hawrylycz, MJ, Ao, N, Ayres, M, Bensinger, A, Bernard, A *et al.* (2007). Genome-wide atlas of gene expression in the adult mouse brain. *Nature* **445**: 168–176.
47. Wu, X, Li, Y, Crise, B, Burgess, SM and Munroe, DJ (2005). Weak palindromic consensus sequences are a common feature found at the integration target sites of many retroviruses. *J Virol* **79**: 5211–5214.
48. Bor, YC, Miller, MD, Bushman, FD and Orgel, LE (1996). Target-sequence preferences of HIV-1 integration complexes *in vitro*. *Virology* **222**: 283–288.
49. Pruss, D, Reeves, R, Bushman, FD and Wolffe, AP (1994). The influence of DNA and nucleosome structure on integration events directed by HIV integrase. *J Biol Chem* **269**: 25031–25041.
50. Fitzgerald, ML and Grandgenett, DP (1994). Retroviral integration: *in vitro* host site selection by avian integrase. *J Virol* **68**: 4314–4321.
51. Pryciak, PM, Sil, A and Varmus, HE (1992). Retroviral integration into minichromosomes *in vitro*. *EMBO J* **11**: 291–303.
52. Mirkin, SM (2006). DNA structures, repeat expansions and human hereditary disorders. *Curr Opin Struct Biol* **16**: 351–358.
53. Wells, RD, Dere, R, Hebert, ML, Napierala, M and Son, LS (2005). Advances in mechanisms of genetic instability related to hereditary neurological diseases. *Nucleic Acids Res* **33**: 3785–3798.
54. Sinden, RR (1994). *DNA Structure and Function*, Academic Press. <http://www.amazon.com/DNA-Structure-Function-Richard-Sinden/dp/0126457506>.

55. Ho, PS (1994). The non-B-DNA structure of d(CA/TG)n does not differ from that of Z-DNA. *Proc Natl Acad Sci USA* **91**: 9549–9553.
56. Wahls, WP, Wallace, LJ and Moore, PD (1990). The Z-DNA motif d(TG)30 promotes reception of information during gene conversion events while stimulating homologous recombination in human cells in culture. *Mol Cell Biol* **10**: 785–793.
57. Tsutsui, KM, Sano, K, Hosoya, O, Miyamoto, T and Tsutsui, K (2011). Nuclear protein LEDGF/p75 recognizes supercoiled DNA by a novel DNA-binding domain. *Nucleic Acids Res* **39**: 5067–5081.
58. Bacolla, A, Larson, JE, Collins, JR, Li, J, Milosavljevic, A, Stenson, PD *et al.* (2008). Abundance and length of simple repeats in vertebrate genomes are determined by their structural properties. *Genome Res* **18**: 1545–1553.
59. Pruss, D, Bushman, FD and Wolffe, AP (1994). Human immunodeficiency virus integrase directs integration to sites of severe DNA distortion within the nucleosome core. *Proc Natl Acad Sci USA* **91**: 5913–5917.
60. Bushman, FD (1994). Tethering human immunodeficiency virus 1 integrase to a DNA site directs integration to nearby sequences. *Proc Natl Acad Sci USA* **91**: 9233–9237.
61. Weidhaas, JB, Angelichio, EL, Fenner, S and Coffin, JM (2000). Relationship between retroviral DNA integration and gene expression. *J Virol* **74**: 8382–8389.
62. Bor, YC, Bushman, FD and Orgel, LE (1995). *In vitro* integration of human immunodeficiency virus type 1 cDNA into targets containing protein-induced bends. *Proc Natl Acad Sci USA* **92**: 10334–10338.
63. Pryciak, PM and Varmus, HE (1992). Nucleosomes, DNA-binding proteins, and DNA sequence modulate retroviral integration target site selection. *Cell* **69**: 769–780.
64. Ciuffi, A and Barr, SD (2011). Identification of HIV integration sites in infected host genomic DNA. *Methods* **53**: 39–46.
65. Langmead, B and Salzberg, SL (2012). Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**: 357–359.
66. Quinlan, AR and Hall, IM (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841–842.

Supplementary Information accompanies this paper on the Molecular Therapy–Nucleic Acids website (http://www.nature.com/mtna)