Lentiviral vector common integration sites in preclinical models and a clinical trial reflect a benign integration bias and not oncogenic selection

Alessandra Biffi¹,², Cynthia C. Bartolomeæ³, Daniela Cesana¹,²,⁴, Natalie Cartier⁵, Patrik Aubourg⁵, Marco Ranzani¹,²,⁴, Martina Cesani¹, Fabrizio Benedicenti¹,², Tiziana Plati¹,², Enrico Rubagotti¹,², Stefania Merella¹,², Alessia Capotondo¹,²,⁴, Jacopo Sgualdino¹,²,⁴, Gianluigi Zanetti⁶, Christof von Kalle³, Manfred Schmidt³, Luigi Naldini¹,²,⁴§, Eugenio Montini¹,²,⁴¥§.

¹San Raffaele Telethon Institute for Gene Therapy (HSR-TIGET), Milan, Italy
²Division of Regenerative Medicine, Stem Cells and Gene Therapy, San Raffaele Scientific Institute, Milan, Italy.
³National Center for Tumor Diseases (Heidelberg, Germany).
⁴Vita Salute San Raffaele University, (Milan, Italy).
⁵INSERM UMR745, University Paris-Descartes, (Paris, France).
⁶Distributed Computing Group, Center for Advanced Studies, Research, and Development in Sardinia (CRS4), (Pula, Italy)

§LN and EM share Senior Authorship

¥Corresponding Author (montini.eugenio@hsr.it)

Copyright © 2011 American Society of Hematology
Abstract

A recent clinical trial for Adrenoleukodystrophy (ALD) showed the efficacy and safety of lentiviral vector (LV) gene transfer in Hematopoietic Stem Progenitor Cells (HSPC). However, several Common Insertion Sites (CIS) were found in patients’ cells, suggesting that LV integrations conferred a selective advantage. We performed high-throughput LV integration site analysis on human HSPC engrafted in immunodeficient mice and found the same CIS reported in ALD patients. Strikingly, most CIS in our experimental model and in ALD patients cluster in megabase-wide chromosomal regions of high LV integration density. Conversely, cancer-triggering integrations at CIS found in tumor cells from γretroviral vector-based clinical trials and oncogene-tagging screenings in mice, always target a single gene and are contained in narrow genomic intervals. These findings imply that LV CIS are produced by an integration bias towards specific genomic regions rather than by oncogenic selection.
Introduction

Stable genetic modification of Hematopoietic Stem/progenitor Cells (HSPC) is achieved with retroviral vectors (RV) that integrate into the cell genome and express a therapeutic transgene\(^1\). Transplantation of genetically modified autologous HSPC provides a therapeutic option for patients with genetic disorders\(^1,3\). However, in clinical trials for X-linked severe combined immunodeficiency (X-SCID) and chronic granulomatous disease (CGD) oncogenesis triggered by gamma (\(\gamma\))–RV-mediated insertional mutagenesis has occurred. Leukemic or myelodysplastic cell clones in patients from these trials harbored RV integrations at common insertion sites (CIS) targeting recurrently LMO2 or MDS1-EVI1, PRDM16, SETBP1 and other genes\(^4-7\).

Alternative to \(\gamma\)RV, HIV-derived self-inactivating (SIN) lentiviral vectors (LV) transduce human HSPC efficiently and display a superior safety profile with respect to \(\gamma\)RVs as shown in \textit{in vitro} and \textit{in vivo} in preclinical mouse models\(^8-11\). Moreover, good efficacy and safety of LV has also been documented in a recent HSPC-based clinical trial for X-linked adrenoleukodystrophy (ALD)\(^3\). However, a careful LV integration site analysis in ALD patients’ derived cells showed that relevant numbers of CIS were present\(^3\). This observation raises concerns\(^12\) because the detection of CIS is a well established hallmark of insertional mutagenesis in mice\(^13,14\) and clinical trials\(^5,7,15\). Thus, it is possible that the occurrence of CIS in the ALD clinical trial is a still silent effect of genotoxicity. To understand if CIS generated by LV integrations are the product of genotoxicity we generated our own dataset of LV integrations in human HSPC and their progeny after engraftment in immunodeficient mice and studied the integration pattern and the clonal repertoire of vector marked cells in \textit{in vitro} culture and \textit{in vivo}. Moreover, we performed an extensive comparison between our dataset and the integrations found in the ALD clinical trial and in other gene therapy trials that reported insertional leukemogenesis, as
well as in mice subjected to retroviral-mediated oncogene tagging. From our own integration data and the meta-analysis of the other integration datasets we provide evidence that the driving force leading to the appearance of CIS in LV transduced HSPC from the ALD clinical trial reflects a previously unappreciated bias of LV in integration site selection rather oncogenic selection.
Material and Methods

*Lentiviral vector production and isolation and transduction of human HSPC.* LV.ARSA and LV.GFP were produced using the pCCLsin.cPPT.hPGK.hARSA.WPREmut6 the pCCLsin.cPPT.hPGK.GFP.Wpre transfer plasmids\(^\text{16}\). VSV-pseudotyped LV concentrated stocks were produced and titered as described\(^\text{17}\). Human HSPC were obtained by positive selection of CD34-expressing cells (CD34 progenitor cell isolation kit, MACS; Miltenyi Biotec, Bergisch Gladbach, Germany) from bone marrow (BM) aspirates, mobilized peripheral blood (MPB) or cord blood (CB) of healthy donors upon informed consent collection in accordance with the Declaration of Helsinki (TIGET01 protocol, approved by San Raffaele Scientific Institute Ethical Committee). Alternatively, purified CD34+ cells from healthy donors’ BM were provided by Lonza (Human Bone Marrow CD34+ Progenitors 2M-101, Lonza). Soon after purification or thawing, cells were placed in culture on retronectin-coated wells (T100A Takara) in CellGro SCGM medium (2001 CellGenix) at a concentration of 1-1.5x10\(^6\) cells/ml in the presence of a standard cocktail of cytokines\(^\text{18}\) for 24-48 hours of pre-stimulation. Cells were then transduced with LV.ARSA or LV.GFP (at a multiplicity of infection [MOI] of 100-200) for 12 hours. One or two hits of transduction were performed. At the end of transduction, cells were counted and collected for clonogenic assays, flow cytometry, and *in vivo* studies. Remaining cells were plated in Iscove’s modified Dulbecco’s medium (IMDM) – 10% fetal bovine serum (FBS) with cytokines (IL-3, 60 ng/\(\mu\)l; IL-6, 60 ng/\(\mu\)l; SCF, 300 ng/\(\mu\)l) and cultured for a total of 14 days. Thereafter, cells were collected for molecular and flow cytometry studies.

*Rag2\(^{-/-}\)Il2rg\(^{-/-}\) mice transplantation and engraftment evaluation.* Rag2\(^{-/-}\)Il2rg\(^{-/-}\) mice were obtained from the Central Institute for Experimental Animals, Nogawa, Japan and maintained in our animal facility according to approved protocols. Three days old mice
were sub-lethally irradiated (450cGy) 24 hours before intravenous injection of un-transduced and un-manipulated or transduced CD34+ cells. HSPC transduction was performed as described above. Ten-twelve weeks after transplant, mice were euthanized and bone marrow, spleen and thymus were collected and multicolor cytofluorimetric analysis to assess human cell engraftment and differentiation performed as previously described\textsuperscript{16}. All procedures were performed according to protocols approved by the Animal Care and Use Committee of the San Raffaele Scientific Institute (IACUC #325 and #353) and communicated to the Ministry of Health and local authorities according to Italian law.

**Quantitative PCR.** Genomic DNA was extracted from CD34+ liquid culture samples with QIAamp DNA Blood Mini Kit-Qiagen, and from murine tissues with the Blood & Cells DNA Midi Kit-Qiagen after o/n digestion with proteinase K (PK)(Roche). LV sequences were detected by quantitative PCR on 50 ng of total genomic DNA (primers and probe: Fw: 5’-TAC TGA CGC TCT CGC ACC-3’; Rv: 5’- TCT CGA CGC AGG ACT CG-3’; Probe: 5’-FAM- ATC TCT CTC CTT CTA GCC TC-MGB) and normalized on hTert gene (primers and probe: Fw: 5’-GGA CAC GTT ATT TAC CCT GTT TCG-3’; Rv: 5’- GGT GAA CCT CGT AAG TTT ATG CAA-3’; Probe: 5’-VIC-TCA GGA CGT CGA GTG GAC ACG GTG-TAMRA-3’). Absolute quantifications were plotted on standard curves prepared with serial dilutions of genomic DNA from the CEMA301 clone\textsuperscript{19}. VCN was then calculated as described\textsuperscript{20}.

**LAM-PCR and genomic integration site analysis.** The sequences of LAM-PCR primers and procedures for LV integration site retrieval have been previously described\textsuperscript{10,21}. Briefly, 10–100 ng of genomic DNA were used as template for LAM-PCR and initiated with a 50-cycle linear PCR and restriction digest using Tsp509I or HpyCH4IV and ligation of a restriction site–complementary linker cassette. The first exponential biotinylated PCR product was captured via magnetic beads and re-amplified by a nested
PCR. LAM-PCR products were separated by Spreadex gel electrophoresis (Elchrom Scientific) to verify the presence and number of bands. LAM PCR were shotgun cloned into the TOPO TA vector (Invitrogen) and sequenced by Sanger sequencing (GATC Biotech), or directly sequenced by 454 pyrosequencing after a PCR re-amplification using oligonucleotides with specific 4-8 nucleotide sequence tags for sample identification. Sequences were validated and classified using specific PERL scripts and aligned to the human genome (freeze March 2006, UCSC) using the UCSC BLAT genome browser (http://genome.ucsc.edu). Genes targeted by vector integrations were considered those nearest to the integration site.

Statistical analysis. Statistical analyses were made by One Way ANOVA for repeated measurements using Bonferroni’s test for post-hoc analysis after significant main effect of the treatment (confidence interval 95%), or with Student’s t test (confidence interval 95%). Over-represented on gene ontology classes were identified by the DAVID-EASE software 2.0 (http://david.abcc.ncifcrf.gov/home.jsp) using the stringency setting “high”. Differences in targeting frequency at gene classes between in vitro and in vivo datasets or between datasets originated by BM-, MPB- and CB-derived CD34+cells were scored by the Fisher’s exact test using the GraphPad Prism Software, version 5.03. Automated generation and statistical evaluation of the chromosomal frequency distributions and the distributions around CIS centers of the different integration datasets was performed with ad hoc scripts using R statistical software version 2.10.1.

Grubbs’ test for outlier analysis was performed with the online “Outlier Calculator” tool (http://www.graphpad.com/quickcalcs/Grubbs1.cfm) using the values of gene integration frequency corrected by the gene-size to which was added a constant value of 100 Kb (calculated using the UCSC Hg18 RefSeq genomic coordinates for transcript start and end ±50 Kb).
For all statistical comparisons unless otherwise specified, p<0.05 was considered significant.
Results

We analyzed the integration profile of different LVs expressing therapeutic $^{20}$ or GFP transgenes in human CD34+ HSPC in vitro and in vivo after engraftment in transplanted $Rag2^{-/-}Il2rg^{-/-}$ mice (Fig. 1A and Supplemental Fig. 1). More than 2,300 unique genomic integration sites were mapped by Linear Amplification Mediated (LAM) PCR $^{21}$ and high-throughput 454 pyrosequencing from 5 distinct in vitro samples and 13 transplanted mice (Supplemental Table 1). LVs displayed the expected tendency to integrate within genes (data not shown). No enrichment in preferentially targeted gene classes was found from in vitro to in vivo (Supplemental Fig. 2). Adopting previously described statistical criteria $^{22}$, 31 CIS were identified in our datasets (Supplemental Table 2), accounting for 5.2 to 8.6% of all integrations. No enrichment in CIS was found from in vitro to in vivo condition (Supplemental Table 2). Interestingly, 85% (11/13) of the CIS hit by ≥4 integrations found in our dataset from human mouse hematochimeras matched the CIS identified in the ALD clinical trial (Table 1). Moreover, five 10-20Mb-wide genomic regions containing 5 or more CIS, among which, at least one overlapped with the integration dataset from the ALD clinical trial. Strikingly, up to 76% of all LV integrations at CIS were clustered in these 5 genomic regions (accounting for 2.5% of the human genome) located in chromosomes 3, 6, 11, and 17 (2 regions) (Fig.1B, Supplemental Fig. 2A and Supplemental Table 3). This LV CIS distribution is clearly different from those obtained analyzing well characterized oncogenic CIS from $\gamma$RV-based clinical trials for X-SCID $^7$ and CGD $^5$ (674 and 760 integrations analyzed, respectively) and oncogene-tagging screenings in mice $^{14,23,24}$ with $\gamma$-retroviruses and Sleeping Beauty (SB) transposon (21,511 integrations analyzed from 17 studies) (Fig. 1C,D and Supplemental Fig. 3B,C). Indeed, oncogenic CIS always appeared as isolated and sharp peaks in the frequency distribution of integrations along the chromosomes.
The only exception was found in SB transgenic mice that showed on chromosome 1 a higher integration frequency over a wide genomic region near the transposon concatemer (Fig. 1E); this enrichment, however, is caused by a “local hopping” effect and not by oncogenic selection. Considering the distribution of integrations mapping in a 5 Mb-wide genomic region centered on each CIS, we found that, in the case of genotoxic CIS, 90% of integrations were located, at most, in a central 100 Kb region (n=50 CIS analyzed), while LV integrations were evenly spread over the whole region (n=81 LV CIS) (Fig. 1G, and Supplemental Fig. 4; p value ranging from $2 \times 10^{-16}$ to $7.7 \times 10^{-5}$, according to bin size and comparison group; Fisher’s Exact test). These findings indicate that LV CIS are embedded in wide regions of high integration density. Interestingly, the distribution of LV CIS integrations within the boundaries of the CIS itself was significantly more spread than for oncogenic CIS (Supplemental Fig. 4C). The tight clustering of integrations at genotoxic CIS suggests that selectable oncogene activation events, independently from the type of vector used, it may occur preferentially with integrations targeting specific genomic regions. Probably, integrations in close proximity of regulatory regions or within specific introns favor oncogene overexpression or the formation of aberrantly spliced oncogenic proteins. To test the validity of this rule also for LV, we induced hematopoietic tumors by injecting a previously described genotoxic LV in Cdkn2a mice and analyzed the integrations in tumors. The CIS generated in this LV-based insertional mutagenesis model, as well as in another in vitro study, displayed the same narrow clustering pattern described above for genotoxic CIS (as shown for the top ranking CIS, the Braf oncogene, Fig. 1F). Of note, cell clones harboring LV CIS integrations were not preferentially enriched in vivo and did not predominate over other repopulating cells, both in our study here (Fig. 2) and in the ALD clinical trial.
The canonical statistical approaches for the identification of biologically significant CIS contained in a given genomic interval assume that integrations are randomly distributed along the genome. To correct the significance of CIS for biases of vector genomic integration, we devised an additional CIS validation step that takes into account the relative frequency of integration at the genomic region surrounding the CIS interval. We decided that the best approach was to measure and compare the integration frequency within the genomic intervals defined by transcription units rather than the entire flanking genomic intervals that may contain large intergenic regions without integrations. Therefore, our analysis is focused on the comparison of integration frequencies at genes, some of which are the culprits of oncogenesis and the preferred targets of different vector platforms. In our rationale, for a CIS to be considered the result of genetic selection (genotoxic), the integration frequency at the CIS target-gene interval must be high enough to be considered a significant outlier with respect to the integration frequency at other genes contained in the flanking genomic regions (genes targeted at least by 1 integration). On the other hand, if the integration frequency at the CIS target-gene is not statistically different from the integration frequency of other flanking genes, it will imply that the CIS is embedded in a wider region of similar integration frequency, and thus likely the product of a vector-specific integration bias. The gene integration frequency is defined as the ratio between the number of integrations targeting a given gene and its size. To determine whether the integration frequency at a CIS target-gene is high enough to be considered a significant outlier with respect to other genes contained in the surrounding regions, we performed the Grubbs’ test for outliers (Supplemental Statistical Material). We applied the Grubbs’ test for outliers to the gene integration frequencies of: 9 γRV CIS reported in the X-SCID and CGD clinical trials;
CIS targeting *Braf* both in mouse *Cdkn2a*<sup>-/-</sup> LV.SF.LTR-marked histiocytic sarcomas and in SB-transposon-marked *Arf*<sup>-/-</sup> sarcomas; and finally on the several LV CIS identified in the 5 genomic regions in the ALD clinical trial and in the human/mouse hematochimeras in this study (Fig. 3 and Supplemental Statistical Material).

The approach does find significant ratio Z outliers for γRV CIS at *LMO2*, *CCND2*, *RUNX1*, *EVI1-MDS1*, *SETBP1* and *PRDM16* (targeted respectively by 7, 9, 5, 94, 9, and 37 integrations) from the X-SCID and CGD clinical trials (Fig. 3A and Supplemental Statistical Material). Other γRV CIS at *EGR* (4 integrations), *BCL2* (4 integrations) and *BACH2* (5 integrations), did not appear to be targeted at a significantly higher frequency with respect to flanking genes. This approach is not influenced by the integration site selection of the different vector platforms as identifies the genotoxic murine CIS at *Braf* targeted both in LV.SF.LTR induced *Cdkn2a*<sup>−/−</sup> histiocytic sarcomas and in SB-transposon induced *Arf*<sup>−/−</sup> sarcomas (Fig. 3B, C).

On the other hand, LV CIS from the ALD trials and our human/mouse hematochimeras, even if targeted by high numbers of integrations (for example 29, 27 and 19 integrations respectively targeting *PACS1*, *FBXL11* and *TNRC6C*), did not show a significantly higher targeting frequency with respect to flanking genes that were also identified as CIS by the canonical statistical analysis (Fig. 3D and Supplemental Statistical Material).
Discussion

We showed that the genomic integration profile of LVs expressing therapeutic and marker transgenes in human hematopoietic cells engrafted in immunodeficient mice is remarkably similar to the integration profile observed in LV-treated ALD patients. These data indicate that xenotransplantation models are a valid surrogate for the study of integration profiles of LVs in human HSPC in vivo. Moreover, the LV CIS found in our datasets overlapped for the 85% with those reported in the ALD clinical trial.

Although the well established role of CIS in genotoxicity has fueled concerns regarding the safety of LV CIS identified in the ALD clinical trial\textsuperscript{12}, our analysis highlights important differences with respect to the known genotoxic CIS identified in malignant cell clones from mouse oncogene-tagging screenings or γRV-based clinical trials. Indeed, the LV CIS in the ALD clinical trial and in our human/mouse hematochimeras clustered in Mb-wide genomic regions with an overall higher integration frequency with respect to other chromosomal regions. Differently, genotoxic CIS are distributed along chromosomes as isolated sharp peaks and always targeting a single gene, the culprit of oncogenesis. Since the features characterizing genotoxic CIS are consistent across different vector platforms (retroviruses, γRV, transposons and genotoxic LV.SF.LTR) and tumor types (hematopoietic and mammary), it suggest that a different mechanisms, other than genetic selection, may drive the formation of LV CIS in our pre-clinical and ALD studies. In support to this notion, it is unlikely that genetic selection would preferentially favor integrations deregulating cancer genes clustered in specific genomic intervals and not the many other well known oncogenes spread along the genome.
Moreover, differently than the known genotoxic CIS integrations marking leukemic or dominant cell clones,\textsuperscript{5,27,28} the LV CIS integrations in the ALD clinical trial and our human/mouse hematochimeras are not enriched from \textit{in vitro} to \textit{in vivo} conditions, or during time after transplant, and are not over-represented (dominant) with respect to other integrations. It should be noted, however, that without further experimental evidence, it is not possible to formally exclude that any of the CIS integrations or even any integration of the dataset, regardless the CIS status or the type of targeted gene, could be the result of selection.

Our findings highlight also the need of more stringent statistical tools for interpreting the presence of CIS identified in future clinical trials. Canonical CIS statistics assume that integrations are distributed in a randomly across the genome and do not take in account the integration biases intrinsic to a given vector. Therefore CIS in genomic regions targeted at high frequency will be considered identical to those in which only one gene is targeted at high frequency. Alternative statistical methods for CIS-validation should consider the size of the datasets analyzed and the local genomic integration biases. We developed a new approach for the validation of CIS significance based on the comparison of the integration frequency at the CIS gene with respect to other genes contained in the surrounding genomic regions. Using the Grubbs’ test for outliers we were able to distinguish well validated genotoxic CIS generated using 3 different vector platforms (genotoxic LV, SB-Transposon and \(\gamma\)-retrovirus/RV). Some CIS from the \(\gamma\)RV-based clinical trials were not found to be outliers by this test. Whether the lack of significance is due to a low sensitivity of this specific test or to a true lack of genotoxicity is unclear. On the other hand, the LV CIS in the ALD clinical trial and our human/mouse hematochimeras mapped in 10-20Mb-wide chromosomal regions together with other genes also targeted by CIS at a similar integration frequency, and were not found to be
outliers by this test. Based on our rationale, a CIS cannot be considered a significant outlier when in the same 10-20Mb wide chromosomal region multiple CIS are present with a similar integration frequency. The reasons why CIS target-genes display a higher integration frequency with respect to others genes within the same interval remain obscure. Possibly, cellular protein-mediated tethering of the lentiviral pre-integration complex at gene dense genomic regions with high transcriptional activity could be responsible for the observed LV integration preferences and CIS formation in our pre-clinical and ALD studies\textsuperscript{35-49}. More refined statistical methods capable of detecting multiple outliers within a population of values may be required to pinpoint multiple genes targeted at a significantly higher frequency with respect to the average gene integration frequency (for example using the Chauvenet's criterion, Peirce's criterion, Bayesian models and others)\textsuperscript{29,30}.

One of the strengths of our outlier-detection approach is that it takes advantage of the integration pattern originating from the same vector-specific dataset and in similar experimental conditions to perform statistical comparisons with respect to the flanking genomic regions, without the need of random- or neutral-control integration profiles. This is important since vector-specific integration profiles from \textit{in vitro} or non leukemic cells from transplanted patients cannot be formally assumed as neutral, as genotoxic integrations may be selected \textit{in vitro} or may be present in “normal” hematopoietic cells \textit{in vivo} before full-blown neoplastic transformation occurs. Moreover, in future studies it will be useful to study larger $\gamma$RV integration datasets\textsuperscript{27,31,32} to possibly improve the strength of the analyses and perform comparisons with CIS identified by Kernel deconvolution-based and k-mean clustering analysis methods\textsuperscript{33,34}.

Overall, our findings highlight a previously unappreciated feature of LV integration that invalidates the predictive value on genotoxicity of standard CIS statistics for this class of vectors. Moreover, our meta-analysis provides a way to distinguish alarming CIS
originating from gain-of-function mutations from those likely originating from biases in integration site selection. The lack of evident signs of genotoxicity during the 2-years follow-up of the ALD clinical trial and in our human/mouse hematochimeras, the widespread distribution of LV integrations at and around CIS, and the integration preference for specific genomic regions altogether suggest that the LV CIS integrations found in the ALD clinical trial are probably the result of an intrinsic integration bias towards selected megabase-wide genomic regions in HSPC.
Acknowledgements

We are kindly indebted to Lucia Sergi Sergi for LV production, Laura Lorioli, Rossano Cesari and Anna Zingale for help with quantitative PCR analysis. This work was supported by grants from the Association for International Cancer Research (AICR 09-0784 to EM), Italian Telethon (TIGET grants to AB, LN and EM), EU (HEALTH-2009-222878, PERSIST to LN), the Italian Ministries of University and Research and of Health to LN, and from the European Leukodystrophy Association (ELA to AB and LN).

Author Contributions

A.B., designed the research, analyzed data and wrote the manuscript. C.C.B., D.C., M.R., M.C., F.B., T.P., A.C. and J.S. performed research and analyzed data. E.R and S.M. performed the bioinformatics and statistical analyses. P.A., N.C., provided original integration data, C.v.K., M.S., L.N. and E.M. designed the research, analyzed data and wrote the manuscript.

Conflict of Interest Disclosure

The authors have no conflict of interest to declare.
References


Table 1

<table>
<thead>
<tr>
<th>N Hits</th>
<th>Genes (Interval Kb)</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>CARD8 (23), NSD1 (90), QRICH1 (31), SAPS2 (47), USP48 (48)</td>
</tr>
<tr>
<td>5</td>
<td>GPATCH8 (53)</td>
</tr>
<tr>
<td>6</td>
<td>FCHSD2 (272)</td>
</tr>
<tr>
<td>7</td>
<td>NPLCOC4 (76), SMARCC1 (95), NF1 (140)</td>
</tr>
<tr>
<td>8</td>
<td>PACS1 (110), HLA (542)</td>
</tr>
<tr>
<td>9</td>
<td>FBXL11 (107)</td>
</tr>
</tbody>
</table>

Table 1. CIS genes targeted multiple times (N Hits) within the integration datasets. The maximum distance between integrations targeting the same CIS is indicated (Interval Kb). Note that only CIS genes targeted by 4 or more integrations are shown. With the exception of SAPS2 and USP48 all other CIS genes found in our experimental dataset matched the CIS of the ALD clinical trial. See Supplemental Table 2 for further details on the identified CIS.
Figure Legends

Figure 1. Identification of LV CIS in human HSPC from hematochimeric mice and ALD clinical trial and comparative analysis of integration distribution within the CIS and in the surrounding chromosomal regions in datasets with documented insertional mutagenesis events.

A) Experimental strategy for LV integration site profiling in human CD34+ HSPC derived from bone marrow (BM), mobilized peripheral blood (MPB) and cord blood (CB). Upon ex vivo transduction with LV expressing a therapeutic (arylsulfatase A, LV.ARSA) or marker (LV.GFP) gene, cells were transplanted into immunodeficient mice (Rag2<sup>−/−</sup> Il2rg<sup>−/−</sup>) and a portion cultured in vitro for 14 days. BM, Thymus (Thy) and Spleen (Spl) from transplanted mice were harvested 12 weeks post-transplant. Vector copy number, engraftment and integration site analysis was then performed on the available samples. B) Frequency distributions of LV integrations at 5 chromosomal regions targeted at high frequency. The bin size used for the chromosomal distributions is 1 Mb. On the y-axis: % of the total integrations of each dataset; on the x-axis: chromosomal coordinates in Mb x 10. Genes at CIS locations are indicated for the ALD dataset and in red when common between the ALD and our datasets from (a). C) Frequency distributions of γRV integrations surrounding validated genotoxic CIS found in X-SCID and CGD clinical trials. D) Frequency distributions of γ-retroviruses or SB-transposon integrations surrounding validated genotoxic CIS found in tumors generated in different insertional mutagenesis studies. E) Frequency distribution of SB-transposon integrations at chromosome 1 near the transposon concatemer locus in transgenic mice. F) Frequency distribution of genotoxic LV integrations targeting: (left) Braf in hematopoietic tumors from Cdkn2a<sup>−/−</sup> mice; (right) the Ghr gene in IL3-independent cell clones from Ref.25.
G,H) Distribution of vector integrations around CIS centers. G) Tukey box-and-whisker graph representing the distance of vector integrations from the center of CIS found in each dataset in a ±2.5 Mb region (x-axis, units in bp). H) Tukey box-and-whisker graph representing the distance of vector integrations from the center of each CIS within the CIS interval. The center of each CIS was calculated as the position closest to the highest number integrations within the CIS interval. The tighter clustering of genotoxic integrations within CIS boundaries, although suggestive of positional constrains for cancer-gene activating integrations, it does not test if the integration frequency at the CIS is significantly different with respect to other regions and therefore cannot be used to discriminate between different CIS types.

Figure 2. Relative retrieval frequency of sequencing reads of integration sites in in vitro and in vivo samples

Retrieval frequency of sequencing reads corresponding to a unique LV integration site from the in vitro culture and the indicated organs of mice transplanted with CD34+ HSPC from BM- (A), MPB- (B), and CB-derived cells (C). Within a red box are represented the integrations at CIS (considered only CIS constituted by at least 4 integrations). LAM PCR products were sequenced by 454-pyrosequencing or Sanger chemistry. Histograms show the percentage of reads for each integration site in the sample dataset. The total number of reads and of unique integration sites (INTS) in each sample dataset is given. Integrations represented by <2% of the total reads in the dataset were pooled and shown in black at the top of each bar (<2%). Integrations represented by >2% of the total sequencing reads are shown individually with the symbol of the targeted gene. Identical integrations found in different organs of the same mouse are shown in green. In (D) the averaged percentage of sequencing reads representing CIS integrations (in red) and non-CIS integrations (in blu) was not statistically different.
Figure 3. Graphical representation of the Grubbs’ test for outliers results on gene integration frequencies in ∼10-20 Mb genomic regions around CIS.

On the y-axis, the Ratio Z that measures how distant is the integration frequency of a given gene with respect to the average of all genes analyzed (genes targeted by at least 1 integration contained within the specified genomic interval). On the x-axis, the chromosomal position of the gene (coordinates in bp). A negative or positive Ratio Z values implicate that the gene is targeted at a frequency below or above the average respectively. The red lines indicate the threshold beyond which the values can be considered significantly different. The red triangles indicate genes considered as CIS in previous publications using the classical statistical approach (in parenthesis the number of integrations targeting each gene).

In A) two examples of Ratio Z of gene integration frequency at genomic regions around CIS form γ RV-based X-SCID and CGD clinical trials. LMO2 (targeted by 7 integrations) and CCND2 (7 integrations) appear to be targeted at a significantly higher frequency by this test (see Supplemental Statistical Material for analyses of other CIS from the same clinical trials). In B) Ratio Z of gene integration frequency at the genomic region around Braf (68 integrations) and the neighboring genes in histiocytic sarcomas form Cdkn2a−/− mice injected with LV.SF.LTR. In C) a very similar integration profile is found at Braf (24 integrations) in sarcomas from Art−/− SB-Transposon/transposase transgenic mice. In D) Ratio Z representation of 2 genomic regions at LV CIS in common between the ALD clinical trial and the hematochimeric model show that none of the identified CIS is a significant outlier.
Lentiviral-vector common integration sites in preclinical models and a clinical trial reflect a benign integration bias and not oncogenic selection

Alessandra Biffi, Cynthia C. Bartolomae, Daniela Cesana, Natalie Cartier, Patrik Aubourg, Marco Ranzani, Martina Cesani, Fabrizio Benedicenti, Tiziana Plati, Enrico Rubagotti, Stefania Merella, Alessia Capotondo, Jacopo Sgualdino, Gianluigi Zanetti, Christof von Kalle, Manfred Schmidt, Luigi Naldini and Eugenio Montini