# Vector integration is nonrandom and clustered and influences the fate of lymphopoiesis in SCID-X1 gene therapy

Annette Deichmann,[1,2,3] Salima Hacein-Bey-Abina,[4,5] Manfred Schmidt,[1,2,3] Alexandrine Garrigue,[4] Martijn H. Brugman,[6] Jingqiong Hu,[1] Hanno Glimm,[1,2] Gabor Gyapay,[7] Bernard Prum,[8] Christopher C. Fraser,[9] Nicolas Fischer,[10] Kerstin Schwarzwaelder,[1,3,11] Maria-Luise Siegler,[1] Dick de Ridder,[12,13] Karin Pike-Overzet,[12] Steven J. Howe,[14] Adrian J. Thrasher,[14,15] Gerard Wagemaker,[6] Ulrich Abel,[3,16] Frank J.T. Staal,[12] Eric Delabesse,[17] Jean-Luc Villeval,[18] Bruce Aronow,[19] Christophe Hue,[4,5] Claudia Prinz,[1] Manuela Wissler,[1,2] Chuck Klanke,[20] Jean Weissenbach,[7] Ian Alexander,[21] Alain Fischer,[4,22] Christof von Kalle,[1,2,3,20] and Marina Cavazzana-Calvo[4,5]

[1]Institute for Molecular Medicine and Cell Research and [2]Department of Internal Medicine I, University of Freiburg, Freiburg, Germany. [3]National Center for Tumor Diseases, Heidelberg, Germany. [4]INSERM U768, Hôpital Necker, and Faculté de Médecine, Université René Descartes Paris V, Paris, France. [5]Département de Biothérapies, Hôpital Necker, Paris, France. [6]Department of Hematology, Erasmus Medical Center, Rotterdam, The Netherlands. [7]Génoscope and CNRS UMR8030, Evry, France. [8]Laboratoire "Statistique et Génome," UMR CNRS 8071, Evry, France. [9]Millennium Pharmaceuticals Inc., Cambridge, Massachusetts, USA. [10]Laboratoire National de Métrologie et D'essais, Trappes, France. [11]Faculty of Biology, University of Freiburg, Freiburg, Germany. [12]Department of Immunology, Erasmus Medical Center, Rotterdam, The Netherlands. [13]Information and Communication Theory Group, Faculty of Electrical Engineering, Mathematics, and Computer Science, Delft University of Technology, Delft, The Netherlands. [14]Molecular Immunology Unit, Institute of Child Health, University College London, London, United Kingdom. [15]Department of Clinical Immunology, Great Ormond Street Hospital for Children NHS Trust, London, United Kingdom. [16]Department of Medical Biostatistics, Tumor Center Heidelberg-Mannheim, Heidelberg, Germany. [17]Laboratoire d'Hématologie CHU Purpan, Toulouse, France. [18]INSERM U790, Institut Gustave Roussy, Villejuif, France. [19]Division of Bioinformatics, Children's Hospital Medical Center, Cincinnati, Ohio, USA. [20]Division of Experimental Hematology, Cincinnati Children's Research Foundation, Cincinnati, Ohio, USA. [21]The Children's Hospital at Westmead and Children's Medical Research Institute, Sydney, New South Wales, Australia. [22]Unité d'Immunologie et d'Hématologie Pédiatriques, Hôpital Necker–Enfants Malades, Paris, France.

**Recent reports have challenged the notion that retroviruses and retroviral vectors integrate randomly into the host genome. These reports pointed to a strong bias toward integration in and near gene coding regions and, for gammaretroviral vectors, around transcription start sites. Here, we report the results obtained from a large-scale mapping of 572 retroviral integration sites (RISs) isolated from cells of 9 patients with X-linked SCID (SCID-X1) treated with a retrovirus-based gene therapy protocol. Our data showed that two-thirds of insertions occurred in or very near to genes, of which more than half were highly expressed in CD34$^+$ progenitor cells. Strikingly, one-fourth of all integrations were clustered as common integration sites (CISs). The highly significant incidence of CISs in circulating T cells and the nature of their locations indicate that insertion in many gene loci has an influence on cell engraftment, survival, and proliferation. Beyond the observed cases of insertional mutagenesis in 3 patients, these data help to elucidate the relationship between vector insertion and long-term in vivo selection of transduced cells in human patients with SCID-X1.**

## Introduction

Retroviruses have been used as efficient gene-delivery vehicles in several gene therapy trials because they integrate stably into the genome, allowing the genetic correction of stem cells, potentially for the entire lifespan of the affected individual (1–3). The availability of the complete human genome sequence has made possible large-scale sequence-based surveys of retroviral integration sites (RISs), which have strongly challenged the notion that retrovirus vector integration may be a semirandom event (4, 5). Schroeder et al. investigated targeting of HIV and HIV-based vectors in a human lymphoid cell line (SupT1) and found that genes were favored integration targets (6). Similarly, Wu et al. examined targeting of murine leukemia virus (MLV) in human HeLa cells and found that MLV strongly favored integration in transcriptional units, with integration focusing near the start of transcription (7). This nonrandom distribution of integrations has been confirmed by Laufs et al. in human bone marrow–repopulating cells in mouse xenografts (8).

A comparative analysis of human primary cell types and cell lines transduced with HIV-1–, avian sarcoma leukemia virus– (ASLV-), or MLV-based vectors showed that each vector type produces a unique pattern of RIS distribution in the human genome (9). These analyses revealed a significant association between integration target sites and transcriptional profiling for HIV-1, but not for ASLV or MLV (9). Thus, the statistics of the integration process of retroviruses, lentiviruses, and derived vectors suggest that a more specific mechanism — e.g., active tethering of the preintegration complex to DNA motifs, DNA binding factors, or other connections to the gene activation or expression status of target cells — are of influ-

**Nonstandard abbreviations used:** γc, common γ chain; CIS, common integration site; GO, gene ontology; kbp, kilobase pair(s); LAM-PCR, linear amplification–mediated PCR; LTR, long-terminal repeat; MLV, murine leukemia virus; Pt, patient; RIS, retroviral integration site; SCID-X1, X-linked SCID; TSS, transcription start site.

**Conflict of interest:** The authors have declared that no conflict of interest exists.

**Citation for this article:** *J. Clin. Invest.* **117**:2225–2232 (2007). doi:10.1172/JCI31659.

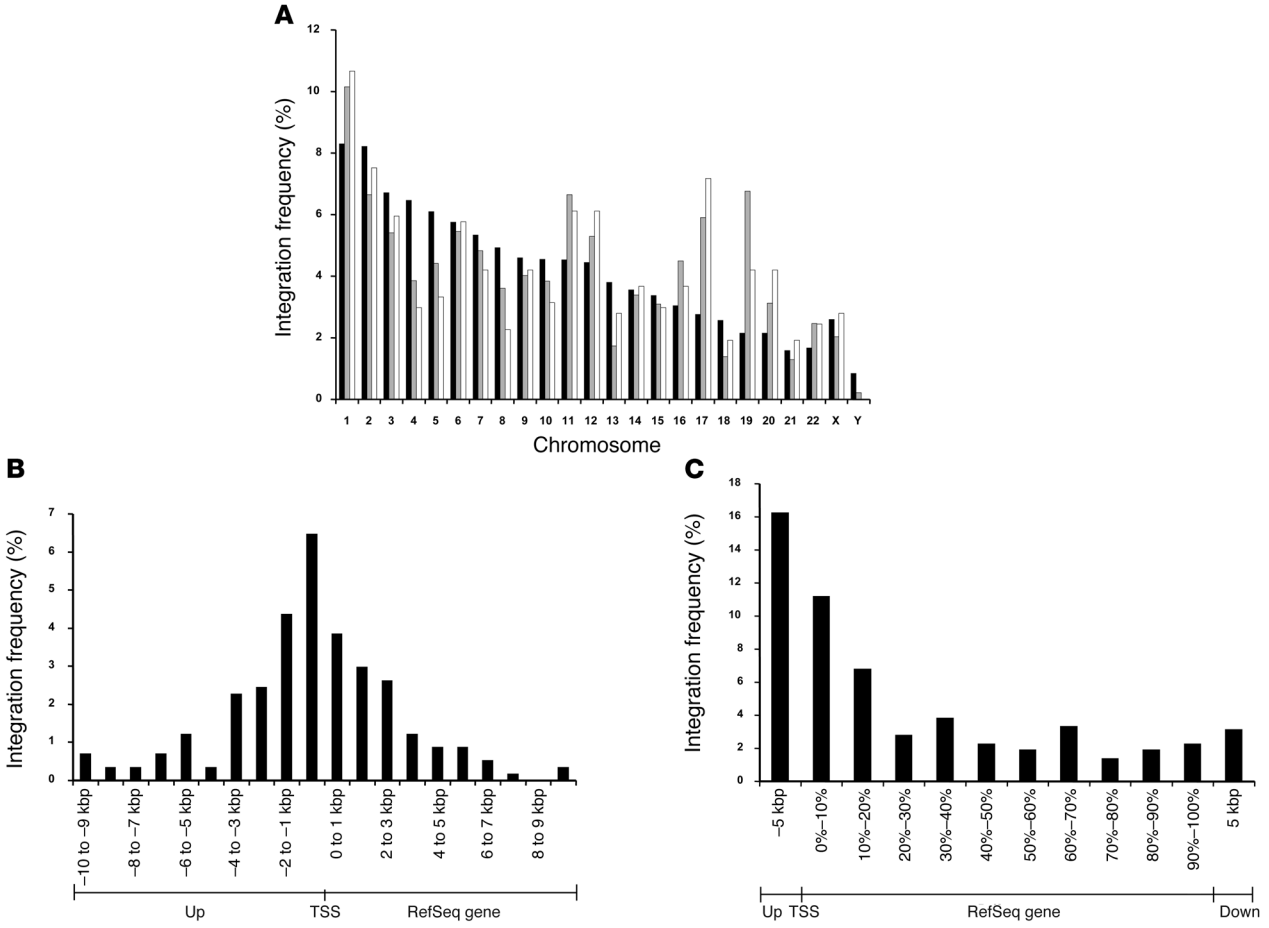


**Figure 1**
RIS distribution analysis of engrafted cells. (**A**) RIS distribution compared with chromosome size and gene content. The displayed chromosome distribution accounts for the double copy number of diploid autosomes. Black bars, size of chromosomes; gray bars, number of known genes; white bars, number of RISs. (**B** and **C**) Vector integration in and near RefSeq genes. RISs were preferentially found near the TSS (**B**) and within gene coding regions (**C**). Negative numbers denote the region upstream (Up) of a gene, positive numbers indicate the gene region downstream of the TSS (RefSeq gene) (**B**) or downstream (Down) of the gene (**C**). (**C**) The position of intragenic hits was mapped according to the percentage of overall gene length.

ence beyond the accessibility of the euchromatin (10–12). In line with this hypothesis, a comparative analysis of retrovirus integration and gene expression status demonstrated reduced integration in genomic sites with highly active transcription (13). A large-scale mapping of RISs in gene-modified T lymphocytes from leukemic patients after allogeneic stem cell transplantation has shown that retroviral vectors integrated preferentially in genes expressed during transduction and that integrations can deregulate gene expression, albeit without obvious side effects (14).

Of the published large-scale in vitro integration site studies, none followed the possible selective advantage induced by virus or vector integration for an individual transduced cell over time. Interestingly, an analysis of MLV retrovirus and SIV lentivirus integration sites in a preclinical nonhuman primate model discovered the presence of common integration sites (CISs) in transcriptional units (15). Recent studies on transduced CD34$^+$ cells have further demonstrated that vector integration is indeed nonrandom, often clustered, and potentially capable of inducing immortalization in vitro, clonal dominance in vivo, or even leukemogenesis in vivo (16–18). Insertion in human gene-modified T lymphocytes occurred preferentially at the transcription start site (TSS), but only a low incidence of CIS insertion was found (14).

Recurrent integration in specific gene loci strongly indicates that the insertion has provided a nonrandom growth or survival advantage to the affected target cell clones (17, 18). Our recent observation in a clinical gene therapy trial for chronic granulomatous disease that cell clones with integrations in *MDS1/EVI1*, *PRDM16*, or *SETBP1* drove a 3- to 4-fold in vivo expansion of the gene-corrected myeloid cell pool emphasizes the importance of analyzing the influence of the integration sites present in transduced cells and their clonal progeny in current gene therapy trials aimed at curing disorders of the myeloid or lymphoid blood cell compartment (19). The occurrence of a lymphoproliferative disease in 3 of our 9 patients showed the biological relevance the integration of replication-defective retroviral vectors may have (20).

Here we demonstrated, by high-throughput integration site analysis and sequencing performed on CD34$^+$ transduced cells and sorted peripheral blood cell samples obtained from patients of

**Table 1**
Overall characteristics of RISs found in 9 patients

| | Pt4, Pt5, Pt10 | Pt1, Pt2, Pt6–Pt9 | Total |
|---|---|---|---|
| Exactly mappable RISs | 210 (100) | 362 (100) | 572 (100) |
| RISs in RefSeq genes | 81 (39) | 135 (37) | 216 (38) |
| RISs in RefSeq genes including the 10-kbp surrounding region | 130 (62) | 226 (62) | 356 (62) |
| RISs near TSSs (± 5 kbp) | 59 (28) | 98 (27) | 157 (27) |
| RISs close to CpG islands (± 1 kbp) | 34 (16) | 66 (18) | 100 (17) |

The time span of investigation for each patient was as follows: Pt1, 15–38 months; Pt2, 13–41 months; Pt4, 6–41 months and pretransplantation sample; Pt5, 13–37 months; Pt6, 4–16 months; Pt7, 11–16 months; Pt8, 10 months; Pt9, 4–12 months; Pt10, 5–12 months. RISs are shown as absolute number (percent) of the exactly mappable sequences for each category. RIS distribution of Pt4, Pt5, and Pt10, which developed leukemia following gene therapy, is shown separately in comparison with RIS distribution in the other patients.

the first X-linked SCID (SCID-X1) gene therapy trial, that integration of retroviral vectors took place preferentially in gene coding regions, was skewed to the transcriptional start site (TSS) of genes, and was significantly correlated with the gene expression pattern of the gene-corrected cell population. Most strikingly, the significant clustering of distinct cellular integration events hitting CISs in different circulating lymphocytes indicates that in vivo selection of transduced cells in the clinical setting occurs in relation to vector insertion and may critically influence an individual cell's repopulation and proliferation capacity.

## Results

*Distribution analysis of retrovirus vector insertions in patients' mature blood cells.* To study the characteristics of retroviral insertion in clinical common γ chain (γc) gene correction, a high-throughput analysis of insertion sites was conducted by linear amplification–mediated PCR (LAM-PCR) (21–23) on the DNA of whole blood leukocytes (554 sites) and purified peripheral blood T cells ($CD3^+$), granulocytes ($CD15^+$), and monocytes ($CD14^+$; a total of 18 sites) collected 4 to 41 months after the reinfusion of autologous $CD34^+$ cells transduced with a γc encoding retrovirus vector. Concerning the purified cells, 6 of the 18 sites were analyzed in detail previously (21). We retrieved 704 unique insertion site sequences from the 9 analyzed patients, of which 572 (81%; Supplemental Table 1; supplemental material available online with this article; doi:10.1172/JCI31659DS1) could be mapped unequivocally to the human genome (see Methods). Chromosomal distribution analysis demonstrated that the frequency of insertion sites detected for each of the 23 human chromosomes correlated well with gene content but not with chromosome size (Figure 1A). Insertions were most frequent on chromosome 1, which is the largest chromosome, and least frequent on chromosomes Y and 18. At the same time, the high insertion site frequency on chromosomes 17 and 19 correlated with a higher-than-average number of genes on these chromosomes. Of the 572 unique RISs, 216 (38%) were located within a RefSeq gene, 157 (27%) were within 5 kilobase pairs (kbp) surrounding the TSS, and 356 (62%) were located in the gene coding sequence or less than 10 kbp away (Figure 1, B and C, Table 1, and Supplemental Table 1). Insertion data sets of the 3 patients (Pt4, Pt5, and Pt10) that developed a vector-associated T cell acute lymphocytic leukemia–like (T-ALL–like) disorder 30–34 months after gene therapy were analyzed separately (20). Their integration pattern was not found to be significantly different for any of the assessable parameters compared with that of the other patients (Table 1).

*RIS distribution in transduced $CD34^+$ cells.* To study the influence of the differentiation process on the distribution of insertion sites, we compared the insertion site distribution of transduced preinjection $CD34^+$ cells (total RISs, 167; mappable RISs, 102) with the profile found in the sorted circulating cell population (total RISs, 191; mappable RISs, 141) of the same patient, Pt4. We did not observe any substantial difference in the frequencies of gene-associated insertions between pre- and posttransplantation cells (49% versus 41%; $P = 0.22$, $\chi^2$ test), of targeting the TSS (within 5 kbp of TSS, 16% versus 26%; $P = 0.05$), of insertions in the proxim-

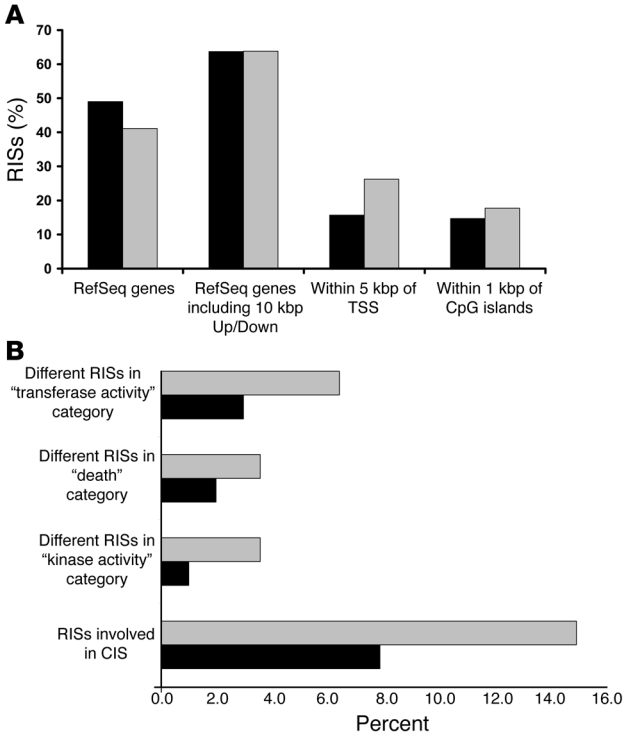


**Figure 2**
Comparison of pre- and posttransplant RIS distribution in Pt4. (**A**) Percentage of RISs detected in the indicated gene regions. (**B**) Distribution of vector-targeted genes (including the surrounding 10-kbp genomic region) with respect to GO and CIS formation. The GO categories were chosen according to the most significantly overrepresented ones retrieved from engrafted cells from all patients. Black bars, pretransplantation samples of Pt4 (102 RISs); gray bars, posttransplantation samples of Pt4 (141 RISs).

**Table 2**
CISs of third and higher order detected in patients

| | Pt1 (56) | Pt2 (101) | Pt4 (141) | Pt5 (52) | Pt6 (23) | Pt7 (94) | Pt8 (79) | Pt9 (9) | Pt10 (17) |
|---|---|---|---|---|---|---|---|---|---|
| **Protooncogenes** | | | | | | | | | |
| *CCND2* | 2 | 1 | 3 | 2 | | 1 | | | |
| *ZNF217* | | | 2 | 1 | | 1 | 3 | | 1 |
| *LMO2* | 1 | | 2 | 1 | | | 1 | | |
| *NOTCH2* | 2 | 1 | | | | | | | |
| *RUNX3* | | | 2 | | | | 1 | | |
| *RUNX1* | 1 | 2 | | | | | | | |
| **Other genes** | | | | | | | | | |
| *C14orf4* | | 1 | 1 | 2 | | | | | |
| *AFTIPHILIN* | | | 2 | | | 1 | | | |
| *FAM9C* | | 2 | | | | | 1 | | |
| *PDE4B* | 1 | | | 1 | | | | | 1 |
| *PRKCBP1* | | 1 | | | | | 2 | | |
| *PTPRC* | | | 1 | 1 | | 1 | | | |
| *TOMM20* | 1 | | | | | 1 | 1 | | |
| *TSRC1* | | | 1 | | 1 | 1 | | | |

The nearest RefSeq gene and the distribution of integrations among the different patients are shown for all CISs formed of at least 3 individual integrants. Numbers in parentheses denote the number of unique integrants retrieved from the individual patient.

ity of RefSeq genes and their 10-kbp upstream and downstream vicinity (64% versus 64%; $P = 0.98$), and of targeting CpG islands (14.7% versus 16.3%; $P = 0.73$; Figure 2).

*Vector integration is clustered in CISs.* For the purpose of analyzing high-throughput insertional mutagenesis models in mice, a nonrandom insertion clustering in the form of retrovirus integration into the same genomic locus on 2 or more different cells has been defined as a CIS. A CIS has been shown to be indicative of a nonrandom functional association of the insertion locus with the transformation event (24–26). To distinguish random coincidence of neighboring integration from nonrandom CIS formation, we followed a more stringent CIS definition as recently defined by Suzuki et al. (26). We classified CISs only by distance, independently of whether vector integrants were inter- or intragenic. We considered 2, 3, or 4 insertions to be CISs if they fell within a 30-kbp, 50-kbp, or 100-kbp window, respectively. CISs of fifth or higher order were defined by a 200-kbp window. Computer simulations showed that with 572 unique mappable RISs, the average number of randomly occurring second-order CISs (formed by 2 individual integrants) was 3.2 (Supplemental Table 2 and Methods). The null hypothesis that the 102 observed CISs of second order were the result of random clustering could be rejected (estimated $P$ value, 0). No CIS of third order (CISs formed by 3 integrants) or higher was obtained in 10,000 simulation runs.

Of the 572 mappable unique insertions found in blood cells, 122 (21.0%) were part of a CIS (Supplemental Table 3), which is 33-fold the value to be expected under random distribution of the RISs. Of the 47 different loci harboring CISs, 38 (81%) were closer than 30 kbp in distance to the next RefSeq gene. Among the 47 different CIS loci, 11 were known protooncogenes, involved in human chromosomal translocations described in acute leukemia or other forms of cancer: *ZNF217*, *VAV-3*, *CCND2*, *LMO2*, *MDS1*, *BCL2L1*, *NOTCH2*, *SOCS2*, *RUNX1*, *RUNX3*, and *SEPT6*. Of these, 9 are well-known transcription factors involved in human hematopoiesis. Fourteen particularly relevant CISs consisted of 3 or more integrants, the majority (10 of 14, 71%) of which localized less than 30 kbp away from genes. Here, protooncogene insertion was found in nearly half (6 of 14, 43%; Table 2). Of note, 3 CISs with 5 (*LMO2*), 8 (*ZNF217*), and 9 insertions (*CCND2*) accounted for 22 (4%) of all independent RISs, suggesting that they confer a strong selective advantage to the cell clones harboring these RISs.

Furthermore, we looked for the appearance of clones during the investigation period. Of all CIS clones, 11 of 122 single clones were detected at different time points, whereas only 28 of 450 non-CIS clones were retrieved more than once over time. Most of them appeared between 6 and 13 months and could also be detected later than 30 months, especially in the case of CIS clones. This shows that constant contribution of single clones to normal hematopoiesis plays an important role. The CIS clones are not exclusively responsible for the success of the gene therapy, but they may play an important role.

In the $CD34^+$ cells of Pt4 prior to transplantation, we identified 4 CISs (7.8%) of second order of the 102 unique RISs (Supplemental Table 3), compared with an expected value of 0.03 CISs. Computer simulations only reached a maximum of 3 CISs in 10,000 runs (mean, 0.098; median, 0; standard deviation, 0.31; $P = 0$; see Methods). This nonrandom integration could indicate that these CISs are particularly accessible, but it was substantially lower than in posttransplantation samples.

We could not distinguish RISs in patients with lymphoproliferation from those without: CISs of third order or higher were spread over these 2 groups of patients. Among the 37% of all integrations derived from lymphoproliferative patients, only 24% of CISs of second order were found, whereas 76% were found in leukemic and healthy patients or only in healthy patients.

*RISs are located next to growth-promoting genes.* To characterize the potential biological influence of vector integration on clonal selection, we used the gene ontology (GO) database and related EASE software (see Methods) to classify each gene into defined functional and biological categories. Any category reflects the percentage of a gene category in the GO database. While we did not find any overrepresented gene classes ($P < 0.05$, Fisher exact test, count

**Table 3**
GO classification

| Level | Category | List hits | *P* |
|---|---|---|---|
| **Molecular function** | | | |
| 2 | Kinase activity | 25 | 0.00018 |
| 2 | Receptor signaling protein activity | 10 | 0.000574 |
| 3 | Protein kinase activity | 20 | 0.000244 |
| 3 | Transferase activity, transferring phosphorous-containing groups | 25 | 0.000373 |
| 3 | DNA binding | 46 | 0.000398 |
| 4 | Phosphotransferase activity, alcohol group as acceptor | 23 | 0.000111 |
| 4 | Protein serine/threonine kinase activity | 15 | 0.000717 |
| **Biological process** | | | |
| 2 | Death | 17 | 0.000657 |
| 3 | Phosphorus metabolism | 24 | 0.000542 |
| 3 | Cell death | 17 | 0.000601 |
| 4 | Phosphate metabolism | 24 | 0.000542 |
| 4 | Intracellular signaling cascade | 26 | 0.00122 |
| 4 | Programmed cell death | 17 | 0.000315 |
| 4 | Cell proliferation | 29 | 0.00162 |
| 5 | Apoptosis | 17 | 0.000305 |
| 5 | Protein amino acid phosphorylation | 18 | 0.00194 |

RefSeq genes that received an insertion hit within the gene or the surrounding 10 kbp were used for GO analysis. Of 356 affected genes identified in engrafted cells, 164 could be analyzed regarding their molecular function, and 189 could not be analyzed regarding the biological process according to GO terms. *P* values were calculated by Fisher exact test. Levels indicate the specificity of the gene category term: the higher the level, the more precise the term of the gene category is, and the more specific the function of its genes. Levels range between 1 and 5; for some genes, there are more than 5 levels. Genes of a higher level also belong to the lower-level categories.

threshold of 3) in the transduced pretransplant samples, insertion analysis of engrafted cells showed highly significant overrepresentation of genes involved in phosphorus metabolism, cell survival, kinase activity, transferase activity, receptor signaling, and DNA binding (Table 3). We did not find any significant differences between patients with and without lymphoproliferation.

Further comparative analysis showed an accumulation of RISs in or near genes listed in the database of the cancer genome project (http://www.sanger.ac.uk/genetics/CGP/; Supplemental Table 1). Of the 356 total genes listed, 31 (9%) vector-targeted genes were known oncogenes. These data underline an integration-related selective advantage of RISs located in the vicinity of growth-promoting genes.

*RIS and CIS loci correlate to the gene expression profile of transduced cells*. To test whether the expression of genes is associated with the likelihood of receiving a retrovirus insertion, we analyzed insertions in gene loci as a function of the corresponding gene expression levels in CD34$^+$ cells, relative to the expression levels of all other genes. RISs in engrafted cells were significantly more frequently among the genes with the highest expression levels in CD34$^+$ cells ($n = 422$; $P < 1 \times 10^{-6}$, Cochran-Armitage test; Figure 3A). We further analyzed insertions in pretransplant CD34$^+$ cells from Pt4. Interestingly, although the association was significant, it was less pronounced than that observed in the in vivo setting ($n = 83$; $P = 4.99 \times 10^{-4}$, Cochran-Armitage test; Figure 3B).

CIS location correlated even better with the genes highly expressed in CD34$^+$ cells (Supplemental Table 3). Of 47 CIS genes, 43 could be analyzed because they were represented on the microarrays. The average expression bin was 6.8. With the exception of *FAM9C*, *PDE4B*, and *TSRC1* (average expression bins, 0.7, 3.3, and 4.66, respectively), 11 of 14 genes associated with CISs of 3 or more integrants were found to be in the highest quartile of expression (average expression bin, 7.1). *LMO2*, *PTPRC*, *TOMM20*, *PRKCBP1*, and *RUNX1* were among the 10% of genes with highest expression, in bin 9.

## Discussion

To understand the biology of insertional gene transfer in clinical trials, we performed high-throughput insertion site mapping on samples derived from a clinical gene therapy trial for SCID-X1. We compared RIS distribution in circulating mature cell populations from patients who had developed a lymphoproliferative adverse event and those who had not. Overall RIS distribution did not differ between the 2 groups. Both revealed the expected distribution features of retroviral vectors, with a strong preference for gene coding regions and symmetrical accumulation close to the TSS. Similar to that previously reported by Wu et al. for HeLa cells (7) and by Laufs et al. for CD34$^+$ cells (8), the frequency of RISs was more closely related to gene density than to overall chromosome size, most frequently targeting chromosomes 1, 17, and 19.

Compared with the distribution in pretransplant cells, in vivo repopulation and normal function of the corrected T cell pool led to a significant skewing of the RIS distribution. Of all RISs detected in posttransplantation blood samples, 21% were found to be clustered, and a much lower CIS frequency in the CD34$^+$ pretransplantation sample (7.8%) was observed. The observed changes in RIS distribution indicate that nonrandom selection or other biological effects of insertions in or near CIS genes have strong influence on the in vivo fate of gene-corrected cell clones.

Because the pre- and posttransplantation samples of Pt4 were comparable in size (102 RISs versus 141 RISs) and the CD34$^+$ cell culture conditions were identical to those used on the CD34$^+$ cells that engrafted and produced the T cells, the results of this analysis are adequate. Several mechanisms may account for the differences between insertion distribution profiles in pre- versus posttransplantation samples. First, the majority of cells in the pretransplantation sample have no repopulating ability. Therefore, the insertion site distribution of this population is not completely representative of repopulating cells from which posttransplantation cells derive. Second, posttransplantation CISs were even more frequently found near genes related to cell growth than were posttransplantation RISs. Consequently, integration sites in lymphocytes and their progenitor cells are not only related to the gene expression status at the time of vector entry into the repopulating target cell, but might additionally confer a selective advantage, most likely as a result of gene activation, in gene loci that govern growth and/or survival of CD34$^+$ cells and T cell precursors.

This observation was further corroborated by our analysis of whether the catalog of gene-associated insertions correlated with the target cells' gene expression pattern. In samples obtained after transplantation, there was an even higher correlation among the level of gene expression present in CD34$^+$ cells, the population initially targeted by the transduction, and the RIS frequency than in the analyzed pretransplant sample. The relevance of this association and its influence on clonal selection of engrafted cells is obvious in CISs with 3 or more RISs, where nearly 80% of CISs affect genes of the highest expression quartile in the engrafted gene-corrected cells.

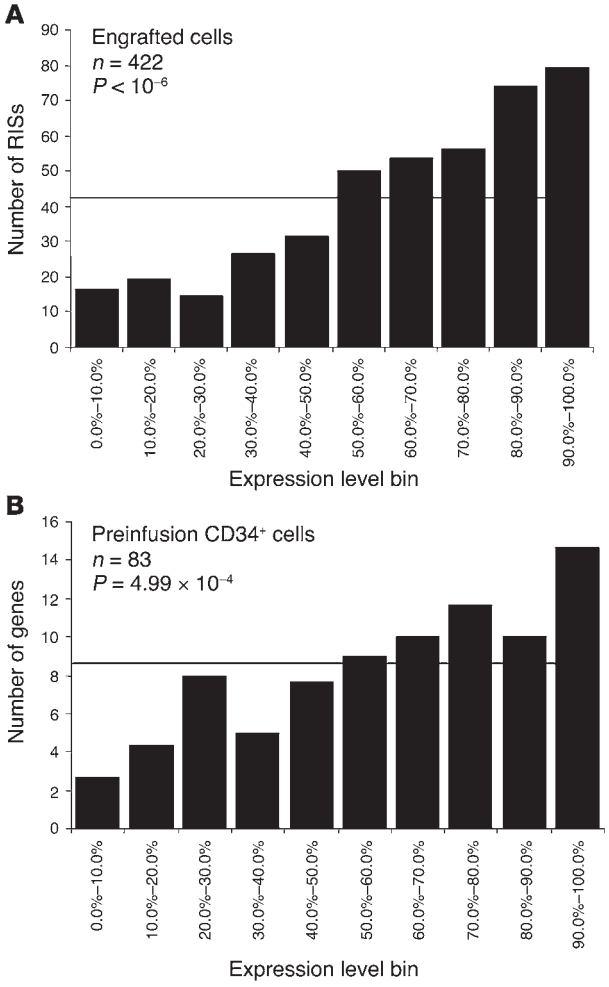


**Figure 3**
Association between vector integration and gene expression. (**A** and **B**) Number of RISs detected in engrafted cells (**A**) and in CD34+ cells prior to reinfusion (**B**) as a function of relative gene expression in stimulated peripheral blood CD34+ cells. For each gene, the probeset with the highest expression value was used. All 20,600 genes present on the array were sorted on expression and divided in 10 percentile categories according to their expression level, so that each category contains 10% of the genes. Values represent the average number of genes in each category based on 3 individual arrays (see Methods).

The results of our GO analysis provide further strong evidence that the biological function of genes at the insertion site is related to the in vivo fate of cell clones. When grouping vector-targeted genes according to their role in cellular physiology, engrafted cells show a clear preponderance of RISs located in or near growth-promoting genes, in particular genes revealing kinase and transferase activity. This feature was not seen with the pretransplant samples, indicating that in vivo selection of clones having integrants in or near growth-promoting genes occurred in our patients.

In line with this observation, more than two-thirds of the detected CIS genes were related to cell signaling and growth regulation or control of cell cycle, tyrosine kinases, or differentiation. The most frequent CIS-associated genes — *CCND2*, a cyclin found deregulated in a number of human cancer cells (27, 28); *ZNF217*, a zinc finger transcription factor hyperexpressed in solid tumors (29); and *LMO2*, a T-ALL related protooncogene (30) — are well known to influence clonal proliferation and survival if activated. Together, these areas represent 3% of all clones but only $7 \times 10^{-7}$% of the genetic code. Aberrant expression in many of these CIS genes in the context of other genetic changes has been linked to human oncogenesis. However, while the presence of CISs indicates that such clones engrafted and/or grew better than others, no evidence of clonal dominance has been detectable in the analyzed samples.

In a T cell gene transfer trial, RIS distribution was similar between clinical in vivo and experimental in vitro samples (14). To test whether pretransplant RIS distribution would have discernible characteristics related to a later lymphoproliferation event, we studied the integration sites in the CD34+ cell population cryopreserved immediately after the transduction phase for Pt4, the first patient who developed a *LMO2*-associated T-ALL–like disease. No *LMO2* RISs, and a low number of CISs, were found among the 102 sequences analyzed in CD34+ cells by LAM-PCR. In contrast, CISs were as frequent in posttransplantation T cells of Pt4 as in those of the other patients, with *CCND2*-related insertions being the most frequent CISs in this patient. In addition, no *LMO2* was detected in a second SCID-X1 trial. The results are published as a related manuscript by Schwarzwaelder et al. (31). Our findings support the concept that insertional activation of CIS genes, even when providing a subtle selective advantage to transduced precursors, will not lead to uncontrolled proliferation in the absence of other genetic changes.

This latter hypothesis is compatible with our recent observation of clonal myeloid cell expansion in a clinical retroviral vector–based gene therapy trial to correct chronic granulomatous disease. We found that a nonrandom integration site distribution had developed by extensive expansion of progenitor cells with *MDS1/EVI1*-, *PRDM16*-, and *SETBP1*-related integration sites in 2 patients. Expression of these genes conferred a selective advantage to the transduced myeloid cells, leading to a 3- to 4-fold self-limiting expansion of the gene-corrected cell fraction (19). Subsequent to the submission of this manuscript, a fourth case of T-monoclonal lymphoproliferation occurred in our group of patients. This lymphoproliferation is under investigation.

Our data indicate that the retrovirus vector integration pattern in T cells following clinical gene transfer is nonrandomly distributed, correlates well with CD34+ target cell gene expression, and is characterized by highly significant clustering into multiple different CISs. These CISs preferentially map to growth-regulating genes expressed in CD34+ cells, highlighting that their integration occurs preferentially in active gene loci and that maintaining their activation in later cell generations by insertion in vector-targeted genes themselves or in the regulatory gene regions likely confers a clonal selection advantage compared with other sites that are rarely affected. Furthermore, the expression as such, but not the intensity of expression, might be influential on insertion. Vector integration in many different sites in our clinical SCID-X1 study has actively influenced the fate of corrected cell clones in vivo. Potential therapeutic advantages associated with the preferential growth of particular clones over time will be the subject of further investigation. Additional biosafety measures designed into vectors could include inactivation of the 3′ long-terminal repeat (3′LTR) enhancer activity, e.g., by use of retrovirus or lentivirus self-inactivating vectors and insulators. Thus, the prospects are excellent that it will be possible in the future to develop safety

measures for gene therapy of severe immunodeficiencies, cancer, and other diseases with limited therapeutic options that avoid or at least minimize unwanted gene activation. The excellent therapeutic success achieved in gene therapy trials can be maintained, while the probability of insertional side effects is substantially decreased.

## Methods

*Patients' cells*. Blood samples were obtained at various time points from patients enrolled in the SCID-X1 gene therapy trial (32). CD3 T cells, CD19 B cells, and CD14 monocytes were selected from patients' PBMCs by immunomagnetic columns (Miltenyi Biotec). Granulocytes (CD15) were sorted by fluorescence-activated cell sorting (BD). A $CD34^+$ cell sample from Pt4 was separated just prior to reinfusion. Genomic DNA was isolated from all cells using commercially available DNA isolation kits (QIAGEN). Informed consent was obtained from parents, and the study was approved by the Comité Consultatif de Protection des Personnes dans la Recherche Biomedicale (CCPPRB), Hôpital Cochin, Paris, France.

*Integration site analysis by LAM-PCR*. DNA derived from patients' blood cells (1–100 ng) were used for integration site sequencing as previously described (21). Biotinylated primers LTRIa (5′-TGCTTACCACAGATATCCTG-3′) and LTRIb (5′-ATCCTGTTTGGCCCATATTC-3′) were used for the preamplification of the vector-genome junctions. After magnetic capture, hexanucleotide priming, and a restriction digest with *Tsp*509I, a linker cassette was ligated at the 5′ end of the genomic sequence. First exponential amplification of the vector-genome junction was performed with linker cassette primer LCI and vector LTR-specific primer LTRII, followed by second exponential PCR with primers LCII and LTRIII (22, 23). LAM-PCR amplicons were purified, shotgun cloned into the TOPO TA vector (Invitrogen), and sequenced (GATC Biotech and Centre National de Sequencage). Alignment of the integration sequences to the human genome was carried out using the University of California Santa Cruz (UCSC) BLAT genome browser (http://genome.ucsc.edu/cgi-bin/hgBlat?command=start). The UCSC and Ensembl database (http://www.ensembl.org) was used to study the relation to annotated genome features. Unmappable sequences were either too short (<20 kbp) or showed no definitive hit or multiple hits on the human genome.

*Definition of CISs and statistics*. For the determination of CISs, we measured the distance between individual integrants independently of being located inside or outside of gene coding regions. We considered 2, 3, or 4 insertions as CISs if they fell within a 30-kbp, 50-kbp, or 100-kbp window from each other, respectively. Of note, 3 clusters of 5, 8, and 9 integrants (next RefSeq gene, *LMO2*, *ZNF217*, and *CCND2*) covered 40 kbp, 170 kbp, and 60 kbp of genomic DNA, respectively. The genomic window for CISs of fifth order and higher was set to 200 kbp.

Computer simulations (10,000 runs) on the haploid size of the human genome ($3.12 \times 10^9$ kbp) were performed to calculate the likelihood of random, coincidental insertions. We counted the number of CISs of second order formed by 2 integrants within a 30-kbp window, the number of CISs of third order formed by 3 integrants within a 50-kbp window, the number of CISs of fourth order formed by 4 integrants within a 100-kbp window, and the number of CISs of higher orders within a 200-kbp window. Of note, CISs of different orders were analyzed independently of each other, e.g., CISs formed by 3 integrants located within 20 kbp were counted as 3 CISs for the calculation of CISs of second order and as 1 CIS for the calculation of CISs of third order (Supplemental Tables 2 and 3).

*Transcription profile in $CD34^+$ cells*. G-CSF–mobilized peripheral blood $CD34^+$ cells from 3 donors were cultured using the same conditions as performed in the original gene therapy trial (1) and served as 3 independent and individual sample sources for further RNA expression analysis. RNA was isolated using Tri Reagent (Sigma-Aldrich) according to the manufacturer's protocol. The mRNA expression levels were determined using Affymetrix U133 Plus 2.0 arrays and normalized as described previously (33). The normalized microarray values were sorted upwardly on expression and divided into 10 equal-sized expression level categories, designated 0 through 9. The presence of the gene closest to a vector integration site as identified by LAM-PCR analysis was determined in each expression level category. A Cochran-Armitage test for trend was performed to determine whether higher expression level categories corresponded to larger numbers of insertions (34). For all gene symbols on the array, the highest expression values were used to describe the gene expression.

*GO analysis*. To classify vector targeted genes according to GO terms, we analyzed RefSeq genes that were hit by vector or had vector integration in the surrounding 10-kbp genomic region. GO analysis was performed using the publicly available EASE software from NIH-DAVID (http://david.abcc.ncifcrf.gov/ease/ease.jsp). The database sorts the genes in categories according to GO terms regarding their "molecular function," "biological process," and "cellular compartment." The gene categories are divided in different levels. Level 1 is a rather general category; this group is composed of many genes. The higher the level, the more precise the parameters, and the more specific the function of its genes. With a level of 3 or 4 there will be a good balance between the amount of listed hits and sufficient specificity. Genes of a higher level category also belong to categories of a lower level. The analysis compares which gene categories were detected more frequently than others compared with their likelihood of detection if insertion was distributed evenly across the entire human genome. Overrepresented gene categories were determined by Fisher exact test. An overrepresentation was given for $P$ values less than 0.05 compared with the whole human genome as a background.

## Acknowledgments

Funding was provided by the European Commission (5th and 6th Framework Programs, Contracts QLK3-CT-2001-00427-INHERINET and LSHB-CT-2004-005242-CONSERT), by NIH grant R01 CA 112470-01, by Deutsche Forschungsgemeinschaft (DFG) grants Ka976/5-3 and Ka976/6-2, by INSERM, l'Assistance Publique des Hôpitaux de Paris (AP-HP), and by Agence Nationale de la Recherche (ANR) grant 05-MRAR.004.

Received for publication January 30, 2007, and accepted in revised form May 29, 2007.

Address correspondence to: Christof von Kalle, National Center for Tumor Diseases, Im Neuenheimer Feld 350, 69120 Heidelberg, Germany. Phone: 49-6221-56-6990; Fax: 49-6221-56-6967; E-mail: christof.kalle@nct-heidelberg.de.

Annette Deichmann, Salima Hacein-Bey-Abina, Manfred Schmidt, and Alexandrine Garrigue contributed equally to this work.
Christof von Kalle and Marina Cavazzana-Calvo are co–senior authors.

1. Cavazzana-Calvo, M., et al. 2000. Gene therapy of human severe combined immunodeficiency (SCID)-X1 disease. *Science.* **288**:669–672.
2. Aiuti, A., et al. 2002. Correction of ADA-SCID by stem cell gene therapy combined with nonmyeloablative conditioning. *Science.* **296**:2410–2413.
3. Gaspar, H.B., et al. 2004. Gene therapy of X-linked severe combined immunodeficiency by use of a pseudotyped gammaretroviral vector. *Lancet.* **364**:2181–2187.
4. Coffin, J.M., Hughes, S.H., and Varmus, H.E. 1997. *Retroviruses*. Cold Spring Harbor Laboratory Press. Plainview, New York, USA. 843 pp.
5. Moolten, F.L., and Cupples, L.A. 1992. A model for predicting the risk of cancer consequent to retroviral gene therapy. *Hum. Gene Ther.* **3**:479–486.
6. Schröder, A.R., et al. 2002. HIV-1 integration in the human genome favors active genes and local hot-

spots. *Cell.* **110**:521–529.
7. Wu, X., Li, Y., Crise, B., and Burgess, S.M. 2003. Transcription start regions in the human genome are favored targets for MLV integration. *Science.* **300**:1749–1751.
8. Laufs, S., et al. 2003. Retroviral vector integration occurs in preferred genomic targets in human bone marrow repopulating cells. *Blood.* **101**:2191–2198.
9. Mitchell, R.S., et al. 2004. Retroviral DNA integration: ASLV, HIV, and MLV show distinct target site preferences. *PLoS Biol.* **2**:e234.
10. Mooslehner, K., Karls, U., and Harbers, K. 1990. Retroviral integration sites in transgenic Mov mice frequently map in the vicinity of transcribed DNA regions. *J. Virol.* **64**:3056–3058.
11. Scherdin, U., Rhodes, K., and Breindl, M. 1990. Transcriptionally active genome regions are preferred targets for retrovirus integration. *J. Virol.* **64**:907–912.
12. Bushman, F.D. 2003. Targeting survival: integration site selection by retroviruses and LTR-retrotransposons. *Cell.* **115**:135–138.
13. Maxfield, L.F., Fraize, C.D., and Coffin, J.M. 2005. Relationship between retroviral DNA-integration-site selection and host cell transcription. *Proc. Natl. Acad. Sci. U. S. A.* **102**:1436–1441.
14. Recchia, A., et al. 2006. Retroviral vector integration deregulates gene expression but has no consequence on the biology and function of transplanted T cells. *Proc. Natl. Acad. Sci. U. S. A.* **103**:1457–1462.
15. Hematti, P., et al. 2004. Distinct genomic integration of MLV and SIV vectors in primate hematopoietic stem and progenitor cells. *PLoS Biol.* **2**:e423.
16. Du, Y., Jenkins, N.A., and Copeland, N.G. 2005. Insertional mutagenesis identifies genes that promote the immortalization of primary bone marrow progenitor cells. *Blood.* **106**:3932–3939.
17. Kustikova, O., et al. 2005. Clonal dominance of hematopoietic stem cells triggered by retroviral gene marking. *Science.* **308**:1171–1174.
18. Calmels, B., et al. 2005. Recurrent retroviral vector integration at the Mds1/Evi1 locus in nonhuman primate hematopoietic cells. *Blood.* **106**:2530–2533.
19. Ott, M.G., et al. 2006. Correction of X-linked chronic granulomatous disease by gene therapy, augmented by insertional activation of MDS1-EVI1, PRDM16 or SETBP1. *Nat. Med.* **12**:401–409.
20. Hacein-Bey-Abina, S., et al. 2003. LMO2-associated clonal T cell proliferation in two patients after gene therapy for SCID-X1. *Science.* **302**:415–419.
21. Schmidt, M., et al. 2005. Clonal evidence for the transduction of $CD34^+$ cells with lymphomyeloid differentiation potential and self-renewal capacity in the SCID-X1 gene therapy trial. *Blood.* **105**:2699–2706.
22. Schmidt, M., et al. 2002. Polyclonal long-term repopulating stem cell clones in a primate model. *Blood.* **100**:2737–2743.
23. Schmidt, M., et al. 2003. Clonality analysis after retroviral-mediated gene transfer to $CD34^+$ cells from the cord blood of ADA-deficient SCID neonates. *Nat. Med.* **9**:463–468.
24. Mikkers, H., et al. 2002. High-throughput retroviral tagging to identify components of specific signalling pathways in cancer. *Nat. Genet.* **32**:153–159.
25. Lund, A.H., et al. 2002. Genome-wide retroviral insertional tagging of genes involved in cancer in Cdkn2a-deficient mice. *Nat. Genet.* **32**:160–165.
26. Suzuki, T., et al. 2002. New genes involved in cancer identified by retroviral tagging. *Nat. Genet.* **32**:166–174.
27. von Eyben, F.E. 2004. Chromosomes, genes, and development of testicular germ cell tumors. *Cancer Genet. Cytogenet.* **151**:93–138.
28. Hideshima, T., Bergsagel, P.L., Kuehl, W.M., and Anderson, K.C. 2004. Advances in biology of multiple myeloma: clinical applications. *Blood.* **104**:607–618.
29. Collins, C., et al. 2001. Comprehensive genome sequence analysis of a breast cancer amplicon. *Genome Res.* **11**:1034–1042.
30. Nam, C.H., and Rabbitts, T.H. 2006. The role of LMO2 in development and in T cell leukemia after chromosomal translocation or retroviral insertion. *Mol. Ther.* **13**:15–25.
31. Schwarzwaelder, K., et al. 2007. Gammaretrovirus-mediated correction of SCID-X1 is associated with skewed vector integration site distribution in vivo. *J. Clin. Invest.* **117**:2241–2249. doi:10.1172/JCI31661.
32. Hacein-Bey-Abina, S., et al. 2002. Sustained correction of X-linked severe combined immunodeficiency by ex vivo gene therapy. *N. Engl. J. Med.* **346**:1185–1193.
33. Dik, W.A., et al. 2005. New insights on human T cell development by quantitative T cell receptor gene rearrangement studies and gene expression profiling. *J. Exp. Med.* **201**:1715–1723.
34. Armitage, P., Berry, G., and Matthews, J.N.S. 2001. *Statistical methods in medical research*. 4th edition. Blackwell Publishing. Malden, Massachusetts, USA/Oxford, United Kingdom. 832 pp.