Bone Marrow Transplantation (2006) 38, 229–235
© 2006 Nature Publishing Group All rights reserved 0268-3369/06 $30.00
www.nature.com/bmt

ORIGINAL ARTICLE

Retroviral vector insertions in T-lymphocytes used for suicide gene therapy occur in gene groups with specific molecular functions

FA Giordano[1], B Fehse[2], A Hotz-Wagenblatt[3], S Jonnakuty[3], C del Val[3], J-U Appelt[1], KZ Nagy[1], K Kuehlcke[4], S Naundorf[4], AR Zander[2], WJ Zeller[1], AD Ho[5], S Fruehauf[5] and S Laufs[1]

[1]*Research Program Innovative Cancer Diagnostics and Therapy, German Cancer Research Center (DKFZ), Heidelberg, Germany;* [2]*Department of Bone Marrow Transplantation, University Hospital Eppendorf, Hamburg, Germany;* [3]*Department of Molecular Biophysics, German Cancer Research Center (DKFZ), Heidelberg, Germany;* [4]*EUFETS AG, Idar-Oberstein, Germany and* [5]*Department of Internal Medicine V, University of Heidelberg, Heidelberg, Germany*

Graft-versus-host disease (GvHD) is a severe complication in the context of allogeneic stem cell transplantation and adoptive immunotherapy. The transfer of a suicide gene into donor T-lymphocytes (TLCs) allows selective elimination of GvHD-causing cells. As retroviral gene transfer into hematopoietic stem cells can induce leukaemia, there is an urgent need also to analyze retroviral integration sites in TLCs. We examined suicide gene-transduced TLCs in four grafts and from four transplanted patients. One-hundred and fifteen integration sites were detected *in vitro*. Of these 90 could be mapped to the human genome; 50% (45) were located in genes and 32% (29) were detected 10 kb upstream or downstream of transcription start sites. We found a significant overrepresentation of genes encoding for proteins with receptor activity, signal transducer activity, transcription regulator activity, nucleic acid binding activity and translation regulator activity. Similar data were obtained from patient samples. Our results point to preferred vector integration patterns, which are specific for the target cell population and probably independent of selection processes. Thus, future preclinical analysis of the integration repertoire with abundant amounts of transduced cells could allow a prediction also for the *in vivo* situation, where target cells are scarce.
Bone Marrow Transplantation (2006) **38,** 229–235. doi:10.1038/sj.bmt.1705424; published online 19 June 2006
Keywords: GvHD; T-lymphocytes; suicide gene therapy; HSV-TK; ligation-mediated PCR; preferred integration

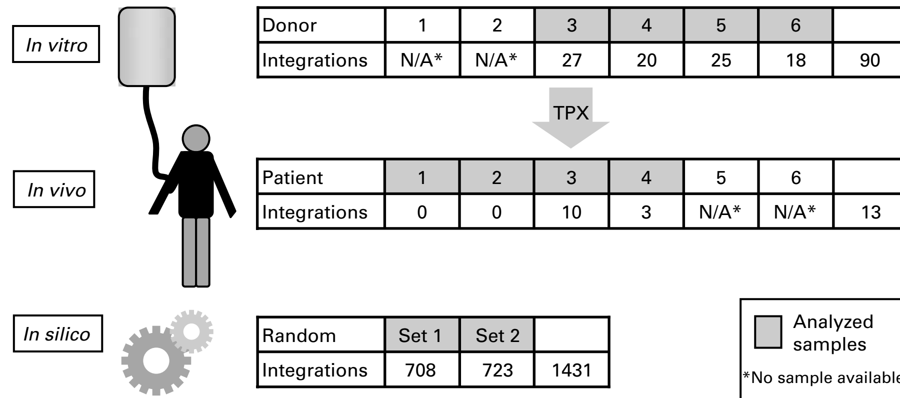
Introduction

Adoptive immunotherapy was shown to be an efficient approach to treat hematological malignancies in the context of allogeneic bone marrow transplantation.[1] Donor lymphocyte infusions (DLIs) improve immune reconstitution and can lead to the destruction of (remaining) malignant cells, termed graft-versus-leukemia effect.[2] Yet, infusion of donor T-lymphocytes (TLCs) is limited by the development of graft-versus-host disease (GvHD), which remains a dreaded complication.[3] For the safety of the patient, it is therefore mandatory to control GvHD once an antileukemic effect has been achieved. One strategy directed towards the control of GvHD is based on the introduction of a 'suicide gene' into TLCs, allowing their later selective removal by administration of prodrugs, which upon activation induce apoptosis.[4] Stable gene integration and expression can be achieved by retroviruses that integrate their vectors in the host genome. One vector–prodrug combination is the herpes simplex virus thymidine kinase (HSV-TK) gene and ganciclovir (GCV) application, whose efficacy and feasibility has been proven in several clinical studies.[5–7]

As in a murine retroviral gene marking study[8] and in a clinical gene therapy trial for X-chromosomal severe combined immunodeficiency (SCID),[9] insertional oncogene activation and subsequent leukemia development have been described; there is an urgent need for assessing the distribution of vector insertion sites also in terminally differentiated cells such as primary TLCs, where alterations of gene expression have been reported previously.[10]

Here we performed integration site analysis on suicide gene-modified (SGM) donor TLC prepared for a clinical phase I/II trial. We investigated cells of four different donors, as well as cells of four different patients, who received DLIs using ligation-mediated polymerase chain reaction (LM-PCR). Thus, for the first time the analysis of genomic integration sites from TLCs used for HSV-TK gene therapy was performed with the novel standardized IntegrationMap task[11] for samples obtained *in vivo* and *in vitro* (Figure 1). Additionally, we generated a random integration pattern ('*in silico*'), which was used for

Correspondence: Dr S Fruehauf, Department of Internal Medicine V, University of Heidelberg, Im Neuenheimer Feld 410, Heidelberg 69120, Germany.
E-mail: stefan_fruehauf@med.uni-heidelberg.de
Received 27 February 2006; revised 12 May 2006; accepted 13 May 2006; published online 19 June 2006



Donor	1	2	3	4	5	6	
Integrations	N/A*	N/A*	27	20	25	18	90

Patient	1	2	3	4	5	6	
Integrations	0	0	10	3	N/A*	N/A*	13

Random	Set 1	Set 2	
Integrations	708	723	1431

Figure 1 Experimental setup survey. Of six donor–patient couples included in this study, DNA was isolated and analyzed with highly sensitive LM-PCR from four donor T-lymphocyte infusion bags (*in vitro* assay) and from four graft recipients (*in vivo* assay). We found 90 integrations *in vitro* and 13 *in vivo*. Both assay results were compared to a randomly generated set consisting of 2000 sequences, where 1431 could be matched to the human genome.

comparison of integration site data. Targeted and adjacent genes were classified using the ExpressMiner tool, which is also a novel standardized algorithm for the description of gene functions.[12]

Materials and methods

Selection and retroviral transduction of TLCs

Following approval by the appropriate authorizing bodies and informed consent of both patients and donors, TLCs of six different donors were harvested by donor leukapheresis before the scheduled transplantation. Donor TLCs were stimulated with 10 ng/ml OKT-3 (Cilag, Neuss, Germany) and cultivated at a density of 2×10^6/ml in the presence of 100 U/ml interleukin (IL)-2 (Chiron, Emeryville, CA, USA) in X-Vivo10 containing 8% autologous serum. After 3 days of prestimulation, three transduction cycles were performed on cell culture flasks pre-loaded with retroviral particles containing the murine leukemia virus (MLV)-based vector Mo3TIN,[13] carrying the HSV-TK and the neomycin resistance (NeoR) gene with an multiplicity of infection of ~ 2 in a specialized good manufacturing proctive (GMP) facility as described before.[14] Suicide gene-modified (SGM) cells were selected from day 6 to 13 using 1 mg/ml G418 and cryopreserved in a freezing solution containing 5% dimethyl sulfoxide, 6% HES, 3.5% human serum albumin and 1% glucose in a concentration of $0.5\text{–}1 \times 10^7$ cells/ml. Before transplantation, all grafts passed safety tests[15] in a GMP facility (European Institute for Research and Development of Transplantation Strategies (EUFETS)). Whole-blood samples of four patients included in the study were taken at different time points after transplantation. Samples were pooled and concentrated owing to the small amount of DNA within each patient sample.

LM-PCR

For detection of retroviral integration sites, DNA was extracted from TLCs stored in infusion bags or whole blood obtained directly from patients (QiaAmp Blood Kit, Qiagen, Hilden, Germany). LM-PCR was performed as described before,[16] whereas for each donor bag, four LM-PCR reactions were performed directly on extracted DNA, and two reactions following pre-amplification of patients' DNA (REPLI-g Kit, Qiagen). Further integration site analysis was not possible owing to the limited amount of patient DNA.

*Random insertion site generation (*in silico *data)*

For later comparison of data in terms of preferential integration pattern, we created a random sequence set consisting of 2000 sequences. Therefore, random number values following a uniform distribution were generated in a way that the values are evenly spread over a given interval. The first interval (1–3121) was then divided into 24 parts, corresponding to the chromosomes with a partition size according to the size of the individual chromosome. The generated value will then give the chromosome number with respect to its size. Another random number (intervals 100–200) was generated to determine the length of the random sequence and a third one (interval: first base of the chromosome to last base of the chromosome) for the start position within the chromosome. With the resulting values for chromosome number, start position and length, the sequences were fetched from the EnsEMBL human genome database (http://www.EnsEMBL.org) using the same EnsEMBL draft the IntegrationMap task uses.[11] Additionally, calculated insertion sites were subjected to the same quality criteria as were the insertion sites found *in vitro* or *in vivo*.

Insertion site sequence analysis

The sequences detected with LM-PCR were first viewed using Chromas 2.23 software (Technelysium Pty Ltd, Tewantin, Australia). Sequence matches were judged to be authentic only if the matching part of the human query sequence was surrounded by the 5′-LTR sequence on the one side and the adapter sequence on the other side.

For fast and standardized analysis of sequences obtained from the LM-PCR reactions, we used the novel IntegrationMap task[11] (available online at http://genius.embnet.dkfz-heidelberg.de/menu/biounit/open-husar or for all services after registration for an account at http://genome.dkfz-heidelberg.de).

IntegrationMap first runs a MEGABLAST analysis using the NCBI 35b assembly. Thus, the location, that is, information about chromosomal position, contig position and strand was calculated. Then, using an application program interface (API) of Ensembl.org (http://www.ensembl.org/Docs, NCBI 35b assembly), the positions of the next gene (i.e. its transcription start site), the next repetitive elements (e.g. short-interspersed nucleotide elements (SINE), long-interspersed nucleotide elements (LINE), simple repeats, LTR elements) and the next CpG island were computed and their distances to the exact integration base were calculated. Thus, the IntegrationMap task retrieves detailed information about whether integrations are located in or close to genes, the name of the gene, the exact localization in the transcriptional units and further parameters like the distance from the transcription start site to the integration. As only those genes annotated in the EnsEMBL or RefSeq database are found by IntegrationMap, the data describe existing and validated genes. Predicted genes as obtained by GENSCAN[17] were not included in this study. Further information, for example, data about CpG Islands, LINEs or SINEs, and their distances to the integration is also displayed. Results of *in vitro* and *in vivo* experiments were then compiled and compared to the random set of 1431 integrations using a χ^2 test.

Gene classification

To classify hit and adjacent genes *in vitro* and *in vivo*, we used a novel standardized task, the ExpressMiner.[12] ExpressMiner first performs queries across many information sources like the NCBI databases – Locuslink and RefSeq,[18] UniGene,[19] Swissprot,[20] Kegg[21] and GeneOntology.[22] The quality of the data was ascertained by reducing the redundancy from the given list of entries, by applying checks to minimize annotation inconsistency across different databases, as well as by using combined queries with different identifiers (gene symbol, Unigene cluster ID, Locuslink ID). When possible, functional annotation, was established using gene ontology (GO),[22] which organizes GO terms in a directed acyclic graph (DAG) in a hierarchical manner. To obtain the GO functional annotation the ExpressMiner task constructed organism-specific 'GO slims' (the slimmed down versions of two ontology categories, molecular function and biological process, respectively). The mapping of genes to GO terms in the context of the DAG structure allows viewing of the GO categories at higher levels of the GO DAG. With all the obtained information, functional profiles were constructed, including information about their biological processes and location of cellular activity of the identified genes. For each GO category assigned to hit or adjacent genes, we performed a matched pair analysis with the corresponding category arising from random data using a χ^2 test.

Results

A phase I/II study of CD34-selected peripheral blood progenitor cell (PBPC) transplantation accompanied by infusion of SGM donor TLCs was performed.[15] To this end, peripheral blood TLCs were isolated from leukapheresis products of six donors. Following activation using IL-2 and OKT-3 cells, were retrovirally transduced[23] with the Mo3TIN vector,[13] carrying the HSV-TK and the neomycin phosphotransferase (NeoR) genes. Transduction efficiency ranged between 15 and 35%. Transduced cells were G418-selected, enriched to 100%, split into portions and cryopreserved. Following myeloablative conditioning, each patient received more than 4×10^6 cells/kg $CD34^+$ and approximately 5×10^6 $CD34^+$ cells/kg body weight SGM donor TLCs.

This study, together with previous work,[24–28] indicates that each vector type combined with different host cells produces a unique pattern of preferred integration target sites, which remains stable also *in vivo*.

Analysis of in vitro *samples*

Of 115 external bands, 90 sequences could be unambiguously mapped and characterized using the novel IntegrationMap tool.[11] In the four DLI bags examined, no identical clones were detected in any of the LM-PCR reactions, pointing to the abundance of clones present. Following primary data collection, retroviral insertion estimation into chromosomes (RISC) scores were calculated as described.[28] The frequency of insertions on chromosome 19 was found most obviously elevated ($P<0.05$), whereas integrations were highly underrepresented on chromosome 5 ($P<0.05$).

Of all 90 mapped insertion sites, 50% (45/90) hit EnsEMBL or RefSeq genes. The gene integration frequency was nearly two times higher than statistically expected, when compared to a gene insertion frequency of 28.5% *in silico* ($n=1431$ mapped clones to the human genome). These findings are in line with previous data, where 40% of retroviral integrations in nonobese diabetic (NOD)/SCID-repopulating cells[28] and 61% of all integrations in MLV-transduced HeLa cells were found in RefSeq genes.[26]

Analysis of the exact integration site within targeted genes showed that 41% (37/90) of the mapped integrations were located in introns; 14% (13/90) resided in the first one. Exons were targeted eight times, of which the first exon was hit three times. In a range of 10 kb upstream and downstream of the integration sites, we located 29 genes (32% of all integrations), two times more than *in silico* (Figure 2).

Using the GO assignment function of the ExpressMiner Task,[12] we assigned 'molecular function' to genes hit by retroviral integration, and to genes found in a distance up to ± 50 kb to the insertion site. Of all annotations in the GO category 'molecular function', we found that genes coding for proteins assigned to 'receptor activity', 'signal transducer activity', 'transcription regulator activity', 'nucleic acid binding' and 'translation regulator activity' to be significantly overrepresented ($P<0.05$) when compared to the *in silico* data. Unlike this, genes coding for proteins assigned to 'transferase activity' and 'kinase activity' were significantly less targeted than expected ($P<0.05$; Figure 3).

Using the IntegrationMap task, retroviral integrations were also analyzed in terms of distances to repetitive elements. We found preferred integration near CpG islands when compared to *in silico* data, but no statistically significant accumulation near SINEs (data not shown) or LINEs (data not shown).

Analysis of in vivo *samples*

When analyzing the recipients of PBSCs and DLIs we found 10 insertion sites in patient 3 ('case 3,'Fehse *et al.*,[15] respectively) and three insertion sites in patient 4 (Table 1). After two LM-PCR reactions on pre-amplified patients' DNA identical clones were found in each of the reactions.

Of all integrations, six vector insertions were located in genes (46% of all integrations). Hit genes were the alpha-sialyltransferase 8D gene (ST8Sia IV), the natural killer cell receptor 2B4 precursor gene (NKR2B4), the PKA-interacting protein gene (breast cancer-associated gene 3 protein, BCA3), the protein kinase C eta-type gene (PKC-L) and two times at different positions, the Dynamin-3 (T-dynamin). In three cases (50% of all integrations in

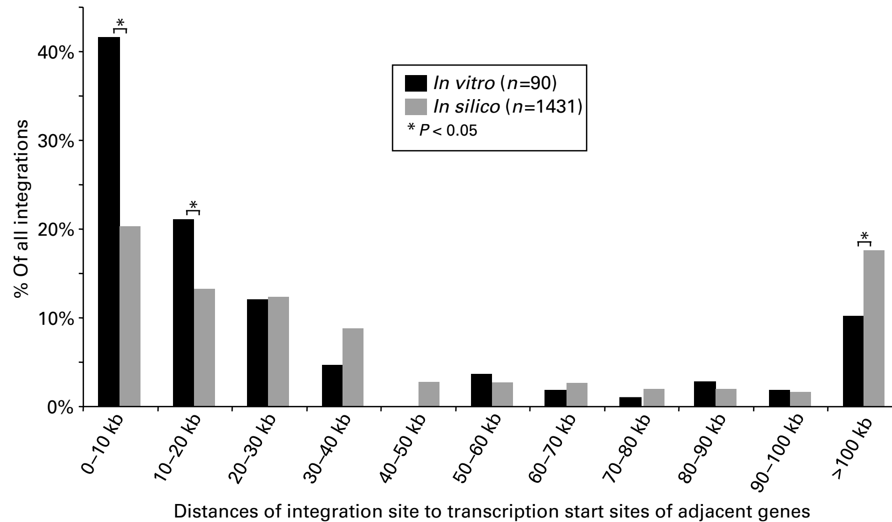


Figure 2 Distances of integration sites upstream and downstream to transcription start sites of genes. For each of the $n = 90$ integrations found in the graft samples (*in vitro* assay) distances to the next transcription start site of genes were calculated and displayed graphically in groups of 10 kb. A significant accumulation of integrations can be seen between 0 and 20 kb. In contrast, in a range of ± 100 kb from the insertion site, a significant lower number of transcription start sites (TSS) than expected was detected.

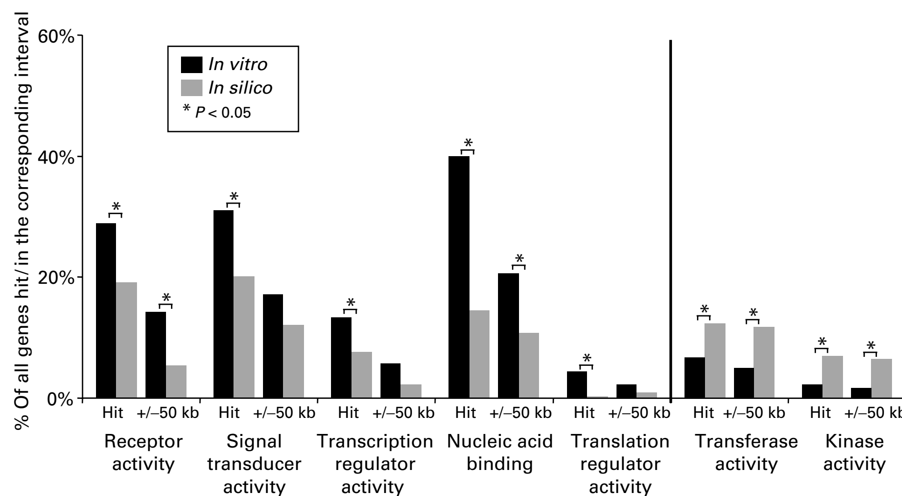


Figure 3 Selection of molecular functions of hit and adjacent genes. Genes were first compiled in sets regarding their distance to the insertion site of the vector. The resulting 'distance sets' were then consisting of hit and ± 50 kb adjacent genes. Following categorization using the GeneMiner task, all available annotations of the GO category 'molecular function' were obtained per distance set. The five groups of gene products found significantly overrepresented as well as the two groups of gene products found significantly underrepresented are shown here for all calculated distance sets. Genes coding for proteins assigned to the functions 'receptor activity', 'signal transducer activity', 'transcription regulator activity' and 'nucleic acid binding' and 'translation regulator activity' were found to be significantly overrepresented ($P \leqslant 0.05$). In contrast, genes coding for proteins assigned to 'transferase activity' and 'kinase activity' were fewer targeted than expected ($P < 0.05$).

Table 1 Patient data overview

Patient no.	*1*	*2*	*3*	*4*
Previous coding Fehse *et al.*[15]	Case 1	Case 2	Case 3	N/A
No. of LM-PCRs	2	2	2	2
Integrations	N/A	N/A	10	3
Integration in genes	N/A	N/A	ST8Sia IV NKR2B4 BCA3 PKC-L DNM-3 (2×)	No genes hit
Adjacent genes ±50 kb	N/A	N/A	CTCF PPCTI CBLB Unknown gene Unknown gene COPB	MC2H2 MCCDS

Samples of all patients were subjected to two LM-PCR reactions and analyzed. Ten integrations were found in patient 3, which corresponds to 'case 3' and three integrations in patient 4. Genes were only found as target in samples with regard to patient 1, whereas one gene was targeted two times. Genes that were not hit but adjacent are displayed when found within a distance of ±50 kb to the integration site.
Abbreviations: BCA3 = breast cancer-associated gene (ENSG00000166452); CBLB = signal-transduction protein CBL-B gene (ENSG00000198956); COPB = coatomer beta subunit gene (ENSG00000129083); CTCF = transcriptional repressor CTCF gene (ENSG00000153922); DNM-3 = dynamin 3 gene (ENSG00000197959); LM-PCR = ligation-mediated polymerase chain reaction; PPCTI = peptidyl-prolyl cis–trans isomerase gene (ENSG00000102974); MC2H2 = member of the C2H2 zinc-finger group gene (ENSG00000179965); MCCDS = member of the CCDS group gene (ENSG00000114423); NKR2B4 = natural killer cell receptor 2B4 precursor gene (ENSG00000122223); PKC-L = protein kinase C eta type gene (ENSG00000027075) ST8Sia IV = alpha-sialyltransferase 8D gene (ENSG00000113532).

genes), the hit genes belonged to the molecular function category 'signal transducer activity' (NKR2B4, BCA3, PKC-L), whereas NKR2B4 is a also assigned to 'receptor activity'. Thus, the insertion frequency into genes, especially in those with signal-transducing activity is significantly higher than expected ($P < 0.05$).

Within a range of ±50 kb to the insertion site, eight genes were detected that were not hit. Three of those (38% of all genes in this range), namely the C2H2-subtype zinc-finger gene, the transcriptional repressor CTCF gene (CCCTC-binding factor gene) and the signal-transduction protein CBL-B gene (RING finger protein 56 gene) were genes coding for zinc-finger proteins and were found assigned to the molecular function 'nucleic acid binding', whereas the transcriptional repressor CTCF has also 'transcription regulator activity', and the signal-transduction protein CBL-B 'signal transducer activity'. Compared to an expected *in silico* insertion frequency of 11% into genes encoding nucleic acid binding proteins, genes of this category are more than three times overrepresented *in vivo*.

Thus, our findings of preferred retroviral vector insertion sites *in vivo* correspond largely those found *in vitro*.

Discussion

It is of great importance from the biological and the safety points of view to identify specific clones and their genomic insertion patterns, including the functions of targeted genes, in patients participating in clinical studies. In this study, we performed retroviral integration site analysis of SGM cells, of donors from a phase I/II clinical HSV-TK trial.[15] We herewith confirm and extend recently published data on suicide gene vector integration in donor TLCs,[10] which showed preferential integration near transcription start sites and promoters. Using a different vector backbone and TLCs from a different clinical study, we extend the current knowledge with a more detailed analysis using a functional classification system of genes bearing integrations to detect host-specific patterns of integration.

We analyzed the gene modified TLCs *in vitro* and *in vivo*, as well as a randomly generated set of integrations with regard to vector integration patterns on the chromosomal and single gene level.

The 90 mapped integration sites showed preferred integration into chromosome 19, whereas chromosome 5 was significantly underrepresented. These results are in line with previous studies, where integrations also occurred with significantly increased frequency into chromosomes 17 and 19 when PBPCs were transduced with the retroviral SF91m3 vector.[16] It supports theories that gene-dense chromosomes like chromosome 19 can be seen as attractive targets for viral vectors. Furthermore, we found 50% of all integrations in EnsEMBL annotated genes, nearly two times more often than calculated by random vector distribution simulation ($n = 1431$ mapped to the human genome). This is also in line with previous data, where 40% of retroviral integrations in RefSeq genes of SF91m3-transduced NOD/SCID-repopulating cells have been found within RefSeq genes.[28]

Of all mapped integration sites, 42% were located closer than 10 kb to transcription start sites of genes. Compared to an *in silico* frequency of 21%, a significant accumulation of integrations near the transcription start site. This, together with data from Mitchell *et al.*,[26] suggests that MLV complexes could interact with factors bound near transcription start sites. In contrast to recent findings of our group, no statistically significant accumulation near SINEs or LINEs was observed.

Recent reports on preferred vector insertion[16,26–28] have also generated intense interest to describe the functions of the targeted genes. The GO annotations assigned to genes

provide a controlled and standardized vocabulary to describe gene and gene products attributes in any organism.[22] For the first time in this context, we used a tool, the ExpressMiner,[12] which can automatically assign GO annotations to targeted genes. This can be seen as an outstanding advantage compared to time-consuming manual data retrieval by querying databases. Using the assignment to all existing categories of 'molecular function' performed by this task, we found a significantly preferred insertion into and near genes coding for proteins with receptor activity, signal transducer activity, transcription regulator activity, nucleic acid binding ability and translation regulator activity (Figure 3).

The data of retroviral integration sites in TLC obtained from patients were generated at the detection limit of the LM-PCR method. For example, the sample of patient 3 (case 3, respectively) contained about 0.1% transduced cells each.[15] For detection of individual clones, a clonal size of at least 2–10 cells is required to allow detection in most PCR protocols.[16,29,30] Owing to other clinical and experimental analyses[15] applied on whole-blood samples of the patient, the amount reserved for our study was 20 ng DNA. Thus, considering a DNA content of 10 pg/cell, one could expect 1–2 integrations per sample. Yet, despite the limited amount of patient DNA samples, we were able to detect 10 integrations in patient 3 and three integrations in patient 4. None of the 13 integrations was found via LM-PCR in the grafts.

Six of those occurred within EnsEMBL or RefSeq genes. To compare insertion site patterns found *in vitro* to those found *in vivo*, we matched the gene groups that were over- and underrepresented and found in the *in vivo* group three of the six hit genes encoding proteins assigned to 'signal transducer activity', which was among the most frequently targeted groups *in vitro* (Figure 3), whereas one was also assigned to the frequently hit 'receptor activity' gene group. As all *intragenetic* insertion sites were detected in a patient who underwent GCV infusion therapy and as this patient developed severe GvHD following administration of donor TLCs,[15] this might point at survival advantage of certain integration-bearing cells.

With respect to the low number of insertion sites, the preferred intragenetic appearance of the vector, the relatively high number of genes closer than ± 10 kb and the preferred insertion in signaling molecule encoding genes and near nucleic acid binding protein encoding genes, the *in vivo* detected pattern can be postulated as consistent with the pattern observed *in vitro*. This means that no specific selection process has been detected in these TLC following transplantation. Preclinical analysis of the integration repertoire with abundant amounts of transduced cells therefore allows a prediction also for the *in vivo* situation, where target cells are scarce.

Thus, our results are confirming and extending recently published data of a different study,[10] where similar integration patterns could be observed owing to same target cell populations, whereas the vector backbone and the transduction conditions were different. We did not have material available for functional studies of individual transduced cells. However, none of our patients developed a clonal disorder so far, suggesting that despite the potentially dangerous interactions with preferred hit and adjacent gene groups, retroviral integration has only little consequence on the safety of TLC transplantation.

Acknowledgements

We gratefully acknowledged the excellent technical assistance of Sigrid Heil, Bernhard Berkus and Hans-Jürgen Engel. We thank Axel Benner from the Central Unit Biostatistics of the German Cancer Research Center for statistical support.

This work was supported by Grant M 20.4 of the HW & J Hector Foundation and by Grant FR 1732/3-1 of the German Research Foundation (DFG).

References

1 Heslop HE, Ng CY, Li C, Smith CA, Loftin SK, Krance RA *et al*. Long-term restoration of immunity against Epstein–Barr virus infection by adoptive transfer of gene-modified virus-specific T lymphocytes. *Nat Med* 1996; **2**: 551–555.
2 Kolb HJ, Schattenberg A, Goldman JM, Hertenstein B, Jacobsen N, Arcese W *et al*. Graft-versus-leukemia effect of donor lymphocyte transfusions in marrow grafted patients. European Group for blood and marrow transplantation working party chronic leukemia. *Blood* 1995; **86**: 2041–2050.
3 Kernan NA, Bartsch G, Ash RC, Beatty PG, Champlin R, Filipovich A *et al*. Analysis of 462 transplantations from unrelated donors facilitated by the national marrow donor program. *N Engl J Med* 1993; **328**: 593–602.
4 Tiberghien P, Reynolds CW, Keller J, Spence S, Deschaseaux M, Certoux JM *et al*. Ganciclovir treatment of herpes simplex thymidine kinase-transduced primary T lymphocytes: an approach for specific *in vivo* donor T-cell depletion after bone marrow transplantation? *Blood* 1994; **84**: 1333–1341.
5 Bordignon C, Bonini C, Verzeletti S, Nobili N, Maggioni D, Traversari C *et al*. Transfer of the HSV-TK gene into donor peripheral blood lymphocytes for *in vivo* modulation of donor anti-tumor immunity after allogeneic bone marrow transplantation. *Hum Gene Ther* 1995; **6**: 813–819.
6 Tiberghien P, Cahn JY, Brion A, Deconinck E, Racadot E, Herve P *et al*. Use of donor T-lymphocytes expressing herpes-simplex thymidine kinase in allogeneic bone marrow transplantation: a phase I–II study. *Hum Gene Ther* 1997; **8**: 615–624.
7 Bonini C, Ferrari G, Verzeletti S, Servida P, Zappone E, Ruggieri L *et al*. HSV-TK gene transfer into donor lymphocytes for control of allogeneic graft-versus-leukemia. *Science* 1997; **276**: 1719–1724.
8 Li Z, Dullmann J, Schiedlmeier B, Schmidt M, Von Kalle C, Meyer J *et al*. Murine leukemia induced by retroviral gene marking. *Science* 2002; **296**: 497.
9 Hacein-Bey-Abina S, Von Kalle C, Schmidt M, Le Deist F, Wulffraat N, McIntyre E *et al*. A serious adverse event after successful gene therapy for X-linked severe combined immunodeficiency. *N Engl J Med* 2003; **348**: 255–256.
10 Recchia A, Bonini C, Magnani Z, Urbinati F, Sartori D, Muraro S *et al*. Retroviral vector integration deregulates gene expression but has no consequence on the biology and function of transplanted T cells. *Proc Natl Acad Sci USA* 2006; **103**: 1457–1462.
11 Giordano FA, Hotz-Wagenblatt A, Lauterborn D, Appelt J-U, Fellenberg K, Nagy KZ *et al*. New bioinformatic strategies to rapidly characterize retroviral integration sites (submitted).

12 Jonnakuty S, del Val C, Glatting KH, Suhai S. ExpressMiner: a tool for mining the gene expression data from microarray experiments. *12th International Conference on Intelligent Systems for Molecular Biology (ISMB 2004) and third European Conference on Computational Biology (ECCB 2004)*, Glasgow, Scotland, UK.
13 Frank O, Rudolph C, Heberlein C, von Neuhoff N, Schrock E, Schambach A *et al*. Tumor cells escape suicide gene therapy by genetic and epigenetic instability. *Blood* 2004; **104**: 3543–3549.
14 Kuhlcke K, Ayuk FA, Li Z, Lindemann C, Schilz A, Schade UM *et al*. Retroviral transduction of T lymphocytes for suicide gene therapy in allogeneic stem cell transplantation. *Bone Marrow Transplant* 2000; **25**: S96–S98.
15 Fehse B, Ayuk FA, Kroger N, Fang L, Kuhlcke K, Heinzelmann M *et al*. Evidence for increased risk of secondary graft failure after *in vivo* depletion of suicide gene-modified T lymphocytes transplanted in conjunction with $CD34^+$-enriched blood stem cells. *Blood* 2004; **104**: 3408–3409.
16 Laufs S, Gentner B, Nagy KZ, Jauch A, Benner A, Naundorf S *et al*. Retroviral vector integration occurs in preferred genomic targets of human bone marrow-repopulating cells. *Blood* 2003; **101**: 2191–2198.
17 Burge C, Karlin S. Prediction of complete gene structures in human genomic DNA. *J Mol Biol* 1997; **268**: 78–94.
18 Pruitt KD, Maglott DR. RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res* 2001; **29**: 137–140.
19 Schuler GD. Pieces of the puzzle: expressed sequence tags and the catalog of human genes. *J Mol Med* 1997; **75**: 694–698.
20 Bairoch A, Apweiler R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res* 2000; **28**: 45–48.
21 Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 2000; **28**: 27–30.
22 Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM *et al*. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 2000; **25**: 25–29.
23 Kuehlcke K, Naundorf S, Laufs S, Fauser AA, Fruehauf S. Optimising retroviral-mediated gene transfer into hematopoietic stem cells under clinically relevant conditions. *Blood* 2002; **100**: 484B–485B.
24 De Palma M, Montini E, Santoni de Sio FR, Benedicenti F, Gentile A, Medico E *et al*. Promoter trapping reveals significant differences in integration site selection between MLV and HIV vectors in primary hematopoietic cells. *Blood* 2004; **105**: 2307–2315.
25 Schroder AR, Shinn P, Chen H, Berry C, Ecker JR, Bushman F. HIV-1 integration in the human genome favors active genes and local hotspots. *Cell* 2002; **110**: 521–529.
26 Mitchell RS, Beitzel BF, Schroder AR, Shinn P, Chen H, Berry CC *et al*. Retroviral DNA integration: ASLV, HIV, and MLV show distinct target site preferences. *PLoS Biol* 2004; **2**: E234.
27 Wu XL, Li Y, Crise B, Burgess SM. Transcription start regions in the human genome are favored targets for MLV integration. *Science* 2003; **300**: 1749–1751.
28 Laufs S, Nagy KZ, Giordano F, Hotz-Wagenblatt A, Zeller WJ, Fruehauf S. Insertion of retroviral vectors in NOD/SCID repopulating human peripheral blood progenitor cells occurs preferentially in the vicinity of transcription start regions and in introns. *Mol Ther* 2004; **10**: 874–881.
29 Gentner B, Laufs S, Nagy KZ, Zeller WJ, Fruehauf S. Rapid detection of retroviral vector integration sites in colony-forming human peripheral blood progenitor cells using PCR with arbitrary primers. *Gene Therapy* 2003; **10**: 789–794.
30 Nagy KZ, Laufs S, Gentner B, Naundorf S, Kuehlcke K, Topaly J *et al*. Clonal analysis of individual marrow-repopulating cells after experimental peripheral blood progenitor cell transplantation. *Stem Cells* 2004; **22**: 570–579.