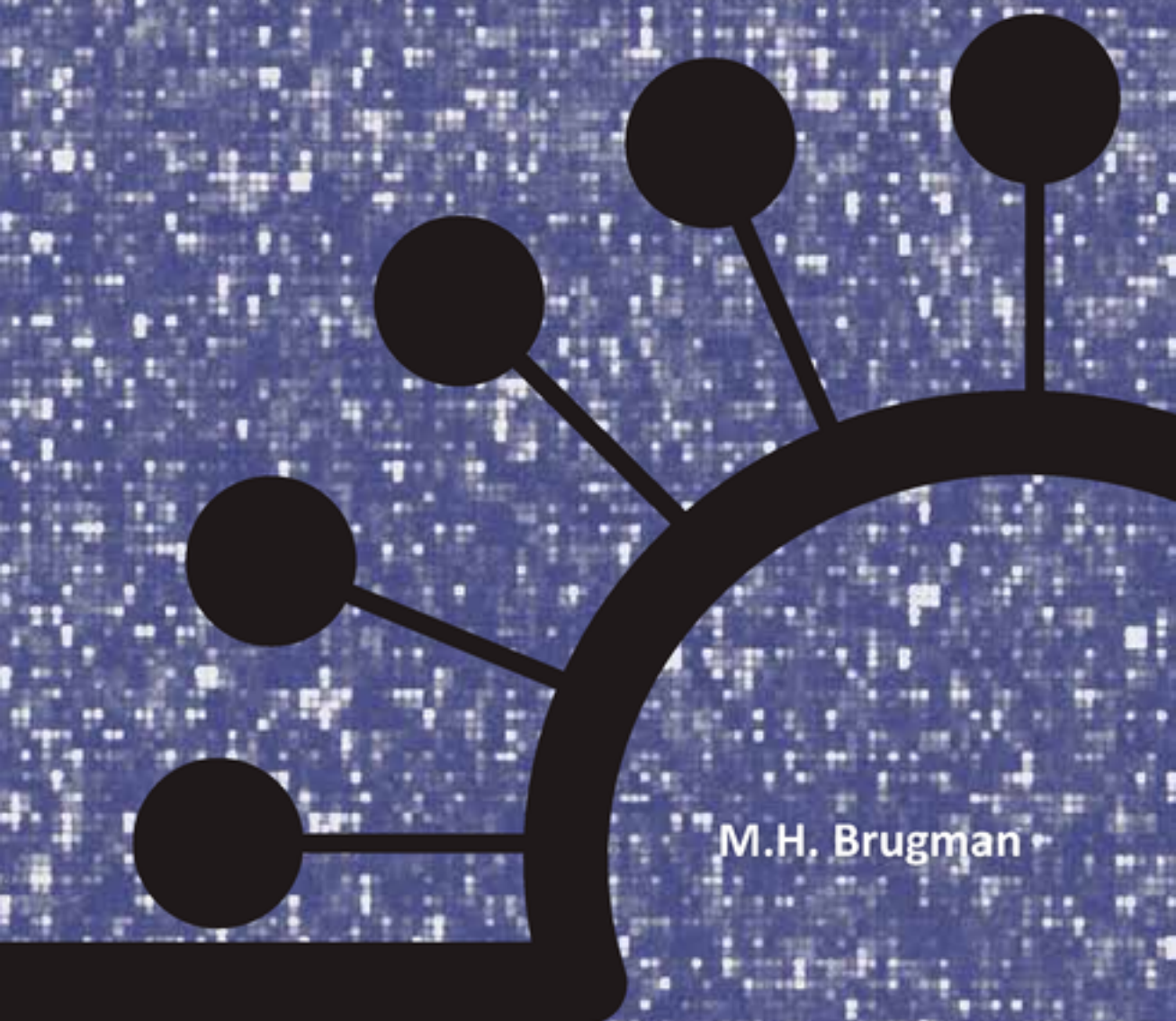
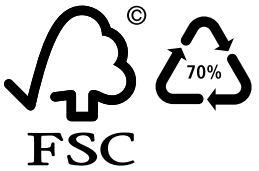
Insertional Oncogenesis after Retroviral Gene Transfer in Hematopoietic Stem Cells



M.H. Brugman

Insertional Oncogenesis after Retroviral Gene Transfer in Hematopoietic Stem Cells

Insertionele Oncogenese na Retrovirale Gen Transfer
in Hematopoietische Stamcellen



Mixed Sources
Productgroep uit goed beheerde bossen, gecontroleerde bronnen en gerecycled materiaal.

Cert no. CU-COC-803902
www.fsc.org
© 1996 Forest Stewardship Council

ISBN: 978-90-9025433-3

Copyright M.H. Brugman, Hannover 2010
No part of this publication may be reproduced or transmitted in any form or by any means without written permission of the copyright owner

Acknowledgements
Funding for this study was provided the Netherlands Health Research Organization ZonMw (www.zonmw.nl), Translational Gene Therapy Program, projects 43100016 and 43400010, by the European Commission's 5th and 6th Framework Programs (http://ec.europa.eu/research), Contracts QLK3-CT-2001-00427-INHERINET, LSHB-CT-2004-005242-CONSERT, LSHB-CT-2006-018933 and LSHB-CT-2006-19038, NIH R01 CA 112470-01 (http://www.nih.gov), by the Deutsche Forschungsgemeinschaft DFG (http://www.dfg.de), grant Ka976/5-3, SCHM 2134/1-1 and SFB738-C3, by the Bundesministerium für Bildung und Forschung BMBF (www.bmbf.de), grant 01GU0601 (TreatID) and 01GU0809 (iGene), by the German Federal Ministry of Education and Research, (BMBF project iGene 01GU0813).

Insertional Oncogenesis after Retroviral Gene Transfer in Hematopoietic Stem Cells

Insertionele Oncogenese na Retrovirale Gen Transfer
in Hematopoietische Stamcellen

Proefschrift

ter verkrijging van de graad van doctor aan de
Erasmus Universiteit Rotterdam
op gezag van de
rector magnificus

Prof.dr. H.G. Schmidt

en volgens besluit van het College voor Promoties.

De openbare verdediging zal plaatsvinden op
Donderdag 17 juni 2010 om 9:30 uur

door

Martinus Hidde Brugman
geboren te Dordrecht

ERASMUS UNIVERSITEIT ROTTERDAM

Promotiecommissie

Promotor:	**Prof.dr. G. Wagemaker**
Overige leden:	**Prof.dr. E.A. Dzierzak**
	Prof.dr. C.H. Baum
	Prof.dr. P.J. van der Spek
Copromotor:	**Dr.ing. M.M.A. Verstegen**

Dit proefschrift is tot stand gekomen met financiële steun van A.K. van der Hidde – Beekman en B.J. Brugman-Groen.

TABLE OF CONTENTS

CHAPTER

Introduction

INTRODUCTION

Abstract

This thesis is focuses on the insertional oncogenesis brought about by gamma-retroviral vector insertions. In hematopoietic gene therapy, gamma-retroviral vectors can be used to deliver therapeutic transgenes into target cells of patients with monogenic disorders, which has been successfully shown in three human diseases. The addition of the therapeutic gene to the host cell genome has the opportunity to cure the disorder. The mechanism that allows insertion of the transgene in the host cell genome can unfortunately also introduce deregulation of the genes surrounding the insertion site, sometimes with leukemia as a result. In the studies described here, the insertion profiles in mouse or human hematopoietic cells were analyzed and the frequency of oncogenic mutations was determined. In addition, software that allows automated determination and annotation of retroviral insertion sites was developed.

Hematopoietic stem cells and blood cell reconstitution

The hematopoietic stem cell (HSC) is a cell that can undergo self-renewal and, by differentiation, give rise to all different cells in the hematopoietic system, from erythrocytes and thrombocytes to the various different cells of the immune system. In transplantation studies, it was shown that transplanting even a low number of these HSC is sufficient to successfully reconstitute hematopoiesis (1). In the mouse, hematopoiesis starts at embryonic day 7.5 with primitive hematopoietic cells originating from the yolk sac. The earliest hematopoietic stem cells (HSC) are formed at embryonic day 10.5 in the aorta-gonad-mesonephros (AGM) region of the developing embryo (2) in a process that is regulated by Runx1, and Gata2, with involvement of the cytokines Bmp4, Hh, IL3 and IL1 (3). The origin of these cells was by *in vitro* experiments shown to be haemogenic endothelium (4). The exact location of these endothelial cells was later shown to be aortic endothelium, using whole embryo microscopy (5). These adult hematopoietic cells are a source of spleen colony forming cells (CFU-S) and are fully capable of reconstituting adult hematopoiesis in mice (6,7). This potential was also shown for cells derived from the umbilical cord blood (UCB) (8), G-CSF mobilized peripheral blood (CPB) (9,10) and transplantation of total adult bone marrow (ABM) in animals (11) and humans (12).

Considerable effort has gone towards identifying which subpopulation of these sources contains most if not all of the stem cells, mainly by studying the expression of surface antigens (13,14). This research resulted in the identification of the surface antigen *CD34* (15), (16) which is a single-chain trans membrane glycoprotein of 105 to 120 kilodalton (17), present on nearly all colony forming units (CFU-GM, BFU-E, CFU-Mix) and marks nearly all hematopoietic colony-forming progenitors (18). From a variety of studies the hematopoietic stem cell is known to be enriched in human $CD34^+$ cells in

the bone marrow (15) and a higher frequency is present in the $CD34^{+}CD38^{-}$ cells (19,20). When $CD34^{+}CD38^{-}$ HSC are isolated and subsequently cultured, the *CD38* marker loses its relevance as a marker for immature hematopoietic stem cells, because the ability of human cord blood cells to repopulate SCID mice dissociates from the *CD38* marker (21). Expression of *CD133* , also a marker for hematopoietic cells (22), is retained and is a marker for immature HSC, that can give rise to hematopoiesis in a NOD/SCID mouse even when these cells have been cultured *in vitro* (23). Even the use of *CD34* as a maker for hematopoietic stem cells is debated, since it was shown that in the mouse, the $CD34^{-}$LSK pool is able to reconstitute mouse hematopoiesis. In the mouse, the LSK (lineage negative, Sca-1 positive, c-Kit positive) cells contain a large fraction of cell with repopulation potential (24). The Lnk protein, which regulates *TPO* signal transduction, seems to be important, since in absence of this protein (in *Lnk*-- cells) repopulation is increased when cells are stimulated with TPO or TPO/SCF (25).

In the mouse, the *CD34* antigen expression on hematopoietic cells is conserved (26). The lineage negative, Sca-1 negative, Thy1.1 low cells that constitute approximately 0.05% of mouse bone marrow, have been shown to be highly enriched in HSC (24) and capable of reconstituting hematopoiesis after transplantation (27,28). When this cell population was further subdivided, the steel factor (SCF) receptor c-Kit was shown to be present on 70-80% of these cells (29). Another marker used to identify HSC is Rhodamine123 (30). Rhodamine123 is an indicator of active mitochondria and therefore, a dim Rhodamine123 stain is associated with the quiescent cells which provides stable, long-term hematopoiesis after transplantation (31). In mice, it was also shown, that when subdividing the LSK cells into populations which either do or do not express Flt3, the Flt3 negative population was able to reconstitute the multipotent long-term hematopoiesis after transplantation. The LSK Flt3+ pool only gave rise to lymphoid lineage restricted reconstitution. Furthermore, stimulation of HSC with Flt3 failed to support survival, whereas c-Kit stimulation does (32). One could therefore argue that the addition of Flt3L to medium used for culturing HSC is at best unnecessary. Several combinations of cytokines were tested by Wognum (33), showing that IL3 and IL11, both commonly used in HSC cultures, reduce the transduction of short- and long term repopulating cells. As the most optimal cytokine cocktail a combination of SCF, FLT3L and TPO was described.

Stem cell signatures and markers

Since hematopoietic stem cells can be used for bone marrow transplantation (BMT), it is important to know what cells make up the true hematopoietic stem cell. Research trying to identify this population of 'true' stem cells has focused on stem cells signature, investigating gene expression in the stem cell, and on stem cell markers, trying to identify cell surface markers that would allow the identification of the stem cells by their phenotype.

Several studies have been directed at the evaluation of gene expression in stem cells derived from different tissues, for example, subtractive analysis of different sources of stem cells, e.g. embryonic stem cells, neuronal stem cells, retinal stem cells and hematopoietic stem cells lead to the formation of gene sets that were hypothesized to contribute to a stem cells signature, or a signature of specific cell lineage (34,35). A later comparison of the data in these papers and analysis with additional data showed that the neuronal and hematopoietic signatures could be nicely reproduced, the stem cell signature was reduced to only one gene that was present in all three datasets. The neuronal and embryonic stem cell signatures showed much higher (236 and 332 respectively) numbers of overlapping genes, which lead to a highly significant prediction of a gene expression profile for these immature, but more differentiated cells (36). One could therefore conclude that the more lineage specific (such as the neuronal and hematopoietics) stem cells do indeed share a signature, while such a signature is harder to determine or even absent when comparing stem cells obtained from different tissues.

Further studies on Lineage$^-$ Sca-1$^+$ cKit$^+$ Thy1.1^{lo} cells and more differentiated cell types in mice (37) and in human CD34$^+$CD38$^-$ cells from mobilized peripheral blood, bone marrow and umbilical cord blood (38) showed that there is a cell signature that is related to the HSC phenotype, which contains *Evi1/Mds*, *Rbpms* and *Cebpb*. Furthermore, both studies observed that the HSC enriched cell population was characterized by notably more expressed genes compared to the more differentiated cell populations. In a similar experimental set-up, but using microarrays that interrogated an increased number of transcript, Chambers (39) reassessed the mouse LSK and more differentiated cell types. This analysis confirmed the earlier findings and found that the genes associated with circadian rhythm and Wnt signal transduction KEGG pathways were expressed in LSK cells. In addition, a transcriptional similarity between T-cell and LSK cells was noted, which lead the authors to make a comparison between the repopulation potential of HSC and the need for T-cells to rapidly proliferate when an immune response is needed. After these signature studies had been performed, a need existed to prove hematopoietic involvement of the identified genes. Two studies showed innovative methods to study the identified genes. Eckfeldt *et al.* identified 41 differentially expressed HSC transcripts and confirmed hematopoietic involvement 16 of the homologs of the identified genes in morpholino knockdown studies in zebrafish (40). Using gene expression data, Deneault *et al.* selected a set of nuclear factors that were subsequently expressed in mouse CD150+CD48-Lin- cells using retroviral vectors. These transduced cells were subsequently tested for their ability to increase engraftment and repopulation of the hematopoietic system. From the 104 genes that were originally selected, 10 showed an increase in repopulation similar or superior to *HoxB4* expressing cells (41). Among these genes was *Prdm16*, which was also found as a retroviral insertion site in the XCGD trial (42).

The precise immunophenotype, as well as the mRNA expression pattern of hematopoietic stem cells has been a focus point of research for considerable time and new immunophenotypical markers for HSC are still being identified (43). $CD34^+$ sorted cells are used as a source for HSC cells in bone marrow transplantations, although regular BMTs are performed using total (unfractionated) bone marrow. HSC transplantations are used in leukemia treatment and for the correction of hematopoietic disorders, although protocols may vary between centers. $CD34^+$ sorting is regularly used in patients with a $CD34^-$ tumor, to purge the leukemic cells from the transplantation, after which the transplant gets reinfused into the patient. Other uses for CD34 selection are depletion in an allogeneous transplantation setting to prevent graft versus host disease, or in autologous transplantation, removal of T-cells that cause autoimmunity to restore normal, non- self- reactive T-cell repertoire. In allogeneous transplantations, most frequently HPCA (hematopoietic progenitor stem cells collected by apheresis) are used without further purification. The same source is also used frequently for allogeneous transplantation, which leads to subsequent immuno-suppression due to the large number of T-cells in the graft.

Hematopoietic stem cells as a target for gene transfer

In hematopoietic gene therapy, providing a functional copy of a gene that is defective in the genome in the HSC, is performed *in vitro*. This means that the target cells need to be cultured *in vitro*, for a period ranging from 1 to 4 days. The culturing of HSCs is associated with a negative effect on the regenerative capacity of these cells in the bone marrow (44). Especially when using retroviral vectors for gene transfer, the target cells need to be stimulated with growth factors to undergo cell division which allows integration of the retrovirus. The use of Flt3L, SCF and TPO for this purpose has been well described, but different combinations of growth factors are also used for HSC stimulation *in vitro* before the transduced cells are transplanted into patients in clinical gene therapy trials (45-47).

The hematopoietic stem cell is a suitable target for gene therapy of diseases of the hematopoietic system. For patients suffering from monogenic disorders such as XSCID (X-linked severe combined immunodeficiency) or ADA-SCID (Adenosine deaminase deficiency), the treatment of choice is bone marrow transplantation. Transplantation with bone marrow obtained from HLA-identical donors is often quite successful (3 year survival 77%) (48), but suitable donors are not always available. Bone marrow transplantation (BMT) with HLA- mismatched donors has met with limited success (3 year survival 54%). Umbilical cord blood is an alternative source of HSC with which successful HSC transplantations have been performed (49). Although positive results have been observed, the T-cell function in SCID patients that have been transplanted with $CD34^+$ cells, after an initial normalization, was shown to decrease over time (50). Furthermore,

conditioning patients for BMT is usually more invasive than the minor conditioning or lack of conditioning required for the transplantation of gene modified stem cells. Retroviral gene therapy is a potent method for long-term mono-allelic introduction of therapeutic genes in patients with monogenic diseases that lead to impaired function in the cells derived from the HSC. And so, in cases were matched related donors are not available, gene therapy on autologous HSC could be a valuable treatment option.

Retroviral gene transfer is normally performed on patient bone marrow cells *ex vivo* (figure 1). After the transduction procedure, the treated cells are reinfused in the patient, which in some cases previously underwent a pretreatment. This pretreatment consists, most often of myoablation by either chemotherapy (e.g. busulfan treatment) or radiotherapy (total body irradiation) and allow hematopoietic reconstitution that

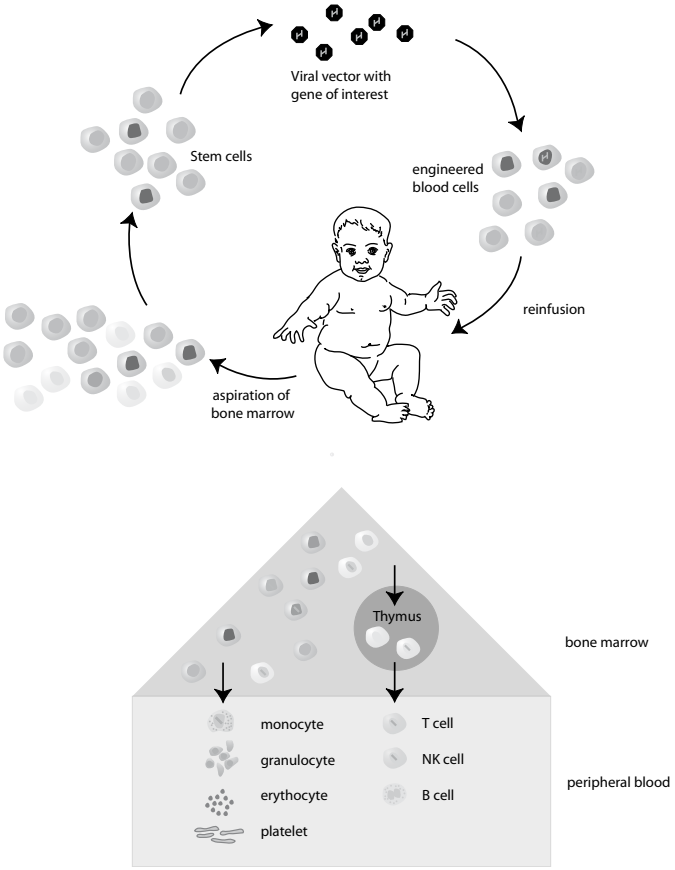


Figure 1: Clinical Gene therapy treatment. $CD34^+$ cells, a cell population enriched in HSC are collected from aspirated bone marrow of young patients. The HSC are transduced with a gamma-retroviral vector carrying the gene of interest (in green) after which the engineered HSC are reinfused into the patient. After the engineered cells engrafted and started producing blood cells, the entire repertoire of hematopoietic cells, now arising from the engineered HSC, carry the corrected gene (in green).

originates from the transplanted cells. This procedure has been successfully demonstrated in patients suffering from X-linked SCID (IL-2 receptor, or common, gamma chain deficiency) (45,46), ADA-SCID (adenosine deaminase deficiency) (47) and chronic granulomatous disease (*GP91PHOX* deficiency, causing a NADPH oxidase defect) (42). Currently, a clinical trial for Wiskott-Aldrich Syndrome is in preparation, with several diseases in pre-clinical stages of preparation (*Btk*, *Rag1* and *Rag2* deficiencies).

Other methods of introduction of therapeutic genes are also available. Several of them provide transient expression, for example using adenoviral vectors or non-viral such as gene delivery by electroporation or lipid mediated plasmid transfection. By using these methods stable integration is achieved only in a very limited amount of cells. Stable insertion is, however, vital for the success of the therapy in dividing cells such as the cells of the hematopoietic system, since these cells require the presence of the therapeutic gene in the entire progeny of the bone marrow cell. In other tissues, where there is less proliferation, non-integrating methods might well cure the disease. A therapeutic system without vector insertion would reduce the chance of gene deregulation by the therapeutic vector, and might therefore be a safer option for those diseases or tissues that do not necessarily need insertion of the vector.

Adeno-associated virus and homologous recombination using zinc finger nucleases (51) are both alternatives to retroviral gene transfer which could also lead to stable integration of the transgene in the host cell and in the latter case provide an option to repair a defective gene, rather than supply an additional copy of the affected gene. When a defective gene is 'repaired' in place, there is no need for insertion of a therapeutic vector, with its promoter and therapeutic gene sequences, thereby theoretically reducing the genotoxicity of such a procedure (Table 1).

The history of gene therapy

The possibility of gene therapy has been debated even before recombinant DNA techniques were available. For example, when the bacteriophage 'filterable agent' was shown to be able to transfer nutritional requirement and drug resistance to Salmonella species (52), the authors coined the term transduction.

> *'The mechanism genetic exchange found in these experiments differs from sexual recombination in Escherichia coli (E. coli) in many aspects so as to warrant a new descriptive term, transduction.'*

Based on experiments with Rous sarcoma virus, which showed that information for the formation of virus progeny as well as morphologic changes could be transfered to target cell, the possible existence of the provirus was suspected (53). This introduced a mechanism working in a manner different from the 'central dogma of molecular biology', namely, that information can be transmitted from RNA to DNA to the genome. The existence of this mechanism was especially clear in the case of RSV, since this

Table 1: Comparison of a hypothetical ideal gene therapy vector and the currently available vectors.

Requirements for an ideal gene therapy vector	Current status
Efficient, non-cytopathic packaging system with low likelyhood of recombination	· Split packaging systems for RV and LV vectors are available
Serum-resistant particles pseudotyped with specific envelopes	· The currently available envelopes include RD114, GALV, VSV-G, ecotropic and amphotropic vectors. Using enveloppe proteins linked to hematopoietic growth factor molecules, it is even possible use the hematopoietic growth factor receptors.
Reverse transcription prior to cell entry	· Not possible for an integrating vector. · Possible for protein/mRNA transfer using viral vectors.
Unrestricted cytoplasmic transport and nuclear import	· Both RV and LV vectors have known restriction factors (Galla, Host restriction of HIV-1 by APOBEC3 and viral evasion through Vif. Niewiadomska AM, Yu XF.Curr Top Microbiol Immunol. 2009;339:1-25. Review.) · Lentivirus can enter the nucleus of non-dividing cells.
Insulated expression cassette and/or specific chromosomal targetting	· Globin vectors currently being tested in clinical trials have chicken HS4 insulator sequences.
Physiological or regulated levels of transgene expression	· In SIN designs, cellular and doxycycline regulated promoters are available. Globin vectors in clinical trial use Globin promoter sequences to regulate expression.
Efficiency in relevant model	· Current RV and LV vectors are efficient in HSC and various other cell types.
Avoidance of horizontal or vertical transmission	· No examples of transmission are currently known.

Adapted from Baum, Mol Ther 2006

virus was already known to have an RNA genome. This and several other experiments (with polyoma and simian virus 40 SV40) brought about the notion that viruses can be regarded as packages that deliver new functions or traits to cells and that these could be inherited by daughter cells. The attributes that make viruses such useful tools for gene therapy were already known before recombinant DNA technology was commonplace.

With the introduction of recombinant DNA techniques in the 1970's, very specific tools allowed the construction of well-defined gene therapy vectors. The first attempts to understand the base pair sequence of the genome were carried out using restriction endonucleases, the first of which had been described by Smith and Wilcox (54). These techniques allowed the creation of restriction maps of a specific locus (55) and a general knowledge about the DNA sequence. Complete elucidation of the DNA sequence was however not possible until 1973, when Gilbert and Maxam (56) described the lac operator DNA sequence determined by a technique based on ^{32}P labeling of DNA fragments and subsequent base-specific chemical decomposition. The resulting cleaved and labeled

material was then separated on a polyacrylamide gel, which in the end allowed identification of the DNA sequence. The method described by Sanger (57) employed DNA amplification by a DNA polymerase and used radio labeled (didesoxyribonucleosidtriphosphate) nucleotides which cannot form the DNA backbone, the phosphodiester bond, thereby terminating the DNA polymerization. Separating these different sized fragments on a gel allowed the identification of the DNA sequence by analyzing the autoradiography of the radioactively labelled fragments.

Generation of the first genetically modified host cells

Retroviruses have a mechanism for insertion of their genome into a host cell that can opportunistically be used to gene delivery.

The initial vectors used dissect the mechanism of retroviral gene delivery were replication deficient SV40 variants that could be generated using the helper dependent complementation principle. These vectors often carried lambda DNA fragments or (58) or similar model DNAs and therefore where 'model viruses' rather than actual therapeutic vectors.

The first vector that actually carried a 'therapeutic gene' was an SV40 plasmid with rabbit beta-globin full length cDNA (59). In 1973, Graham and van der Eb introduced an optimized calcium phosphate transfection technique (60), after which the now reproducible calcium phosphate transfection method became widely used. Using calcium phosphate transfection, Cline introduced dihydrofolate reductase (*DHFR*) into cultured mouse bone marrow cells and could successfully transplant these cells and prove their resistance to methotrexate (61). Based on these preclinical results and the results obtained in similar experiments, where Herpes Simplex Virus thymidine kinase (HSV-TK) was expressed (62), Cline and colleagues attempted the cure of beta-thalassemia by means of *ex vivo* calcium phosphate transfection of human bone marrow with beta-globin and subsequent transplantation. Cline and his colleagues were chided for not adhering to the regulatory procedures necessary for the approval of human clinical trials at UCLA (63) and the National Institutes of Health (USA) (64,65) after which the experiments were halted.

With the current knowledge about rarity of HSC in total bone marrow and the relatively low proportions of cells that integrate the plasmid DNA stably into the genome, it seems unlikely that the Cline experiments would have been able to transduce enough hematopoietic stem cells to allow long-term gene correction in the hematopoietic system. For gene therapy in HSC to be successful, the therapeutic gene needs to be expressed continuously and at the appropriate level. Stable insertion into the host genome allows the therapeutic gene to be expressed in all cells arising from the HSC.

The problem of stable expression of the transgene by integration into the host genome was solved by the development of retroviral vectors in the 1980s. These RNA

viruses were engineered to express only part of the viral genome, and retain the features that make them attractive tools for gene transfer: relative efficient cell entry, reverse transcription of the viral RNA genome into DNA and stable integration of this DNA into the host cell genome. The first retroviral vectors were synthesized using replication competent wild type retrovirus, that supplied the engineered virus with the necessary proteins for cell entry, reverse transcription and integration (66,67). The risk for generating replication competent retrovirus was reduced soon thereafter, when the need for helper virus was overcome by using packaging lines that supplied the required proteins *in trans* (68,69) The first experiments with these viruses attempted to deliver the human hypoxanthine-guanine phosphoribosyltransferase (*HPRT*) cDNA to HPRT defective cell lines (70). In murine hematopoietic progenitors (71) and pluripotent cells (72), the first marking experiments were carried out using the selectable Neo marker, which confers resistance to the neomycin analogue G418 and allows the transduced cells to be positively selected. This approach allowed the tracking of the transduced colonies by culturing them in the presence of G418.

The use of retroviral vectors provided an additional tool to track transduced cells, since the semi-random integration of retrovirus provides unique virus-genome boundaries. When the number of insertions per cell is low the unique boundary sequences can therefore be considered to each mark an individual clone. Several techniques were developed to identify these boundaries, using restriction digest of the genomic DNA, ligation and PCR amplification of the ligated DNA circles and using primer pairs located in the virus DNA and in a known sequence that is attached in a DNA linker at the restriction site. Various polymerase chain reaction (PCR)-based methods have been developed for this purpose, including inverse PCR (73,74), vectorette- and splinkerette-PCR [(75-77), ligation-mediated (LM) PCR (78-84). Later, more sensitive methods such as Linear amplification mediated (LAM)-PCR (85) were developed, which became widely used. With the availability of efficient retroviral systems for introduction of therapeutic genes into cells of patients with monogenic diseases, work was started on more and more preclinical models for specific human diseases (86). Attempts were made to express the genes for beta-globin (87) for hemoglobinopathies, adenosine deaminase (*ADA*) for ADA combined immunodeficiency (88), glucocerebrosidase for Gaucher disease (89), and beta-glucuronidase for mucopolysaccharidosis type VII (Sly syndrome) (90). The latter showed for the first time that the phenotype of an inherited disease in a mouse model could actually be reversed using retroviral gene therapy.

Marker experiments in carried out in larger animals, such as dogs and non-humane primates showed that gene expression from a retrovirus in the hematopoietic stem cells in large animals could be more difficult than expressing genes in the mouse HSC (91,92). Difficulties with expressing genes could be connected to the fact that gammaretrovirus only integrates in cells that underwent mitosis, and that the induction of

mitosis reduces the repopulation capacity of the hematopoietic stem cells (93,94). This notion led to a diversity of experiments that were aimed at increasing the repopulation of transduced stem cells, either by optimizing the growth factors that are used during the *in vitro* gene transfer, or by trying new vector systems such as adeno-associated viruses (95), HIV derived lentiviruses (96) or foamy virus (97). These developments led to protocols that were used in the first clinical trials.

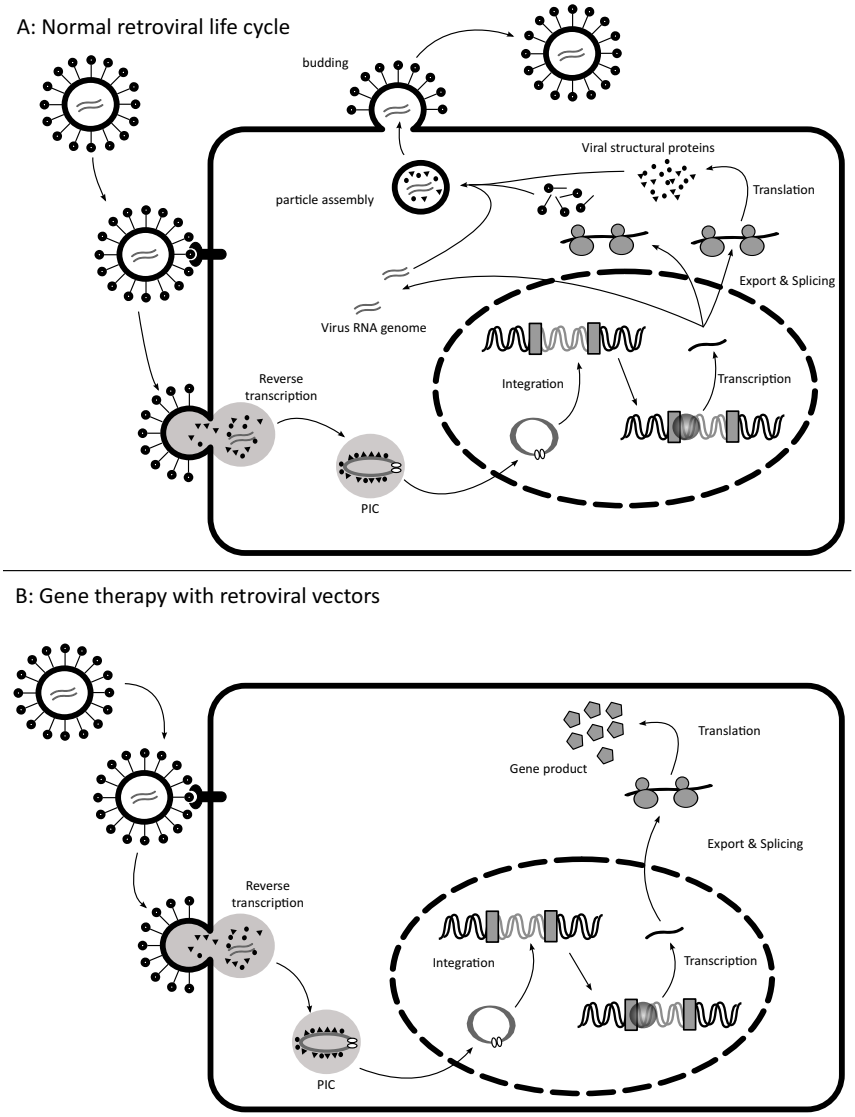
The first example of successful reconstitution of patient bone marrow and clinical improvement was observed in X-SCID patients with IL2R-gamma deficiency (45). These patients received a transplantation with *ex vivo* transduced HSC in which the therapeutic IL2R-gamma gene was incorporated. Later, similar success was shown in a clinical gene therapy trial for ADA-deficiency. ADA deficiency is a gene defect in adenosine deaminase that leads to lymphopenia due to the accumulation of the of adenosine and adenine deoxyribonucleotides (dAXP) in plasma and tissues, which eventually leads to immunodeficiency because the immature lymphoid cells are specifically sensitive to the toxic effects of the metabolites (98).

The retroviral life cycle

The first step in the normal gamma-retroviral life cycle is the entry into the host cell. The host cell is recognized by the virus or viral vector by the envelope glycoproteins present on the virus. Through a series of conformational changes, the lipid bilayers of the virus fuse and the viral content ends up inside the host cell. In a second step, the virus capsid is degraded (uncoated) after which the RNA genome is reverse transcribed into DNA by the viral reverse transcriptase (99). The viral DNA can then be integrated in the host cell genome by the viral integrase. Although the integration site selection is considered semi-random, different retroviruses have different insertion profiles (100) (Figure 2).

In the retrovirus life cycle, the integrated viral genome can now be transcribed and translated using the host cell machinery, giving rise to the viral capsid, polymerase, integrase, reverse transcriptase and envelope proteins. In addition, the viral genome RNA transcribed. This genome contains the packaging signal ψ, which causes the viral proteins to aggregate and form new virus particles. These particles then bud off from the host cell and form the new virus particles. (Adapted from (101)).

In gamma-retroviral gene therapy, once integration has happened, the vector is in stably in place and the therapeutic gene can be transcribed from the vector. The first generations of gamma-retroviral vectors drove transcription from the virus LTR, which was prone to silencing (102) and because the viral LTR is present at both ends of the vector backbone, the LTR allowed transcription of surrounding genes. An improvement in retroviral vectors came with the self-inactivating (SIN) architecture. Here, a deletion in the 3' LTR is introduced, which inactivates the LTR promoter. Since the U3 region of



Adapted from Baum *et al.* Molecular Therapy 2006 and Cepko and Pear, Current Protocols 2001.

Figure 2: The retroviral life cycle. (A) shows the normal retroviral life cycle. A virus attaches to a receptor recognized by the Env proteins. After fusion, the RNA genome is reverse transcribed into DNA, which then forms the pre-integration complex (PIC). After translocation to the nucleus and integration of the PIC into the host cell genome by the viral integrase, the virus genome is transcribed and partially spliced. The spliced RNA translated into the viral proteins that are needed for the generation of new viral particles. These particles are assembled from the RNA genome and the viral proteins translated by the host cell. The newly formed particles bud off from the host cell and can infect other cells. (B) In gene therapy, the retrovirus mechanism is exploited to deliver the vector into the host cell DNA. Since the vector does not contain any viral sequences, no viral proteins or genome are formed and therefore no infectious virus is generated. After integration of the vector, the host cell will express the transgene that was present in the vector.

the 3' LTR is copied to the 5' LTR during reverse transcription (Figure 3), the integrated vector cannot drive expression from the LTR. Instead, an internal promoter of choice can be added to drive expression of the transgene (103).

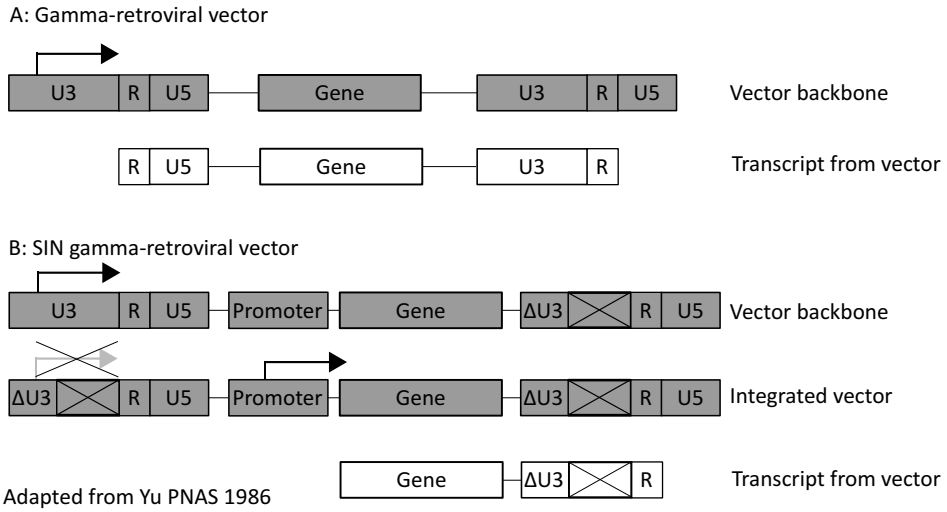


Figure 3: Architecture differences between gamma-retroviral vectors using normal and SIN design. (A) Regular gamma-retroviral vectors express the transgene from the LTR U_3 element. In white, the transcript that is formed is shown. (B) Self-inactivating vectors have a deletion in the U_3 in located in the 3' end of the vector. After integration, the ΔU_3 region from the 3' LTR is copied over to the 5' LTR effectively removing promoter and enhancer activity from the backbone. The vector can now utilize its internal promoter, which can be tissue specific or physiologic, to drive expression of the transgene (in white).

Leukemogenesis due to mutations introduced by viral vectors

Extensive research into the mechanisms of human oncogenesis raised the notion that oncogenesis is a process that requires multiple steps. Tumor cells need to be self sufficient in growth signals, be insensitive to anti-growth signals, evade apoptosis, have limitless replicative capacity, be able to induce angiogenesis and be able to invade tissues and metastasize (104). For leukemias not all these steps are of critical necessity, specifically the requirements for angiogenesis and metastasis do not need to be met. Furthermore, comparison of mouse and human studies made it clear that the six independent steps needed for oncogenic transformation in humans are not all required in the mouse model. Depending on the mouse model, in some cases only two lesions are sufficient for oncogenic transformation. Whether this is caused by the higher metabolic rate in mice, a more advance DNA damage control mechanism in humans or different metabolism of toxic compounds remains to be elucidated (105). Several studies in cell lines and patients showed that leukemogenesis indeed requires more than one lesion (106-108). In cells of patients treated with retroviral gene therapy one or several of these steps could be brought about by the transactivational properties of the retroviral promoter. In normal

retrovirus constructs, the expression of the viral genes or, in gene transfer vectors, the transgene, is driven by the long terminal repeats (LTR). The LTR region consists of a U3, R and U5 region. Transcription starts at the 5' U3, where sequences for initiation of transcription, basal promoter elements and upstream enhancers are situated. The 3' U5 region facilitates RNA cleavage and polyadenylation. Since LTR regions are present at both ends of the integrated virus, the promoter and enhancer regions in the 3' U3 of the virus can also drive transcription of genes surrounding the virus integration site (VIS) (109), which was shown to occur in the *dLNGFR* transduced mouse (110) and the common gamma chain transduced patients that presented with lymphoproliferative disease (111).

Microarray analysis and its application in retroviral gene therapy safety research

Together with the development of more efficient ways of transducing cells, toxic side effects of transduction of cells with viral vectors were observed (110-112). The method of action was determined to be the interference of the promoter sequences present in the viral LTR with the surrounding genomic locus. For analysis of the effect of an integrated virus on the surrounding locus, RT-qPCR can be used, but analysis large integration datasets is facilitated by the availability of DNA microarrays. Microarray technology, which allows the simultaneous measurement of mRNA expression of all genes in an organism in one sample, was originally described for Arabidopsis thalia by Schena (113) and subsequently for human Jurkat cells (114).

Researchers soon realized the power of this tool and new research strategies employing microarray techniques surfaced. The first papers using microarray techniques, clearly showed that the wealth of information that could be gathered using this technique needed more elaborate methods for analysis of the gene expression data. The initial experiments used DNA microarrays, where plasmids were spotted on glass slides. This design required the hybridization of a sample and a RNA reference, which usually were marked with difference fluorescent dyes, because the affinity of the spotted plasmids varied between the different targets RNA. Affymetrix introduced a lithographic method of generating small 25 bp probe sequences, which reduced the variability in probe affinity, thereby allowing 'single color' experiments, where just the RNA of the sample can be hybridized, without the need for a reference.

Affymetrix provides Microarray Analysis System (MAS) as a part of their GeneChip Operating System. (Affymetrix. Microarray Suite User Guide. Affymetrix. 2001,Version 5[1]). The microarray design as supplied by Affymetrix is considered technically advanced and reproducible, but it has been shown that the perfect match/ mismatch probe setup, that should allow measurement of the expression of the gene and the probe specific background, actually allows measuring of gene expression with the mismatch probes,

1 http://www.affymetrix.com/support/technical/manuals.affx

albeit at lower levels (115). The background correction and normalization method proposed by (116) Robust Analysis of Microarrays (RMA) which uses only the perfect match probesets to calculate gene expression, has gained popularity, as has GCMRA, a modified model, that does take the probeset composition into account. Which of these models is most suitable for background correction is still a matter of debate (117), although RMA seems to be better suited for high expression values, while GCRMA performs better for low expression values (116).

Once properly normalized, the gene expression for each probeset needs to be tested for significant differential expression. The most uncomplicated approach would be to perform t-tests for each probeset for the groups that should be compared. Performing these tests on the tens of thousands (45101 for Affymetrix Mouse 430 2 array) probesets present on one array immediately introduces the need for to correct the resulting p-values for the fact that we perform so many tests. Several solutions have been proposed, such as the classical Bonferroni correction (divide alpha by the number of tests (m) performed), the method due to Holm (divide alpha by the number of test performed (m) minus the rank number + 1) and more sophisticated methods, such as FDR control and FWER (118). Since each probeset in an Affymetrix array is composed of 11 [2] probes, additional information is present in the array that is summarized when normalization is applied. The data present in each probe can used to generate an model for the expression of a probeset and by ANOVA, statical differences can be identified (119). This method is particularly attractive, since it allows reaching statements about significant expression while using less than the conventional three arrays for each condition, an experimental setup that is less costly and allows more information to be obtained in situations, where the population of cells is difficult to obtain (120). Once the gene expression data are suitably corrected for background, normalized and tested for significance, conclusions should be drawn from the difference in expression values for different genes. Classically, conclusions would only be drawn based on the extend of difference that was observed (113), but it was felt that this approach was lacking, since it is not really expected that subtle changes in gene expression should be ignored. It is however necessary to be certain about the accuracy of the measurement, so expression differences should be tested for significance. The significantly differentially expressed genes can be related to the biological processes they take part in, the molecular functions they perform in a cell or in which part of the cell they reside. These relations are deposited in the Gene Ontology database[3] and several tools for testing over representation in the subclasses of these relations have been developed (e.g. DAVID, EASE (121), FatiGO (122,123), the R topGO package (124), GODist(125), ermineJ (126)). Again, the statistical tests employed by each

2 http://www.affymetrix.com/support/tech, Human Genome U133 array and Mouse Genome 430 array)

3 http://geneontology.org

of these tools are different and range from Fisher Exact tests without (DAVID, EASE) or with multiple testing correction (FatiGO) to more elaborate methods, such as permutation testing (topGO, GODist, ermineJ). The outcome of these tools is usually a list of under- or overrepresented genes based on a background gene set. Such functions can be difficult to interpret, since an over representation of a certain function class might not have any meaning in a phenotypical sense. In addition it is possible to use annotation tools (e.g. Ingenuity Pathway Analysis ((Mountain View, CA), PantherDB (127)) that query gene relation annotation, with the added opportunity to predict whether a set of genes is involved in a specific biological process or a disease.

Gene expression arrays as designed be Affymetrix and others are useful for measurement of gene expression in a relatively small form factor, allowing gene expression analysis on one genechip. The recent developments in gene expression measurement platforms included the exon array and tiling array. While on a gene expression array mRNA transcripts are interrogated by 1 to 8 (for the Affymetrix Mouse 430 2.0 Array) probesets, Exon Arrays use one probe per known or predicted exon (using predictions from GENSCAN, Ensembl and Vega[4]) This provides a higher level of detail and allows the accurate identification of expressed splice variants (128). Tiling arrays, which cover the entire genome with a resolution of one probe per approximately 35 base pairs even allows identification of transcription of parts of the genome that are not known or predicted to be expressed. The major application of these tiling array is, however, in methods such as chromatine immunoprecipitation, where complex DNA samples can be hybridized and their relative enrichment over control samples can be measured, allowing elaborate binding studies.

MLV oncogenesis studies

Murine Leukemia Virus (MLV) is a frequently utilized tool to introduce gene deregulation in mechanistic studies for cancer. Several studies used the mutagenic properties of wild type MLV to generate leukemia in either wild type mice (129-133) or in mouse models with predisposition for cancer development. The mice used in these studies carried either knockout mutations with possible oncogenic effect, such as Cdkn2a$^-$ for myeloid leukemia (134), CD3e$^-$ for T-cell leukemia (135) and p27Kip1$^-$ for lymphomas (136). Other mouse models introduced transgenes that were shown to be caused by translocations in human leukemias, such as Cbfb-MYH11 found in AML with inv(16)(p13q22) (137) and NUP98_HoxA9 found in CML with t(7;11)(p15;p15) (138) to obtain sensitivity for a certain type of leukemia and treated these mice with wild type virus to identify mechanism for the development of these diseases. In several cases, these approaches led to new insight in the mechanism of AML and lymphoid leukemia formation. The use of these models

4 http://www.affymetrix.com/products_services/arrays/specific/exon_arrays.affx

for safety evaluation is faced with difficulties, because the models were developed using replication competent wild type retrovirus, which will result in high numbers of proviral integrations per cell, while *ex vivo* transduction protocols for BM transduction typically aim to obtain very low numbers of integrations per cell. While these models generated very useful data on the integration properties of MLV provirus, the data obtained might be difficult to interpret for safety studies as the number of integrations introduced by wild type virus is much higher than by the virus dose used in *ex vivo* gene therapy. The studies that used wild type virus also failed to show that the candidate genes for leukemia formation are directly responsible for this effect. Most often the leukemias that were retrieved from the mice consisted of multiple clones or clones with multiple integrations. To prove the oncogenic potential of a certain retroviral integration, the subsequent disruption of the affected gene needs to be shown on mRNA and protein level. Confirmation of the oncogenic potential could be proved by overexpression or inhibition of the expression of this suspected gene in a mouse BM transduction model.

Mechanism of retrovirus integratConsiderable effort has been put into the evaluation of the integration mechanism of both MLV and HIV integrases and pre-integration complexes, both from the field of virology and hematology. Although this lead to some contradicting evidence concerning the role of sequence preference (139,140) in virus integration, the data on the preference of MLV for promoter regions and the HIV lentivirus preference for in-gene integration in active transcription areas have been well established in different cell lines and HSC from different sources and species (78,141-146). The determining factor guiding MLV to the transcription start site (TSS) and HIV to in-gene regions seems to be the viral integrase, as was shown in viruses where the MLV integrase replaced the HIV integrase in lentivirus and vice versa (147). In microarray mRNA expression studies in human (38,148) and mouse (149), the gene expression profiles of the different immature subsets of HSC have been determined. This evidence could be used to corroborate the preference of retrovirus integration in HSC with their mRNA expression pattern (145,149). A recent report (100) compared all data published thus far on integration of retroviral vectors (HTLV-1, ASLV, FV, MLV, SIV, and HIV) with regards to the genomic features (Gene density, TSS distribution, CpG density and DNase hypersensitive sites surrounding the insertion) that were characteristic of each vector insertion pattern and concluded that different vectors with similar insertion patterns can have different insertional mutagenesis risks.

Retrovirus as gene therapy tool

First evidence of the possible oncogenic potential of retroviral transduction in a gene therapy setting emerged from a study in mice were mouse BM were transduced with a retroviral vector carrying *dLNGFR* (a truncated form of the Low-affinity Nerve Growth Factor Receptor) (110). This was a side effect that was expected when using retrovirus,

as was shown before in several studies in cell lines (150). However, while the incidence of malignant transformation was estimated to be 1 in 10^7 insertions, no exact data are available for the incidence of malignant transformation that results in a true oncogenic event in the transplanted patient. To address this question, several studies in mouse models were initiated. One study describes the introduction of the *MDR* gene or *dsRED* in C57BL6 lineage depleted BM cells to assess the effect of low or high numbers of insertion events. In this study high and low doses of retrovirus were used for transduction and transplantation of 58 mice. In the 22 mice transplanted with high dose retrovirus transduced cells, 8 mice developed leukemias. The mice transplanted with low dose virus did not develop leukemia, although they showed expression of the respective transgenes (151). In a follow up study, C57BL6 lineage depleted BM was transduced with truncated *CD34*, *ΔCD34* or *dsRED* with a gamma-retroviral vector and transplanted into a total of 44 lethally irradiated recipients. In this study low virus doses were used that result in on average one retrovirus integration per cell. In the mice that received these cells, transcriptional deregulation of genes near retrovirus insertion sites and dominance of certain clones was shown (152). After these initial observations in mice (84), an acute myeloid leukemia was observed in a rhesus macaque (153). Typically, the leukemias in these studies occur very late after transplantation. Therefore, several models have been developed that reduce the time to observation, allowing more rapid safety analysis of newly designed vectors. An *in vitro* model proposed by Modlich (154) employs a 2-weeks expansion and limiting dilution after transduction of lineage depleted mouse bone marrow, which reduces the time to analysis to approximately 4 weeks. Montini (155) showed that a *Cdnk2*$^{-/-}$ knockout mouse model could also be used to evaluate the safety of retroviral integrations, although the model has limited sensitivity because of the high background of the mouse model, and the time to leukemia is still rather long (200 days). Another study then showed that a self-inactivating (SIN) gamma-retroviral design, in which the transgene is expressed from an internal promoter rather than the LTR, is 20 times safer than the normal retroviral design with similar vectors (156). When lentiviral and gamma-retroviral vectors were compared in both models, the clonogenic capacity seemed dependent on the strength of the promoter, showing that lentiviral architecture is not inherently safer than gamma-retroviral architecture. Therefore, the TSS distribution does not seem to be the determining factor in safety as the promoter strength is. From these data it could be hypothesized, that in a situation where high expression of the transgene is needed for efficacy, one would rather choose a treatment that leads to multiple integrations with a weaker promoter (e.g. EIFα or PGK), than one integration with a very strong promoter (e.g. SF). The attention on oncogenic events after retroviral transduction followed by transplantation was raised after a patient, who received transduced cells developed a lymphoproliferative disease (111). Later that year a second patient was described. Both patients were shown to have a retroviral integration

near the *LMO2* T-cell oncogene (157). To date five patients presented with leukemia. In all cases a retrovirus integration near the *LMO2* oncogene was involved, although other possible oncogenes were also identified.

In vivo expansion of Hematopoietic Stem Cells

In vivo expansion of Hematopoietic Stem cells has been a focus point of research, not only because it would reduce the number of donor cells needed for stem cell transplantations, but also because it would make for example cord blood stem cell transplantations better suitable for adults recipients. Although UCB transplantations have been shown to be comparable in terms of survival and engraftment, an ongoing problem with cord blood stem cell transplantations is the limited dose of cells available and the resulting delayed engraftment (158,159). Moreover, gene therapy would benefit from *ex vivo* expansion of the transduced cells: less cells would need to be modified, which is desirable from a safety perspective. *Ex vivo* expansion protocols would also allow time to identify the virus integrations present in a transplant and assess the risk associated with transplanting cells carrying these integrations. Initial efforts using UCB cells met with the problem that the cells that expanded were mostly the mature cells, that do not engraft after transplantation, while the immature cells that do engraft failed to respond to cytokine stimulation(160).

Originally, HSC were cultured on stromal cells (Dexter cultures,(161)) The first liquid culture systems that were used to expand HSC using slow perfusion and local oxygenation showed that 10-20 fold enrichments in progenitor cells would be possible (162,163). $CD34^+$ progenitor cell liquid cultures also seemed promising, since they showed an expansion of progenitor cells by 50 fold in the presence of IL-1beta, IL-3, IL-6, G-CSF, GM-CSF and SCF. The adherence to fibronectin also had a positive effect on regenerative capacity of human HSC (164). Mouse HSC could be expanded in SCF, Flt3L and TPO (165-167). These preclinical attempts to expand the cells that eventually repopulate patients did unfortunately not yet translate into suitable protocols for use in clinical situations, since the increases in repopulating cells in humans were smaller than those observed in preclinical models (168,169). The automated culturing systems with local oxygenation and slow perfusion also did not result in large increases in repopulating cells (170,171). Extensive research has been performed to identify and analyze the difficulties that prevent the promising preclinical results with repopulating cells into clinically useful culturing conditions. Several stem cell assays have been developed, such as the *in vitro* colony assays for CFU-E, CFU-GEMM etc. and although these models are reproducible, it remains unclear whether the cells that are measured in these assays are also the cells that contribute to hematopoietic reconstitution. Even the mouse transplantation models, such as competitive repopulation units (CRU) and SCID-repopulating cells (SRC or huSRC) are not completely adequate, since they either

use a mouse models that does not reflect the complexity of the human immune system response (CRU) or do not reproduce human marrow homing ability (SRC)(172). The problems with *ex vivo* culturing that were identified included cell cycle abnormalities, acquired homing defects and the induction of apoptosis. Glimm *et al.* (173) showed that cytokine exposure resulted in G1 entry, which reduced the engraftment capacity. *Ex vivo* culturing procedures were also shown to result in a delayed engraftment of $CD34^+$ umbilical cord blood cells in NOD/SCID mice (174,175). In $CD34^+$ cells, the expression of VLA-5, an adhesion protein important in homing, was increased after liquid culture in the presence of Flt3L, TPO and SCF(176), but this did not result in increased engraftment (177).

Defective homing capacity was observed after *ex vivo* expansion of murine bone marrow cells using a cytokine cocktail consisting of IL-3, IL-6, IL-11 and SCF (178,179). This effect is possibly caused by the loss of the alpha4beta1 and alpha5beta1 integrins (176) in culture. Another frequently reported result of *ex vivo* culturing procedures is apoptosis induction. The cultured HSC upregulate CD95 (Fas) which makes them more sensitive to FasL apoptosis signals. Apparently, these signals are encountered when the cells home to the bone marrow, since engraftment of cultured cells can be increased by a monoclonal blocking Fas/CD95 antibody (180).

The same results were observed in fresh UCB or PBSC $CD34^+$, where FLIP (FLICE like inhibitory protein), which can block Fas-meditated apoptosis, was highly expressed cells, showing that apoptosis induction through Fas should be escaped for successful engraftment (181). Other changes that prime the HSC for apoptosis have also been identified, such as a decrease in the anti-apoptotic Bcl-2 protein (182) and an increased caspase activation (183).

Rationale for the studies described in this thesis

Gene therapy has proven to be a suitable approach for patients suffering from ADA SCID (47) and XSCID (45,46), that otherwise have limited options for treatment. For ADA SCID gene replacement therapy has been developed, but it is, due to its cost, not always available. For XSCID, bone marrow transplantation are only an option in the presence of suitable bone marrow donors. The use retroviral vectors to deliver therapeutic genes *ex vivo* to hematopoietic stem cells is a powerful method to equip them with therapeutic transgenes. The method is clinically relevant, even though five cases of malignant proliferation in the XSCID trials were observed, which were caused by the insertion of the viral vector near an oncogene. The ADA SCID clinical phase I/II trail in Milan, aimed at a different disease, but using a similar gamma-retroviral backbone (47) to date has not met with these problems. The difference in the occurrence of malignant proliferation between these trials suggests that minor differences (e.g. growth factors and culturing conditions used in the transduction procedure or perhaps the transduced

target cell population) in the protocols could be the factors that determine whether malignant transformations will or will not occur and thereby determine whether gene therapy protocols using gamma-retroviral vectors are safe.

We addressed the safety of retroviral vector systems for gene delivery into hematopoietic stem cells by performing a large scale study in mice. In this study, we converted increased the observation time in experiments where we expressed a signaling gene, wild-type Stat5b or the marker gene EGFP to be able to analyze the relation between gamma-retroviral insertion and leukemia occurrence. The mice that received transplants with the transduced cells were monitored for over seven months after which their bone marrow cells were transplanted into secondary recipients. This procedure is thought to introduce the replicative stress associated with life-long maintenance of the hematopoietic system and therefore presents a model of the maximum amount of cell division that a transplanted transduced cell will have to go through when transplanted into a recipient in a clinical gene therapy protocol. Furthermore, the virus dose (MOI) used in these experiments was deliberately kept low, to achieve an average of one integration per cell. This approach would limit collaboration between different virus integrations in one cell, that would possibly increase the leukemia incidence in our model. The malignancies that were observed were carefully analyzed and retransplanted, which allowed identification of single transformed clones and the location of the retroviral integration in several cases. The genes surrounding the retrovirus integrations were subsequently identified. The resulting list of genes provides a summary of the integration behavior of the gamma-retroviral vector, when used to transduced mouse hematopoietic stem cells. This integration behavior in mice is compared to integration patterns in human and rhesus monkey cells, which provides evidence that retrovirus integrates into specific loci within the genomes of hematopoietic stem cells of these species with remarkable similarities in the insertion behavior, in the function of the genes nearby and in the common insertion sites that were involved.

REFERENCES

1. van BEKKUM D, van PUTTEN L, de VRIES M. Antihost reactivity and tolerance of the graft in relation to secondary disease in radiation chimeras. Ann N Y Acad Sci. 1962;99:550-63.
2. Medvinsky A, Dzierzak E. Definitive hematopoiesis is autonomously initiated by the AGM region. Cell. 1996 Sep 20;86(6):897-906.
3. Orelio C, Haak E, Peeters M, Dzierzak E. Interleukin-1-mediated hematopoietic cell regulation in the aorta-gonad-mesonephros region of the mouse embryo. Blood. 2008 Dec 15;112(13):4895-4904.
4. Eilken HM, Nishikawa S, Schroeder T. Continuous single-cell imaging of blood generation from haemogenic endothelium. Nature. 2009 Feb 12;457(7231):896-900.
5. Boisset J, van Cappellen W, Andrieu-Soler C, Galjart N, Dzierzak E, Robin C. *In vivo* imaging of haematopoietic cells emerging from the mouse aortic endothelium. Nature. 2010 Mar 4;464(7285):116-120.

6. de Bruijn MF, Peeters MC, Luteijn T, Visser P, Speck NA, Dzierzak E. CFU-S(11) activity does not localize solely with the aorta in the aorta-gonad-mesonephros region. Blood. 2000 Oct 15;96(8):2902-2904.
7. Müller AM, Medvinsky A, Strouboulis J, Grosveld F, Dzierzak E. Development of hematopoietic stem cell activity in the mouse embryo. Immunity. 1994;1(4):291-301.
8. Gluckman E, Broxmeyer HA, Auerbach AD, Friedman HS, Douglas GW, Devergie A, et al. Hematopoietic reconstitution in a patient with Fanconi's anemia by means of umbilical-cord blood from an HLA-identical sibling. N Engl J Med. 1989;321(17):1174-8.
9. Socinski MA, Cannistra SA, Elias A, Antman KH, Schnipper L, Griffin JD. Granulocyte-macrophage colony stimulating factor expands the circulating haemopoietic progenitor cell compartment in man. Lancet. 1988;1(8596):1194-8.
10. Haas R, Ho AD, Bredthauer U, Cayeux S, Egerer G, Knauf W, et al. Successful autologous transplantation of blood stem cells mobilized with recombinant human granulocyte-macrophage colony-stimulating factor. Exp Hematol. 1990;18(2):94-8.
11. VAN BEKKUM DW, VOS O. Immunological aspects of homo- and heterologous bone marrow transplantation in irradiated animals. J Cell Physiol Suppl. 1957;50(Suppl 1):139-56.
12. KURNICK NB, MONTANO A, GERDES JC, FEDER BH. Preliminary observations on the treatment of postirradiation hematopoietic depression in man by the infusion of stored autogenous bone marrow. Ann Intern Med. 1958;49(5):973-86.
13. Bernard A, Boumsell L, Dausset J, Milstein C, Schlossman S. Leucocyte Typing: Human Leucocyte Differentiation Antigens Detected by Monoclonal Antibodies. Springer Verlag Berlin; 1984.
14. Haynes B, Nadler L, Bernstein I, Reinherz E. Human Myeloid and Hematopoietic Cells. Springer-Verlag, New York; 1988.
15. Civin CI, Strauss LC, Brovall C, Fackler MJ, Schwartz JF, Shaper JH. Antigenic analysis of hematopoiesis. III. A hematopoietic progenitor cell surface antigen defined by a monoclonal antibody raised against KG-1a cells. J Immunol. 1984;133(1):157-65.
16. Silvestri FF, Banavali SD, Hulette BC, Civin CI, Preisler HD. Isolation and characterization of the CD34+ hematopoietic progenitor cells from the peripheral blood of patients with chronic myeloid leukemia. Int J Cell Cloning. 1991;9(5):474-90.
17. Simmons DL, Satterthwaite AB, Tenen DG, Seed B. Molecular cloning of a cDNA encoding CD34, a sialomucin of human hematopoietic stem cells. J Immunol. 1992;148(1):267-71.
18. Andrews RG, Singer JW, Bernstein ID. Monoclonal antibody 12-8 recognizes a 115-kd molecule present on both unipotent and multipotent hematopoietic colony-forming cells and their precursors. Blood. 1986;67(3):842-5.
19. Terstappen LW, Huang S, Safford M, Lansdorp PM, Loken MR. Sequential generations of hematopoietic colonies derived from single nonlineage-committed CD34+CD38- progenitor cells. Blood. 1991;77(6):1218-27.
20. Verstegen MM, van Hennik PB, Terpstra W, van den Bos C, Wielenga JJ, van Rooijen N, et al. Transplantation of human umbilical cord blood cells in macrophage-depleted SCID mice: evidence for accessory cell involvement in expansion of immature CD34+CD38- cells. Blood. 1998;91(6):1966-76.
21. Dorrell C, Gan OI, Pereira DS, Hawley RG, Dick JE. Expansion of human cord blood CD34(+)CD38(-) cells in *ex vivo* culture during retroviral transduction without a corresponding increase in SCID repopulating cell (SRC) frequency: dissociation of SRC phenotype and function. Blood. 2000 Jan 1;95(1):102-110.
22. Yin AH, Miraglia S, Zanjani ED, Almeida-Porada G, Ogawa M, Leary AG, et al. AC133, a novel marker for human hematopoietic stem and progenitor cells. Blood. 1997;90(12):5002-12.
23. Brenner S, Ryser MF, Whiting-Theobald NL, Gentsch M, Linton GF, Malech HL. The late dividing population of gamma-retroviral vector transduced human mobilized peripheral blood progenitor cells contributes most to gene-marked cell engraftment in nonobese diabetic/severe combined immunodeficient mice. Stem Cells. 2007;25(7):1807-13.
24. Spangrude GJ, Heimfeld S, Weissman IL. Purification and characterization of mouse hematopoietic stem cells. Science. 1988;241(4861):58-62.

25. Seita J, Ema H, Ooehara J, Yamazaki S, Tadokoro Y, Yamasaki A, et al. Lnk negatively regulates self-renewal of hematopoietic stem cells by modifying thrombopoietin-mediated signal transduction. Proc Natl Acad Sci U S A. 2007;104(7):2349-54.
26. Brown J, Greaves MF, Molgaard HV. The gene encoding the stem cell antigen, CD34, is conserved in mouse and expressed in haemopoietic progenitor cell lines, brain, and embryonic fibroblasts. Int Immunol. 1991;3(2):175-84.
27. Ikuta K, Kina T, MacNeil I, Uchida N, Peault B, Chien YH, et al. A developmental switch in thymic lymphocyte maturation potential occurs at the level of hematopoietic stem cells. Cell. 1990;62(5):863-74.
28. Smith LG, Weissman IL, Heimfeld S. Clonal analysis of hematopoietic stem-cell differentiation *in vivo*. Proc Natl Acad Sci U S A. 1991;88(7):2788-92.
29. Ikuta K, Weissman IL. Evidence that hematopoietic stem cells express mouse c-kit but do not depend on steel factor for their generation. Proc Natl Acad Sci U S A. 1992;89:1502-6.
30. Spangrude GJ, Brooks DM, Tumas DB. Long-term repopulation of irradiated mice with limiting numbers of purified hematopoietic stem cells: *in vivo* expansion of stem cell phenotype but not function. Blood. 1995;85(4):1006-16.
31. Kim M, Cooper DD, Hayes SF, Spangrude GJ. Rhodamine-123 staining in hematopoietic stem cells of young mice indicates mitochondrial activation rather than dye efflux. Blood. 1998 Jun 1;91(11):4106-4117.
32. Adolfsson J, Borge OJ, Bryder D, Theilgaard-Mönch K, Astrand-Grundström I, Sitnicka E, et al. Upregulation of Flt3 expression within the bone marrow Lin(-)Sca1(+)c-kit(+) stem cell compartment is accompanied by loss of self-renewal capacity. Immunity. 2001;15(4):659-69.
33. Wognum AW, Visser TP, Peters K, Bierhuizen MF, Wagemaker G. Stimulation of mouse bone marrow cells with kit ligand, FLT3 ligand, and thrombopoietin leads to efficient retrovirus-mediated gene transfer to stem cells, whereas interleukin 3 and interleukin 11 reduce transduction of short- and long-term repopulating cells. Hum Gene Ther. 2000;11:2129-41.
34. Ivanova NB, Dimos JT, Schaniel C, Hackney JA, Moore KA, Lemischka IR. A stem cell molecular signature. Science. 2002;298:601-4.
35. Ramalho-Santos M, Yoon S, Matsuzaki Y, Mulligan RC, Melton DA. : transcriptional profiling of embryonic and adult stem cells. Science. 2002;298:597-600.
36. Fortunel NO, Otu HH, Ng H, Chen J, Mu X, Chevassut T, et al. Comment on " 'Stemness': transcriptional profiling of embryonic and adult stem cells" and "a stem cell molecular signature". Science. 2003 Oct 17;302(5644):393; author reply 393.
37. Terskikh AV, Miyamoto T, Chang C, Diatchenko L, Weissman IL. Gene expression analysis of purified hematopoietic stem cells and committed progenitors. Blood. 2003;102:94-101.
38. Georgantas RW, Tanadve V, Malehorn M, Heimfeld S, Chen C, Carr L, et al. Microarray and serial analysis of gene expression analyses identify known and novel transcripts overexpressed in hematopoietic stem cells. Cancer Res. 2004;64:4434-41.
39. Chambers SM, Boles NC, Lin KK, Tierney MP, Bowman TV, Bradfute SB, et al. Hematopoietic fingerprints: an expression database of stem cells and their progeny. Cell Stem Cell. 2007 Nov;1(5):578-591.
40. Eckfeldt CE, Mendenhall EM, Flynn CM, Wang TF, Pickart MA, Grindle SM, et al. Functional analysis of human hematopoietic stem cell gene expression using zebrafish. PLoS Biol. 2005;3:e254.
41. Deneault E, Cellot S, Faubert A, Laverdure J, Fréchette M, Chagraoui J, et al. A functional screen to identify novel effectors of hematopoietic stem cell activity. Cell. 2009 Apr 17;137(2):369-379.
42. Ott MG, Schmidt M, Schwarzwaelder K, Stein S, Siler U, Koehl U, et al. Correction of X-linked chronic granulomatous disease by gene therapy, augmented by insertional activation of MDS1-EVI1, PRDM16 or SETBP1. Nat Med. 2006;12(4):401-9.
43. Bornhäuser M, Eger L, Oelschlaegel U, Auffermann-Gretzinger S, Kiani A, Schetelig J, et al. Rapid reconstitution of dendritic cells after allogeneic transplantation of CD133+ selected hematopoietic stem cells. Leukemia. 2005;19(1):161-5.
44. van Hennik PB, Verstegen MM, Bierhuizen MF, Limon A, Wognum AW, Cancelas JA, et al. Highly efficient transduction of the green fluorescent protein gene in human umbilical cord blood stem cells capable of

cobblestone formation in long-term cultures and multilineage engraftment of immunodeficient mice. Blood. 1998;92:4013-22.
45. Hacein-Bey-Abina S, Le Deist F, Carlier F, Bouneaud C, Hue C, De Villartay J, et al. Sustained correction of X-linked severe combined immunodeficiency by *ex vivo* gene therapy. N Engl J Med. 2002;346(16):1185-93.
46. Gaspar HB, Parsley KL, Howe S, King D, Gilmour KC, Sinclair J, et al. Gene therapy of X-linked severe combined immunodeficiency by use of a pseudotyped gammaretroviral vector. Lancet. 2004;364:2181-7.
47. Aiuti A, Slavin S, Aker M, Ficara F, Deola S, Mortellaro A, et al. Correction of ADA-SCID by stem cell gene therapy combined with nonmyeloablative conditioning. Science. 2002;296:2410-3.
48. Antoine C, Muller S, Cant A, Cavazzana-Calvo M, Veys P, Vossen J, et al. Long-term survival and transplantation of haemopoietic stem cells for immunodeficiencies: report of the European experience 1968-99. Lancet. 2003;361:553-60.
49. Grewal SS, Barker JN, Davies SM, Wagner JE. Unrelated donor hematopoietic cell transplantation: marrow or umbilical cord blood? Blood. 2003;101:4233-44.
50. Patel DD, Gooding ME, Parrott RE, Curtis KM, Haynes BF, Buckley RH. Thymic function after hematopoietic stem-cell transplantation for the treatment of severe combined immunodeficiency. N Engl J Med. 2000;342:1325-32.
51. Urnov FD, Miller JC, Lee YL, Beausejour CM, Rock JM, Augustus S, et al. Highly efficient endogenous human gene correction using designed zinc-finger nucleases. Nature. 2005;435:646-51.
52. ZINDER ND, LEDERBERG J. Genetic exchange in Salmonella. J Bacteriol. 1952;64(5):679-99.
53. TEMIN HM. THE EFFECTS OF ACTINOMYCIN D ON GROWTH OF ROUS SARCOMA VIRUS *IN VITRO*. Virology. 1963;20:577-82.
54. Smith HO, Wilcox KW. A restriction enzyme from Hemophilus influenzae. I. Purification and general properties. J Mol Biol. 1970;51(2):379-91.
55. Danna K, Nathans D. Specific cleavage of simian virus 40 DNA by restriction endonuclease of Hemophilus influenzae. Proc Natl Acad Sci U S A. 1971;68(12):2913-7.
56. Gilbert W, Maxam A. The nucleotide sequence of the lac operator. Proc Natl Acad Sci U S A. 1973;70(12):3581-4.
57. Sanger F, Coulson AR. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J Mol Biol. 1975;94(3):441-8.
58. Ganem D, Nussbaum AL, Davoli D, Fareed GC. Propagation of a segment of bacteriophage lamda-DNA in monkey cells after covalent linkage to a defective simian virus 40 genome. Cell. 1976;7(3):349-59.
59. Mulligan RC, Howard BH, Berg P. Synthesis of rabbit beta-globin in cultured monkey kidney cells following infection with a SV40 beta-globin recombinant genome. Nature. 1979;277(5692):108-14.
60. Graham FL, van der Eb AJ. Transformation of rat cells by DNA of human adenovirus 5. Virology. 1973;54(2):536-9.
61. Cline MJ, Stang H, Mercola K, Morse L, Ruprecht R, Brown J, et al. Gene transfer in intact animals. Nature. 1980;284(5755):422-5.
62. Mercola KE, Stang HD, Browne J, Salser W, Cline MJ. Insertion of a new gene of viral origin into bone marrow cells of mice. Science. 1980;208(4447):1033-5.
63. Wade N. UCLA gene therapy racked by friendly fire. Science. 1980;210(4469):509-11.
64. Wade N. Gene therapy pioneer draws Mikadoesque rap. Science. 1981;212(4500):1253.
65. Wade N. Gene therapy caught in more entanglements. Science. 1981 Apr 3;212(4490):24-25.
66. Wei CM, Gibson M, Spear PG, Scolnick EM. Construction and isolation of a transmissible retrovirus containing the src gene of Harvey murine sarcoma virus and the thymidine kinase gene of herpes simplex virus type 1. J Virol. 1981;39(3):935-44.
67. Tabin CJ, Hoffmann JW, Goff SP, Weinberg RA. Adaptation of a retrovirus as a eucaryotic vector transmitting the herpes simplex virus thymidine kinase gene. Mol Cell Biol. 1982;2(4):426-36.
68. Miller DG, Adam MA, Miller AD. Gene transfer by retrovirus vectors occurs only in cells that are actively replicating at the time of infection. Mol Cell Biol. 1990;10(8):4239-42.

69. Pear WS, Nolan GP, Scott ML, Baltimore D. Production of high-titer helper-free retroviruses by transient transfection. Proc Natl Acad Sci U S A. 1993;90(18):8392-6.
70. Miller AD, Jolly DJ, Friedmann T, Verma IM. A transmissible retrovirus expressing human hypoxanthine phosphoribosyltransferase (HPRT): gene transfer into cells obtained from humans deficient in HPRT. Proc Natl Acad Sci U S A. 1983;80(15):4709-13.
71. Joyner A, Keller G, Phillips RA, Bernstein A. Retrovirus transfer of a bacterial gene into mouse haematopoietic progenitor cells. Nature. 1983;305(5934):556-8.
72. Williams DA, Lemischka IR, Nathan DG, Mulligan RC. Introduction of new genetic material into pluripotent haematopoietic stem cells of the mouse. Nature. 1984;310(5977):476-80.
73. Dunbar CE, Young NS. Gene marking and gene therapy directed at primary hematopoietic cells. Curr. Opin. Hematol. 1996 Nov;3(6):430-437.
74. Ochman H, Gerber AS, Hartl DL. Genetic applications of an inverse polymerase chain reaction. Genetics. 1988;120(3):621-3.
75. Devon RS, Porteous DJ, Brookes AJ. Splinkerettes--improved vectorettes for greater efficiency in PCR walking. Nucleic Acids Res. 1995 May 11;23(9):1644-1645.
76. Eggert H, Bergemann K, Saumweber H. Molecular screening for P-element insertions in a large genomic region of Drosophila melanogaster using polymerase chain reaction mediated by the vectorette. Genetics. 1998 Jul;149(3):1427-1434.
77. Tolar J, Osborn M, Bell S, McElmurry R, Xia L, Riddle M, et al. Real-time *in vivo* imaging of stem cells following transgenesis by transposition. Mol. Ther. 2005 Jul;12(1):42-48.
78. Wu X, Li Y, Crise B, Burgess SM. Transcription start regions in the human genome are favored targets for MLV integration. Science. 2003;300:1749-51.
79. Mueller PR, Wold B. *In vivo* footprinting of a muscle specific enhancer by ligation mediated PCR. Science. 1989 Nov 10;246(4931):780-786.
80. Pfeifer GP, Steigerwald SD, Mueller PR, Wold B, Riggs AD. Genomic sequencing and methylation analysis by ligation mediated PCR. Science. 1989 Nov 10;246(4931):810-813.
81. Izsvák Z, Ivics Z. Two-stage ligation-mediated PCR enhances the detection of integrated transgenic DNA. BioTechniques. 1993 Nov;15(5):814-818.
82. Schmidt M, Hoffmann G, Wissler M, Lemke N, Mussig A, Glimm H, et al. Detection and direct genomic sequencing of multiple rare unknown flanking DNA in highly complex samples. Hum Gene Ther. 2001;12:743-9.
83. Laufs S, Guenechea G, Gonzalez-Murillo A, Zsuzsanna Nagy K, Luz Lozano M, del Val C, et al. Lentiviral vector integration sites in human NOD/SCID repopulating cells. J Gene Med. 2006 Oct;8(10):1197-1207.
84. Kustikova OS, Modlich U, Fehse B. Retroviral insertion site analysis in dominant haematopoietic clones. Methods Mol. Biol. 2009;506:373-390.
85. Schmidt M, Schwarzwaelder K, Bartholomae C, Zaoui K, Ball C, Pilz I, et al. High-resolution insertion-site analysis by linear amplification-mediated PCR (LAM-PCR). Nat. Methods. 2007 Dec;4(12):1051-1057.
86. Lagresle-Peyrou C, Yates F, Malassis-Séris M, Hue C, Morillon E, Garrigue A, et al. Long-term immune reconstitution in RAG-1-deficient mice treated by retroviral gene therapy: a balance between efficiency and toxicity. Blood. 2006;107(1):63-72.
87. Dzierzak EA, Papayannopoulou T, Mulligan RC. Lineage-specific expression of a human beta-globin gene in murine bone marrow transplant recipients reconstituted with retrovirus-transduced stem cells. Nature. 1988;331(6151):35-41.
88. Lim B, Apperley JF, Orkin SH, Williams DA. Long-term expression of human adenosine deaminase in mice transplanted with retrovirus-infected hematopoietic stem cells. Proc Natl Acad Sci U S A. 1989;86(22):8892-6.
89. Correll PH, Kew Y, Perry LK, Brady RO, Fink JK, Karlsson S. Expression of human glucocerebrosidase in long-term reconstituted mice following retroviral-mediated gene transfer into hematopoietic stem cells. Hum Gene Ther. 1990;1(3):277-87.

90. Wolfe JH, Sands MS, Barker JE, Gwynn B, Rowe LB, Vogler CA, et al. Reversal of pathology in murine mucopolysaccharidosis type VII by somatic cell gene transfer. Nature. 1992;360(6406):749-53.
91. Bodine DM, McDonagh KT, Brandt SJ, Ney PA, Agricola B, Byrne E, et al. Development of a high-titer retrovirus producer cell line capable of gene transfer into rhesus monkey hematopoietic stem cells. Proc Natl Acad Sci U S A. 1990;87(10):3738-42.
92. van Beusechem VW, Kukler A, Heidt PJ, Valerio D. Long-term expression of human adenosine deaminase in rhesus monkeys transplanted with retrovirus-infected bone-marrow cells. Proc Natl Acad Sci U S A. 1992;89(16):7640-4.
93. Tisdale JF, Hanazono Y, Sellers SE, Agricola BA, Metzger ME, Donahue RE, et al. *Ex vivo* expansion of genetically marked rhesus peripheral blood progenitor cells results in diminished long-term repopulating ability. Blood. 1998;92(4):1131-41.
94. Dunbar CE, Takatoku M, Donahue RE. The impact of *ex vivo* cytokine stimulation on engraftment of primitive hematopoietic cells in a non-human primate model. Ann N Y Acad Sci. 2001;938:236-44; discussion 244.
95. Schimmenti S, Boesen J, Claassen EA, Valerio D, Einerhand MP. Long-term genetic modification of rhesus monkey hematopoietic cells following transplantation of adenoassociated virus vector-transduced CD34+ cells. Hum Gene Ther. 1998;9(18):2727-34.
96. Miyoshi H, Smith KA, Mosier DE, Verma IM, Torbett BE. Transduction of human CD34+ cells that mediate long-term engraftment of NOD/SCID mice by HIV vectors. Science. 1999;283(5402):682-6.
97. Hirata RK, Miller AD, Andrews RG, Russell DW. Transduction of hematopoietic cells by foamy virus vectors. Blood. 1996;88(9):3654-61.
98. Giblett ER, Anderson JE, Cohen F, Pollara B, Meuwissen HJ. Adenosine-deaminase deficiency in two patients with severely impaired cellular immunity. Lancet. 1972;2(7786):1067-9.
99. Baltimore D. RNA-dependent DNA polymerase in virions of RNA tumour viruses. Nature. 1970 Jun 27;226(5252):1209-1211.
100. Derse D, Crise B, Li Y, Princler G, Lum N, Stewart C, et al. Human T-cell leukemia virus type 1 integration target sites in the human genome: comparison with those of other retroviruses. J Virol. 2007;81(12):6731-41.
101. Coffin JM, Varmus HE. Retroviruses. CSHL Press; 1997.
102. Lund A, Duch M, Pedersen F. Transcriptional Silencing of Retroviral Vectors. J. Biomed. Sci. 1996 Dec;3(6):365-378.
103. Yu SF, von Rüden T, Kantoff PW, Garber C, Seiberg M, Rüther U, et al. Self-inactivating retroviral vectors designed for transfer of whole genes into mammalian cells. Proc Natl Acad Sci U S A. 1986 May;83(10):3194-3198.
104. Hanahan D, Weinberg RA. The hallmarks of cancer. Cell. 2000;100:57-70.
105. Rangarajan A, Weinberg RA. Opinion: Comparative biology of mouse versus human cells: modelling human cancer in mice. Nat Rev Cancer. 2003;3:952-9.
106. Fialkow PJ, Janssen JW, Bartram CR. Clonal remissions in acute nonlymphocytic leukemia: evidence for a multistep pathogenesis of the malignancy. Blood. 1991;77:1415-7.
107. Nichols J, Nimer SD. Transcription factors, translocations, and leukemia. Blood. 1992;80:2953-63.
108. Prassolov V, Meyer J, Brandenburg G, Hannemann J, Bergemann J, Ostertag W, et al. Functional identification of secondary mutations inducing autonomous growth in synergy with a truncated interleukin-3 receptor: implications for multi-step oncogenesis. Exp Hematol. 2001;29:756-65.
109. Fan H. Leukemogenesis by Moloney murine leukemia virus: a multistep process. Trends Microbiol. 1997;5:74-82.
110. 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.
111. Hacein-Bey-Abina S, Von Kalle C, Schmidt M, McCormack MP, Wulffraat N, Leboulch P, et al. LMO2-associated clonal T cell proliferation in two patients after gene therapy for SCID-X1. Science. 2003;302:415-9.

112. Hacein-Bey-Abina S, Garrigue A, Wang GP, Soulier J, Lim A, Morillon E, et al. Insertional oncogenesis in 4 patients after retrovirus-mediated gene therapy of SCID-X1. J. Clin. Invest. 2008 Sep;118(9):3132-3142.
113. Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science. 1995;270(5235):467-70.
114. Schena M, Shalon D, Heller R, Chai A, Brown PO, Davis RW. Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. Proc Natl Acad Sci U S A. 1996;93(20):10614-9.
115. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics. 2003;4(2):249-64.
116. Wu Zhijin IRA. A Model Based Background Adjustment for Oligonucleotide Expression Arrays. Johns Hopkins University, Dept. of Biostatistics Working Papers. 2004;
117. Millenaar FF, Okyere J, May ST, van Zanten M, Voesenek LACJ, Peeters AJM. How to decide? Different methods of calculating gene expression from short oligonucleotide array data will give different results. BMC Bioinformatics. 2006;7:137.
118. Ge Y, Dudoit S, Speed T. Resampling-based multiple testing for microarray data analysis. TEST. 2003;12:1-44.
119. de Ridder D, Staal FJ, van Dongen JJ, Reinders MJ. Maximum significance clustering of oligonucleotide microarrays. Bioinformatics. 2006;22:326-31.
120. Dik WA, Pike-Overzet K, Weerkamp F, de Ridder D, de Haas EF, Baert MR, et al. New insights on human T cell development by quantitative T cell receptor gene rearrangement studies and gene expression profiling. J Exp Med. 2005;201:1715-23.
121. Dennis G, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, et al. DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol. 2003;4:P3.
122. Al-Shahrour F, Díaz-Uriarte R, Dopazo J. FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. Bioinformatics. 2004;20(4):578-80.
123. Al-Shahrour F, Minguez P, Vaquerizas JM, Conde L, Dopazo J. BABELOMICS: a suite of web tools for functional annotation and analysis of groups of genes in high-throughput experiments. Nucleic Acids Res. 2005;33(Web Server issue):W460-4.
124. Alexa A, Rahnenfuhrer J. topGO: topGO: Enrichment analysis for Gene Ontology. R package version 1.10.1. [Internet]. 2003;Available from: http://www.bioconductor.org/packages/2.0/bioc/html/topGO.html
125. Ben-Shaul Y, Bergman H, Soreq H. Identifying subtle interrelated changes in functional gene categories using continuous measures of gene expression. Bioinformatics. 2005;21(7):1129-37.
126. Lee HK, Braynen W, Keshav K, Pavlidis P. ErmineJ: tool for functional analysis of gene expression data sets. BMC Bioinformatics. 2005;6:269.
127. Thomas PD, Kejariwal A, Guo N, Mi H, Campbell MJ, Muruganujan A, et al. Applications for protein sequence-function evolution data: mRNA/protein expression analysis and coding SNP scoring tools. Nucleic Acids Res. 2006;34(Web Server issue):W645-50.
128. Abdueva D, Wing MR, Schaub B, Triche TJ. Experimental comparison and evaluation of the Affymetrix exon and U133Plus2 GeneChip arrays. PLoS ONE. 2007;2(9):e913.
129. Akagi K, Suzuki T, Stephens RM, Jenkins NA, Copeland NG. RTCGD: retroviral tagged cancer gene database. Nucleic Acids Res. 2004;32:D523-7.
130. Justice MJ, Morse HC3, Jenkins NA, Copeland NG. Identification of Evi-3, a novel common site of retroviral integration in mouse AKXD B-cell lymphomas. J Virol. 1994;68(3):1293-300.
131. Kim R, Trubetskoy A, Suzuki T, Jenkins NA, Copeland NG, Lenz J. Genome-based identification of cancer genes by proviral tagging in mouse retrovirus-induced T-cell lymphomas. J Virol. 2003;77:2056-62.
132. Joosten M, Vankan-Berkhoudt Y, Tas M, Lunghi M, Jenniskens Y, Parganas E, et al. Large-scale identification of novel potential disease loci in mouse leukemia applying an improved strategy for cloning common virus integration sites. Oncogene. 2002;21:7247-55.

133. Martín-Hernández J, Sørensen AB, Pedersen FS. Murine leukemia virus proviral insertions between the N-ras and unr genes in B-cell lymphoma DNA affect the expression of N-ras only. J. Virol. 2001 Dec;75(23):11907-11912.
134. Lund AH, Turner G, Trubetskoy A, Verhoeven E, Wientjens E, Hulsman D, et al. Genome-wide retroviral insertional tagging of genes involved in cancer in Cdkn2a-deficient mice. Nat Genet. 2002;32:160-5.
135. Bijl J, Sauvageau M, Thompson A, Sauvageau G. High incidence of proviral integrations in the Hoxa locus in a new model of E2a-PBX1-induced B-cell leukemia. Genes Dev. 2005;19(2):224-33.
136. Hwang HC, Martins CP, Bronkhorst Y, Randel E, Berns A, Fero M, et al. Identification of oncogenes collaborating with p27Kip1 loss by insertional mutagenesis and high-throughput insertion site analysis. Proc Natl Acad Sci U S A. 2002;99:11293-8.
137. Castilla LH, Perrat P, Martinez NJ, Landrette SF, Keys R, Oikemus S, et al. Identification of genes that synergize with Cbfb-MYH11 in the pathogenesis of acute myeloid leukemia. Proc Natl Acad Sci U S A. 2004;101(14):4924-9.
138. Iwasaki M, Kuwata T, Yamazaki Y, Jenkins NA, Copeland NG, Osato M, et al. Identification of cooperative genes for NUP98-HOXA9 in myeloid leukemogenesis using a mouse model. Blood. 2005;105(2):784-93.
139. Wu X, Luke BT, Burgess SM. Redefining the common insertion site. Virology. 2006;344:292-5.
140. Holman AG, Coffin JM. Symmetrical base preferences surrounding HIV-1, avian sarcoma/leukosis virus, and murine leukemia virus integration sites. Proc Natl Acad Sci U S A. 2005;102:6103-7.
141. Ho ES, van Leeuwen B, O, ', Neill HC. Association of repeat sequences with integrated retroviruses in a murine leukaemia cell line. Leuk Res. 1996;20:421-7.
142. 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;101:2191-8.
143. Laufs S, Nagy KZ, Giordano FA, 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;10:874-81.
144. 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;110:521-9.
145. 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.
146. Hematti P, Hong BK, Ferguson C, Adler R, Hanawa H, Sellers S, et al. Distinct genomic integration of MLV and SIV vectors in primate hematopoietic stem and progenitor cells. PLoS Biol. 2004;2:e423.
147. Lewinski MK, Yamashita M, Emerman M, Ciuffi A, Marshall H, Crawford G, et al. Retroviral DNA integration: viral and cellular determinants of target-site selection. PLoS Pathog. 2006;2(6):e60.
148. Ng YY, van Kessel B, Lokhorst HM, Baert MR, van den Burg CM, Bloem AC, et al. Gene-expression profiling of CD34+ cells from various hematopoietic stem-cell sources reveals functional differences in stem-cell activity. J Leukoc Biol. 2004;75:314-23.
149. Weidhaas JB, Angelichio EL, Fenner S, Coffin JM. Relationship between retroviral DNA integration and gene expression. J Virol. 2000;74:8382-9.
150. Stocking C, Bergholz U, Friel J, Klingler K, Wagener T, Starke C, et al. Distinct classes of factor-independent mutants can be isolated after retroviral mutagenesis of a human myeloid stem cell line. Growth Factors. 1993;8:197-209.
151. Modlich U, Kustikova OS, Schmidt M, Rudolph C, Meyer J, Li Z, et al. Leukemias following retroviral transfer of multidrug resistance 1 (MDR1) are driven by combinatorial insertional mutagenesis. Blood. 2005;105:4235-46.
152. Kustikova OS, Wahlers A, Kuhlcke K, Stahle B, Zander AR, Baum C, et al. Dose finding with retroviral vectors: correlation of retroviral vector copy numbers in single cells with gene transfer efficiency in a cell population. Blood. 2003;102(12):3934-7.
153. Seggewiss R, Pittaluga S, Adler RL, Guenaga FJ, Ferguson C, Pilz IH, et al. Acute myeloid leukemia is associated with retroviral gene transfer to hematopoietic progenitor cells in a rhesus macaque. Blood. 2006;107(10):3865-7.

154. Modlich U, Bohne J, Schmidt M, von Kalle C, Knoss S, Schambach A, et al. Cell-culture assays reveal the importance of retroviral vector design for insertional genotoxicity. Blood. 2006;108:2545-53.
155. Montini E, Cesana D, Schmidt M, Sanvito F, Ponzoni M, Bartholomae C, et al. Hematopoietic stem cell gene transfer in a tumor-prone mouse model uncovers low genotoxicity of lentiviral vector integration. Nat Biotechnol. 2006;24:687-96.
156. Zychlinski D, Schambach A, Modlich U, Maetzig T, Meyer J, Grassman E, et al. Physiological promoters reduce the genotoxic risk of integrating gene vectors. Mol Ther. 2008;16(4):718-25.
157. McCormack MP, Rabbitts TH. Activation of the T-cell oncogene LMO2 after gene therapy for X-linked severe combined immunodeficiency. N Engl J Med. 2004;350:913-22.
158. Gluckman E, Rocha V, Arcese W, Michel G, Sanz G, Chan K, et al. Factors associated with outcomes of unrelated cord blood transplant: guidelines for donor choice. Exp Hematol. 2004;32(4):397-407.
159. Migliaccio AR, Adamson JW, Stevens CE, Dobrila NL, Carrier CM, Rubinstein P. Cell dose and speed of engraftment in placental/umbilical cord blood transplantation: graft progenitor cell content is a better predictor than nucleated cell quantity. Blood. 2000;96(8):2717-22.
160. Rice A, Flemming C, Case J, Stevenson J, Gaudry L, Vowels M. Comparative study of the *in vitro* behavior of cord blood subpopulations after short-term cytokine exposure. Bone Marrow Transplant. 1999;23(3):211-20.
161. Meagher RC, Salvado AJ, Wright DG. An analysis of the multilineage production of human hematopoietic progenitors in long-term bone marrow culture: evidence that reactive oxygen intermediates derived from mature phagocytic cells have a role in limiting progenitor cell self-renewal. Blood. 1988;72(1):273-81.
162. Caldwell J, Palsson BO, Locey B, Emerson SG. Culture perfusion schedules influence the metabolic activity and granulocyte-macrophage colony-stimulating factor production rates of human bone marrow stromal cells. J Cell Physiol. 1991;147(2):344-53.
163. Schwartz RM, Palsson BO, Emerson SG. Rapid medium perfusion rate significantly increases the productivity and longevity of human bone marrow cultures. Proc. Natl. Acad. Sci. U.S.A. 1991 Aug 1;88(15):6760-6764.
164. Dao MA, Hashino K, Kato I, Nolta JA. Adhesion to fibronectin maintains regenerative capacity during *ex vivo* culture and transduction of human hematopoietic stem and progenitor cells. Blood. 1998;92(12):4612-21.
165. Bhatia M, Wang JC, Kapp U, Bonnet D, Dick JE. Purification of primitive human hematopoietic cells capable of repopulating immune-deficient mice. Proc Natl Acad Sci U S A. 1997;94:5320-5.
166. Conneally E, Cashman J, Petzer A, Eaves C. Expansion *in vitro* of transplantable human cord blood stem cells demonstrated using a quantitative assay of their lympho-myeloid repopulating activity in nonobese diabetic-scid/scid mice. Proc Natl Acad Sci U S A. 1997;94(18):9836-41.
167. Piacibello W, Sanavio F, Severino A, Danè A, Gammaitoni L, Fagioli F, et al. Engraftment in nonobese diabetic severe combined immunodeficient mice of human CD34(+) cord blood cells after *ex vivo* expansion: evidence for the amplification and self-renewal of repopulating stem cells. Blood. 1999;93(11):3736-49.
168. Shpall EJ, Quinones R, Giller R, Zeng C, Baron AE, Jones RB, et al. Transplantation of *ex vivo* expanded cord blood. Biol Blood Marrow Transplant. 2002;8(7):368-76.
169. Pecora AL, Stiff P, Jennis A, Goldberg S, Rosenbluth R, Price P, et al. Prompt and durable engraftment in two older adult patients with high risk chronic myelogenous leukemia (CML) using *ex vivo* expanded and unmanipulated unrelated umbilical cord blood. Bone Marrow Transplant. 2000;25(7):797-9.
170. Koller MR, Manchel I, Maher RJ, Goltry KL, Armstrong RD, Smith AK. Clinical-scale human umbilical cord blood cell expansion in a novel automated perfusion culture system. Bone Marrow Transplant. 1998;21(7):653-63.
171. Jaroscak J, Goltry K, Smith A, Waters-Pick B, Martin PL, Driscoll TA, et al. Augmentation of umbilical cord blood (UCB) transplantation with *ex vivo*-expanded UCB cells: results of a phase 1 trial using the AastromReplicell System. Blood. 2003;101(12):5061-7.

172. Hofmeister CC, Zhang J, Knight KL, Le P, Stiff PJ. *Ex vivo* expansion of umbilical cord blood stem cells for transplantation: growing knowledge from the hematopoietic niche. Bone Marrow Transplant. 2007;39(1):11-23.
173. Glimm H, Oh IH, Eaves CJ. Human hematopoietic stem cells stimulated to proliferate *in vitro* lose engraftment potential during their S/G(2)/M transit and do not reenter G(0). Blood. 2000;96(13):4185-93.
174. Guenechea G, Segovia JC, Albella B, Lamana M, Ramírez M, Regidor C, et al. Delayed engraftment of nonobese diabetic/severe combined immunodeficient mice transplanted with *ex vivo*-expanded human CD34(+) cord blood cells. Blood. 1999;93(3):1097-105.
175. Szilvassy SJ, Meyerrose TE, Ragland PL, Grimes B. Homing and engraftment defects in *ex vivo* expanded murine hematopoietic cells are associated with downregulation of beta1 integrin. Exp Hematol. 2001;29(12):1494-502.
176. Ramírez M, Segovia JC, Benet I, Arbona C, Güenechea G, Blaya C, et al. *Ex vivo* expansion of umbilical cord blood (UCB) CD34(+) cells alters the expression and function of alpha 4 beta 1 and alpha 5 beta 1 integrins. Br. J. Haematol. 2001 Oct;115(1):213-221.
177. Giet O, Huygen S, Beguin Y, Gothot A. Cell cycle activation of hematopoietic progenitor cells increases very late antigen-5-mediated adhesion to fibronectin. Exp Hematol. 2001;29(4):515-24.
178. Peters SO, Kittler EL, Ramshaw HS, Quesenberry PJ. Murine marrow cells expanded in culture with IL-3, IL-6, IL-11, and SCF acquire an engraftment defect in normal hosts. Exp Hematol. 1995;23(5):461-9.
179. Peters SO, Kittler EL, Ramshaw HS, Quesenberry PJ. *Ex vivo* expansion of murine marrow cells with interleukin-3 (IL-3), IL-6, IL-11, and stem cell factor leads to impaired engraftment in irradiated hosts. Blood. 1996;87(1):30-7.
180. Liu B, Buckley SM, Lewis ID, Goldman AI, Wagner JE, van der Loo JCM. Homing defect of cultured human hematopoietic cells in the NOD/SCID mouse is mediated by Fas/CD95. Exp Hematol. 2003;31(9):824-32.
181. Kim H, Whartenby KA, Georgantas RW3, Wingard J, Civin CI. Human CD34+ hematopoietic stem/progenitor cells express high levels of FLIP and are resistant to Fas-mediated apoptosis. Stem Cells. 2002;20(2):174-82.
182. Domen J, Cheshier SH, Weissman IL. The role of apoptosis in the regulation of hematopoietic stem cells: Overexpression of Bcl-2 increases both their number and repopulation potential. J Exp Med. 2000;191(2):253-64.
183. Wang LS, Liu HJ, Xia ZB, Broxmeyer HE, Lu L. Expression and activation of caspase-3/CPP32 in CD34(+) cord blood cells is linked to apoptosis after growth factor withdrawal. Exp Hematol. 2000;28(8):907-15.

CHAPTER 2

Materials and Methods

Animals

Male BALB/c and female α-thalassemic BALB/c TH/- mice were bred and housed under FELASA+ conditions (Federation of European Laboratory Animal Science Associations. Mice in the animal facilities were tested and found negative for: Parvovirus, Minute virus of mice, mouse hepatitis virus, pneumonia virus of mice, Sendai virus, Theiler's encephalomyelitis virus, Reo 3, Rota virus, Anthropods, gastrointestinal helminthes, *Giardia* spp, *Entamoeba muris*, *Tritrichomonas* spp, *Eimeria* spp, *Spironucleus* spp.. For FELASA+ quality, mice are also screened and negative for *Staphylococcus aureus*, *Pseudomonas aeruginosa*, Proteus, *Klebsiella pneumoniae*, *Klebsiella oxytoca* (1). Mice were housed in individually ventilated cages in the Erasmus MC animal facility (Erasmus MC, Rotterdam, The Netherlands). Transplanted animals were housed in stainless steel (Beyer & Eggelaar metaalindustrie B.V., Utrecht, The Netherlands) or plastic isolator units (Harlan Isotec, Bicester, UK), according to the requirements of The Netherlands Commission on Genetic Modification (COGEM). Experiments were evaluated on ethical grounds according to Dutch Law on Animal experiments (DEC). Male wild type BALB/c mice were used as donors and α-thalassemic female mice (8-12 weeks old) as recipients. The α-thalassemic phenotype, with smaller red cell volume, allows measuring of reconstitution of the bone marrow with normal sized red cells in peripheral blood by flow cytometry.

Total Body Irradiation

Prior to transplantation, recipient mice received a sub-lethal dose of 6 Gy of total body irradiation, (^{137}Cs source, Gammacell, Atomic Energy of Canada, Ottawa, Canada) with dose rate between 0.8136 Gy/min and 0.7619 Gy/min.

Transduction

Mouse bone marrow (BM) cells were isolated by density gradient centrifugation and transduced with retrovirus vectors packaged in amphotropic Phoenix cells with either pLZRS-IRES-EGFP or pLZRS-wtStat5-IRES-EGFP (Kind gift of A. Miyajima). In short, a stable packing cell line was created for both vectors, generating a supernatant with an approximate titer of 10^6 transducing units/ml. Supernatant containing retrovirus was harvested overnight in enriched Dulbecco's medium[5], (2,3) 0.45 µm filtered to

5 Stem cell medium: Dulbecco's modified Eagle's medium (Gibco, Life Technologies Inc., Paisley, Scotland) supplemented with $2.8x10^{-4}$ M L-alanine, $3.3x10^{-4}$ M L-aspargine, $2.3x10^{-4}$ M L-aspartic acid, $5.8x10^{-4}$ M L-cysteine, $5.1x10^{-4}$ M L-glutamic acid, 3.5 $x10^{-4}$ M L-proline, $1.5x10^{-5}$ M cholesterol, 4 µM cytidine, 4 µM adenosine, 4 µM uridine, 3.5 µM guanosine, 4.4 µM 2'-deoxycytidine, 4 µM 2'deoxyadenosine, 4 µM thymidine, 3.7 µM 2'deoxyguanosine, $1.2x10^{-7}$ M d-biotin, 1% fraction V BSA, (all Sigma-Aldrich, Zwijndrecht, The Netherlands), $1.8x10^{-8}$ M vitamin B12, 10^{-3} M sodiumpyruvate, $1.9x10^{-2}$ M glucose (Merck), $4.4x10^{-2}$ M $NaHCO_3$, 10^{-4} M β-mercapto-ethanol, 10^{-7} M Na_2SeO_3, $1.5x10^{-5}$ M linoleic acid (Merck, Darmstadt, Germany), 0.1 g/l penicillin (Yamanouchi, Leiderdorp, The Netherlands), 10^5 IE/l streptomycin (Fisiopharma, Milano, Italy), $2x10^{-6}$ M iron saturated transferrin (Serologicals Proteins Inc. Kankakee, IL, USA), at an osmolarity of 300 mOsm/l.

remove cell debris and diluted 1:1 with fresh enriched Dulbecco's medium. Male wild type donor mouse bone marrow was isolated and enriched for mononuclear cells by percoll (Amersham Bioscience, Uppsala, Sweden) density gradient centrifugation. The isolated cells were cultured for 2 days in enriched Dulbecco's medium supplemented with 100 ng/ml murine SCF (R&D systems), 10 ng/ml murine TPO (kindly provided by Genentech, South San Francisco, CA, USA) and 50 ng/ml Flt3L (kindly provided by Amgen, Thousand Oaks, CA, USA), after which the cells were plated at a density of $5x10^5$ cells /ml on retronectin (CH-296, 48 μg/ml, Takara Shuzo, Otsu, Japan) coated plates (4) pre-incubated with retrovirus, the medium was replaced with supernatant containing retrovirus and 100 ng/ml murine SCF, 50 ng/ml murine TPO and 10 ng/ml Flt3L, on which the cells were cultured 2 more days. A ratio of 1 transducing particle per target cell was used (MOI=1). On day 4, EGFP expression was measured by flowcytometry (FACS Calibur, Becton Dickinson) and the cells were prepared for transplantation.

Transplantation

Transduced mouse bone marrow cells, with an average transduction efficiency of ~20% as measured by EGFP expression, were injected intravenously in graded cell numbers (10^4, $3x10^4$, 10^5, $3x10^5$ and 10^6 cells/mouse), dissolved in 250 μl HBSS, in 6 Gy irradiated (^{137}Cs source, Gammacell, Atomic Energy of Canada, Ottawa, Canada) female α-thalassemic BALB/c mice. The mice were observed daily and bled monthly to determine chimerism and transgene expression in the peripheral blood. When showing signs of disease (no eating, drinking, apathic behavior, fuzzy coat, kyphosis), the mice were anesthetized by isoflurane (Pharmachemie, Haarlem, The Netherlands) inhalation, peripheral blood was drawn, the mice were killed and cells were isolated from femoral bone marrow and spleen for cryopreservation and DNA (Nucleospin Tissue system, Machery-Nagel, Düren, Germany) and RNA isolation (RNABee, Tel-Test Inc., Friendswood, Texas, USA). An obduction was performed and kidney, liver, spleen, heart, lung, brain, stomach, coecum, ileum, colon and (enlarged) lymph nodes were fixed in formalin to analyze tissue morphology. 4 μm sections were made from paraffin embedded tissue, that were stained with hematoxylin and eosin according to standard protocols. Peripheral blood smears were May Grünwald Giemsa (Merck, Darmstadt, Germany) stained.

Testing for Recombinant Retrovirus

To test for the presence of replication competent retrovirus (RCR) we performed *in vitro* and *in vivo* tests on the virus producing cell lines and on the samples that were obtained from diseased mice.

In vitro replication competent retrovirus (RCR) analysis

Supernatant containing retrovirus particles was incubated with HeLa or Rat-2 cells in the presence of 4 μg/ml polybrene (hexadimethrin bromide, Sigma-Aldrich, Zwijndrecht, The Netherlands). After incubation, the supernatant containing retrovirus particles was replaced with enriched DMEM[6] (Gibco) +10% FCS and the cells were cultured for two further weeks after which supernatant was collected and used to assay the capability of EGFP transfer. This was done by using the supernatant of the transduced cells for transduction of fresh HeLa or Rat-2 cells in the presence of 4 μg/ml polybrene. At confluency the cells were analyzed by flow cytometry for EGFP expression.

In vivo RCR analysis

Spleen cells were isolated from diseased animals. A fraction ($3x10^6$ – 10^7) of these cells was irradiated with 20 Gy (^{137}Cs source), which results in repopulation defective cells. Any recombinant virus particles present will not be affected by this radiation dose and will still be able to infect the recipients, resulting in leukemia. The cells were transplanted into 6 Gy irradiated BALB/c recipients. We observed no recurrence of the disease in these animals until 6 months after transplantation, suggesting that the diseased animals did not carry a replication competent retrovirus.

RT-qPCR RCR analysis

Reverse transcriptase activity was determined in protein isolates of spleen cells of mice presented with hematopoietic malignancy as described (5). Mouse spleen cells were pelleted by centrifugation (1600 rpm, 5 minutes) and resuspended in 200 μl protease inhibitor cocktail (Boehringer). The protein content was determined and samples were stored at –20°C. 10 μg protein was incubated with 6 ng BMV template RNA (Promega), 10 nmol dNTP, 200 nmol MgCl2, 1.25 U AmpliTaq Gold (Applied Biosystems), 4 U RNAseOUT recombinant ribonuclease inhibitor (Invitrogen) , 15 pmol of each primer and 5 pmol probe (Eurogentec) and 150 ng activated calf thymus DNA (Sigma).

The product of reverse transcription of BMV RNA was amplified in an ABI Prism 7900 Sequence Detection System (Applied Biosystems) using real-time PCR (BMV forward primer BMV reverse primer and a 5'FAM (6-carboxyfluorescein) and 3'TAMRA (6-carboxy-tetramethyl-rhodamine) labelled BMV probe. PCR conditions were 30 min-

6 Enriched DMEM: Dulbecco's modified Eagle's medium (Gibco, Life Technologies Inc., Paisley, Scotland) supplemented with $2.8x10^{-4}$ M L-alanine, $3.3x10^{-4}$ M L-aspargine, $2.3x10^{-4}$ M L-aspartic acid, $5.8x10^{-4}$ M L-cysteine, $5.1x10^{-4}$ M L-glutamic acid, 3.5 $x10^{-4}$ M L-proline, $1.2x10^{-7}$ M d-biotin, (all Sigma-Aldrich, Zwijndrecht, The Netherlands), $1.8x10^{-8}$ M vitamin B12, 10^{-3} M sodiumpyruvate, $1.9x10^{-2}$ M glucose, $4.4x10^{-2}$ M $NaHCO_3$, (Merck, Darmstadt, Germany), 0.1 g/l penicillin (Yamanouchi, Leiderdorp, The Netherlands), 10^5 IE/l streptomycin (Fisiopharma, Milano, Italy), at an osmolarity of 300 mOsm/l.

utes at 48°C followed by 40 cycles of 1 minute 94°C, 30 seconds 60°C and 30 seconds of 72°C and a final 10 minutes of 72°C.

Protein samples from AM12 SF91 EGFP retrovirus producer cell lines were used as a positive control. Superscript II reverse transcriptase was used to calibrate the reaction in a range of 10^{-1} to 10^{-8} units reverse transcriptase. All tested samples had less than $4x10^{-6}$ (range $6.8x10^{-9}$ to $3.2x10^{-6}$) units reverse transcriptase activity, compared to $5x10^{-5}$ units for the virus producer cell line.

Hematological and Phenotypical Analysis

The transplanted mice were bled monthly by retro-orbital vein puncture (200 µl), and their peripheral blood values as well as the red blood cell size, EGFP expression in erythrocytes, leukocytes and thrombocytes were measured by flow cytometry. When the mice were moribund, peripheral blood was obtained from anesthetized diseased animals by retro-orbital puncture, after which they were euthanised by cervical dislocation. Peripheral blood values were measured on a Scil-Vet animal blood counter (ABX diagnostics, Montpellier, France). Femurs and spleen cells were isolated. Also spleen, kidney and visible lymph nodes were isolated for histology. Femurs were flushed with HBSS (Sigma) to obtain BM cells. Erythrocytes in peripheral blood and spleen were lysed in 155 mM NH4Cl, 1 mM KHCO3 (Merck), 10 µM EDTA (Sigma) for 10 minutes on ice. BM and spleen cells were strained over a nylon sieve in 5 ml HBSS to obtain single cell suspensions, 10^6 cells from peripheral blood, bone marrow and spleen were incubated for 30 minutes on ice with anti-CD4PE and anti-CD8-APC or antiB220-APC (BD Pharmingen, San Jose, CA, USA) and anti-CD11b PE (Immunotech, Marseille, France) monoclonal antibodies in HBSS with 0.5% wt/vol BSA(Sigma-Aldrich) and 0.05% wt/vol NaN3 (Merck) and 2% vol/vol mouse serum, and washed twice. EGFP and marker expression were measured in the presence of 1.5 nM propidium iodide (Sigma), which allows exclusion of dead cells in the analysis by flow cytometry.

Hematopoietic clonogenic progenitor assays

Bone marrow cells were cultured in BFU-E and CFU-C optimized semisolid medium. In short, $5x10^4$ cells were plated in the appropriate semisolid medium and cultured at 37°C in a humidified atmosphere containing 10% CO_2. For BFU-E erythropoietin (4U/ml Behringwerke AG, Marburg, Germany) and mouse SCF (100ng/ml R&D Systems, Minneapolis) were added. CFU-C were cultured in the presence of mouse IL-3 (30 ng/ml), mouse SCF (100ng/ml, both R&D Systems, Minneapolis, USA) and a 300x dilution ConA adsorbed fraction of pregnant mouse uterus extract (CSF). 10 days after plating the colonies were counted and expressed as colonies per 10^5 cells plated as described before (2,6).

Retransplantations of leukemic cells

To further characterize the malignancies and dilute contaminating non-malignant clones, $3x10^6$ or 10^7 spleen cells or 10^6 bone marrow cells were transplanted in 3 Gy irradiated BALB/c mice. When the mice were moribund the mice they were killed and analyzed as described.

Integration site analysis

Retrovirus integration sites were analyzed using 500 ng DNA from peripheral blood, bone marrow or spleen by LAM-PCR as described before (7). In short, DNA underwent 100 cycles linear amplification using a biotinylated primer (LTRI, Eurogentec, Seraing, Belgium). The products were captured with streptavidin coated magnetic beads (Kilobase binder kit, Dynal, Oslo, Norway). A second strand was generated using hexanucleotides and Klenow enzyme (Roche, Mannheim, Germany) at 37°C for 60 minutes. The double stranded product was cleaved for 1 hour at 65°C with Tsp509I (New England Biolabs, Ipswich, MA, USA) and a complementary oligonucleotide cassette (LC BOX1 and LC BOX2, below) was ligated (Fastlink ligase, Epicentre technologies, Madison, WI, USA). Nested PCR was performed to obtain the retrovirus integration sites (primer pairs: LTRI1 and LC1 and LTRIII and LCII).

Nucleotide sequencing

Complex LAM-PCR products were ligated into TOPO-tk plasmids (Invitrogen, Breda, The Netherlands), which were transformed into TOP10 E. coli. The bacteria were grown overnight on 1.5% wt/vol agar Luria-Bertani plates and single colonies were collected and placed in 0.75% wt/vol agar LB. From these samples the LAM- PCR inserts were sequenced (GATC, Konstanz, Germany) using the M13 sequences in the TOPO-tk plasmid. In the resulting sequences, the presence of the primer binding sites of the virus LTR and the linker cassette was verified. The genomic sequence between LTR and linker cassette was isolated and masked for repeat sequences (using rodent or mammalian libraries where applicable). All resulting DNA sequences larger than 25 bp were aligned to the mouse genome (Ensembl NCBI v36 mouse genome assembly). For each successful alignment, the distance between the virus integration and the nearest surrounding genes was calculated (using the Ensembl NCBI v36 genomic locations)

Copy number Determination

The number of virus integrations was determined in DNA obtained from peripheral blood from mice with high (>90%) numbers of cells with EGFP expression or cells without EGFP expression, but with predominant expression of one of the lineage markers (CD4, CD8, CD11b or B220). *EGFP* and *IL2* were amplified by PCR with SYBR Green

mastermix (Applied Biosystems, Warrington, UK) using EGFP primers and IL2 primers (42,43). Delta Ct values were calculated to correct for loading differences.

Y Chromosome PCR

A Y chromosome PCR on *YMT* was used to determine donor or recipient origin of the leukemia samples. Using 1.25 U Qiagen Taq-polymerase, 20μM dNTP, *MyoD* (1 μM) and *YMT* (2 μM) primers, 100 ng DNA from the sample was denatured (5 minutes at 94°C), after which the sample was amplified in 40 cycles (30 seconds 94°C, 30 second 58°C, 30 seconds 72°C). After a final elongation (5 minutes 72°C). The resulting product was visualized on a 1% agarose gel, showing a 344 bp *YMT* product and a 226 *MyoD* internal control.

Gene expression analysis of mouse hematopoietic stem cells

For each experimental group bone marrow cells from tibia and femurs of 30 C57Bl6 mice or 9 C57Bl6/ CD45.1 mice were isolated and strained over nylon. Lineage$^-$ Sca- 1$^+$c-kit$^+$ cells, that are similar to the CD34$^+$CD38$^-$ cell population in humans, were isolated by first depleting the BM cells for lineage committed cells (Lineage depletion kit, Miltenyi Biotec GmbH, Bergisch Gladbach, German) using the Automacs cell separation system (Miltenyi Biotec), subsequently, the cells were positively selected for CD117 expression (CD117 selection kit, Miltenyi Biotec). From this fraction the cells expressing both CD117 and Sca-1 (BD Biosciences, Pharmingen) were sorted (FACS DiVa, Becton Dickinson, San Jose, CA, USA). Reanalysis of the sorted cells confirmed >95% purity of the sorted cells.

RNA of half of the sorted cells was isolated directly after sorting, using Qiagen's RNEasy system and half after 2 days of culture in enriched Dulbecco's medium with 100 ng/ml murine SCF, 10 ng/ml murine TPO and 50 ng/ml Flt3L at 37°C and 10% CO2. CDNA was generated according to the manufacturer's recommendations (Qiagen Quantiscript), amplified (MEGAscript T7 kit, Ambion, Huntingdon, UK) and labeled (Affymetrix 2-step synthesis kit) as described before. Labeled cDNA was hybridized to Mouse Genome U430 2.0 arrays (Affymetrix) and analysed on an Genechip G7 reader(Affymetrix). Expression of genes neighbouring a previously identified virus integrations site were located and their expression was determined.

The 45101 probesets on the Mouse Genome U430 2.0 array were sorted based on the expression values, 10 bins of equal size were generated and the presence of the virus integration sites (represented by the affyIDs) was determined (Perl 5.8.2) Using the Perl scripts, the intersections of the retrieved virus integration site dataset with the RTCGD database (8) (version mm8, mouse, February 2007[7]), the Sanger Institute Cancer Gene Census database (Accessed November 2, 2006[8]) and the Jackson Labs MGI Mammalian

7 http://rtcgd.abcc.ncifcrf.gov/

8 http://www.sanger.ac.uk/genetics/CGP/Census/

phenotype database (Accessed April 19, 2006[9], using Hematopoietic system phenotype, Immunological phenotype and tumor phenotype caused by targetted mutations only). Lastly, the Gene Symbols were entered into Ingenuity Pathway Analysis[10] to retrieve the Functions and Diseases and Canonical Pathway annotations. Statistics were calculated using R (version 2.4.1- 2.10[11])

List of primers

Primer	Sequence (5' to 3')
LTRI	Biotin-AGCTGTTCCATCTGTTCCTGACCTT
LTRII	GACCTTGATCTGAACTTCTC
LTRIII	Biotin-TTCCATGCCTTGCAAAATGGC
M13 forward	GTAAAACGACGGCCAG
M13 reverse	CAGGAAACAGCTATGAC
EGFP forward	TCCTTGAAGAAGATGGTGCG
EGFP reverse	AAGTTCATCTGCACCACCG
IL2 forward	CTAGGCCACAGAATTGAAAGATCT
IL2 reverse	GTAGGTGGAAATTCTAGCATCATCC
LC I	GACCCGGGAGATCTGAATTC
LC II	GATCTGAATTCAGTGGCACAG
LC BOX1	GACCCGGGAGATCTGAATTCAGTGGCACAGCAGTTA GG
LC BOX2	AATTCCTAACTGCTGTGCCACTGAATTCAGATC
YMT1	CTGGAGCTCTACAGTGATGA
YMT2	CAGTTACCAATCAACACATCAC
MyoD1	TTACGTCCATCGTGGACAGC
MyoD2	TGGGCTGGGTGTTAGTCTTA
BMV Forward	TCTTGAGTTAGACCACAACGTTCCT
BMV Reverse	TGCGCTTGTCTCTGTGTGAGA
BMV Probe	(5'FAM)TCTGCTCGAGGAGAGCCCTGTTCC(3' TAMRA)

References

1. Nicklas, W. et al. Recommendations for the health monitoring of rodent and rabbit colonies in breeding and experimental units. Lab Anim 36, 20-42 (2002).
2. Merchav, S. & Wagemaker, G. Detection of murine bone marrow granulocyte/macrophage progenitor cells (GM-CFU) in serum-free cultures stimulated with purified M-CSF or GM-CSF. Int. J. Cell Cloning 2, 356-367 (1984).
3. Wagemaker, G. & Peters, M.F. Effects of human leukocyte conditioned medium on mouse haemopoietic progenitor cells. Cell Tissue Kinet 11, 45-56 (1978).

9 http://www.informatics.jax.org/searches/MP_form.shtml

10 www.ingenuity.com

11 http://www.r-project.org/

4. Moritz, T. et al. Fibronectin improves transduction of reconstituting hematopoietic stem cells by retroviral vectors: evidence of direct viral binding to chymotryptic carboxy-terminal fragments. Blood 88, 855-62 (1996).
5. Lovatt, A. et al. High throughput detection of retrovirus-associated reverse transcriptase using an improved fluorescent product enhanced reverse transcriptase assay and its comparison to conventional detection methods. J. Virol. Methods 82, 185-200 (1999).
6. Wagemaker, G. & Visser, T.P. Erythropoietin-independent regeneration of erythroid progenitor cells following multiple injections of hydroxyurea. Cell Tissue Kinet 13, 505-517 (1980).
7. Schmidt, M. et al. Detection and direct genomic sequencing of multiple rare unknown flanking DNA in highly complex samples. Hum Gene Ther 12, 743-9 (2001).
8. Akagi, K., Suzuki, T., Stephens, R.M., Jenkins, N.A. & Copeland, N.G. RTCGD: retroviral tagged cancer gene database. Nucleic Acids Res 32, D523-7 (2004).

CHAPTER 3

Retroviral vector integrations relate to hematopoietic stem cell gene expression patterns

Adapted from: Martijn H. Brugman[1,3], Manfred Schmidt[4], Carla Oerlemans-Bergs[1], Sigrid Swagemakers[2], Dick de Ridder[5], Christof von Kalle[4], Monique Verstegen[1] and Gerard Wagemaker[1]. Retroviral vector integrations relate to hematopoietic stem cell gene expression patterns. *Submitted for publication.*

[1] Department of Hematology, [2]Department of Bioinformatics, Erasmus MC Rotterdam, The Netherlands.
[3] Department of Experimental Hematology, Hannover Medical School, Hannover, Germany
[4] NCT and Translational Oncology department, DKFZ, Heidelberg, Germany.
[5] Information & Communication Theory Group, Delft University of Technology, Delft, The Netherlands.

ABSTRACT

Integrating gamma-retroviral vectors are effective tools for gene transfer into both mouse and human hematopoietic stem cells and have been used to perform clinical gene therapy. These clinical trials, although highly successful in efficiency, showed severe adverse effects of virus insertion in a limited number of cases, which has been explained by insertional mutagenesis. In this study, the insertion pattern of gamma-retroviral vectors in mouse hematopoietic stem cells was investigated. The gamma-retroviral vector showed a preference for insertion in (highly) expressed genes near transcription start sites. As much as 83% of the insertions occurred near expressed genes in the target cells and a highly significant proportion of these genes were proto-oncogenes. The vectors did not only insert near a single highly expressed gene, but rather showed a preference for loci where all genes were highly expressed. The gamma-retroviral integrations in hematopoietic cells reflect the high expression of genes with general cell functions as well as a pattern specific for the hematopoietic cells used as targets for the transduction, in that genes with hematopoietic and immune related function are frequent targets for integration. Common insertions were identified, but did not follow the general pattern of strong preference for highly expressed genes but rather occurred near genes with widely varying expressions, suggesting that other mechanisms such as selective advantage conferred to stem cells or their progeny might be involved.

Cell function and preference for highly expressed genes was conserved between mouse and human, implying that properly conducted mouse studies are predictive for safety in humans. This study predicts that transduction of human hematopoietic cells with gamma-retroviral vectors would result in an overrepresentation of hits near proto-oncogenes, in cells with hematopoietic and immune functionality. In addition, since gene expression in the target cell seems to govern the distribution of insertions, a significant overlap between individual, independent clinical trials using the gamma-retroviral backbone is expected and demonstrated.

INTRODUCTION

Integrating viral vectors provide long-term transgene expression required for gene therapy of inherited disorders by a single therapeutic intervention. As has been amply demonstrated in initial clinical trials (1-3) viral integration of a therapeutic gene may result in insertional oncogenesis (4) when situated in a region near proto-oncogenes or tumor-suppressor genes. Multiple factors, such as genetic background of cells (5), culture conditions, vector type and dose, location of insertion and level of expression of the transgene or surrounding genes (6) , determined by the virus LTR or endogenous promoter, are involved in clonal selection after transgene insertion into somatic cells, which may either result in non-malignant clonal dominance or in malignant transformation after secondary mutagenic events. In a transplant setting also cell dose and stressors such as the conditioning regimen are selection factors (7). Insertional mutagenesis following therapeutic transgenesis has been described extensively in preclinical mouse models (8), rhesus monkey models (9) and clinical trials (2,3). Oncogenesis by retroviruses has been described in detail (10,11) and many studies use this property to identify potential proto-oncogenes (12-16). Such studies provide a valuable source of genes correlated with leukemia, but since these datasets are compiled from studies employing replicating viruses, resulting in multiple insertions in leukemic clones, identification of the primary leukemic event has met with difficulty and the attention in these studies has focused on clusters of genes (12,17) to identify proto-oncogenes rather than on single insertions.

Reports on the safety of retrovirus insertion stressed the importance of testing safety of the retroviral vector dose (18), as well as the therapeutic transgene (19). Therapeutic gene transfer uses replication deficient vectors with relatively low infectious particle doses that preferably result in one vector copy per cell. By employing such procedures, multiple retroviral integrations will most probably not occur.

The integration profile of gamma-retroviral vectors has been studied in preclinical models (6,18,20,21) in human cord blood cells (22) and in sample obtained from gene therapy trials (23-25) and show remarkable similarities. Gamma-retroviral vectors predominantly insert near transcription start sites (TSS) and have a preference for expressed genes. Open chromatin configuration is suggested to be important for guiding retrovirus integrations (26) and open chromatin is related to expression of the genes in that region of the chromosome.

Replication deficient gene therapy vectors have been studied in several safety studies where target cells ranged from cell lines (27), to umbilical cord blood cells (28,29) and mouse bone marrow cells (4,18). Also several *in vitro* models have been proposed (30,31).

Lineage-depleted ($Lin^{-/-}$)/Sca-1^{+}/c-Kit^{+} (LSK) cells are a cell population enriched in hematopoietic cells, known to be able to repopulate an irradiated recipient (32). In the present study we identified transduced cells by their genome-vector boundary sequences

and analyze the integration pattern of a gamma-retroviral vector in long-term murine bone marrow repopulating cells, transduced using a 4-day stimulation with murine SCF, murine TPO and Flt3L to facilitate integration of the retrovirus as described before (33). These cells were subsequently transplanted in irradiated recipients. The study was originally set up to identify the effect of expression of the signal transduction protein wtStat5 in murine hematopoietic cells, but was adapted to focus on the insertion profiles in the transplanted mice. Integration sites, identified by LAM-PCR and sequencing, were related to gene expression levels of highly purified mouse stem cells (LSK cells). We identified clusters of common integration and analyzed the functions of the genes near insertion sites to document the insertion pattern of the gamma-retroviral backbone.

In the current study, we were interested to find the relation between integration and gene expression, the occurrence of common insertion sites and the functions of the genes that were hit by the integrating virus. We also sought to assess whether mouse integration studies are predictive for human hematopoietic stem cell transduction using the overlap with clinical trial datasets using similar vector backbones as determining factor.

MATERIALS AND METHODS

Animals and cells

Specific pathogen free (SPF), 8-12 weeks old (C57BlxCBA F1 hybrid (BCBA)) BALB/c mice, were bred and housed under SPF conditions at the Experimental Animal Facility of Erasmus MC (Rotterdam, Netherlands) in compliance with Dutch ethics and animal welfare regulations. Murine BM cells were isolated from male donors by flushing the femurs with Hank's Balanced Salt Solution (HBSS, Gibco, Breda, The Netherlands). Mononucleated cells were selected using Percoll gradient centrifugation (Amersham Biosciences) and subsequently pre-stimulated in a serum-free enriched version of Dulbecco's modified Eagle's medium (DMEM; Gibco, Gaithersburg, MD) (34-36) in the presence of mouse stem cell factor (mSCF; 100 ng/ml), fms-like tyrosine kinase-ligand (Flt3-L; 50 ng/ml) and mouse thrombopoietin (mTPO; 10 ng/ml), for 2 days (37). The pre-stimulation was followed by 2 days of supernatant infection on Falcon 1008 dishes (35 mm) that were coated with the human recombinant retronectin fragment CH296 (10 μg/cm^2, Takara Shuzo, Otsu, Japan) as described (38) and preloaded with filtered (0.45 μm) virus. EGFP expression was analyzed by flow cytometry (FACS Calibur, Becton Dickinson, San Jose, CA)(33). Transduced cells (using pLZRS-IRES-EGFP and pLZRS-stStat5-IRES-EGFP retroviral vectors) were subsequently transplanted intravenously into irradiated (6 Gy) female (α-thalassemic) BALB/c mice. The number of transplanted cells was 10^4, $3x10^4$, 10^5, $3x10^5$, and 10^6 per recipient group. As a control, irradiated mice were injected with BM cells treated similarly with non-virus containing medium.

Monthly blood samples were analyzed for the presence of EGFP in erythrocytes, leukocytes and thrombocytes by flow cytometry, and total blood counts were determined in EDTA tubes for small volumes (Becton Dickinson, San Jose, CA) using a blood counter (ABC-VET, ABX Diagnostics, Montpellier, France). At a time interval of 4 months after transplantation, BM of the mice showing the highest EGFP levels (experiment 1 only) was collected for marrow repopulating ability (MRA) analysis and long-term repopulating ability (LTRA) assays. Furthermore, integration analysis (LAM-PCR and sequencing) was performed to analyze the integration pattern of the retroviral vectors in the mouse genome and the outgrowth of transduced clones.

RCR analysis

The initial virus preparations were tested for the absence of replication competent retrovirus (RCR) according to Markowitz (39). To do so, NIH 3T3 cells transduced with the virus stocks used in the animal experiments were cultured in DMEM +10% FCS for 14 days, passaging cells every 3 days. Fourteen days after transduction supernatant of the transduced cells was collected, passed through a 0.45 mm filter and added to $5x10^5$ NIH 3T3 in the presence of 4 mg/ml polybrene for 24 hours, after which cells were cultured in DMEM + 10% FCS. When the cells reached confluence, they were trypsinized and analyzed for GFP expression by flow cytometry.

To show absence of RCR after in mice that presented with hematopoietic malignancies, spleen cells of three mice that were found with hematopoietic malignancies were 20 Gy irradiated and $3x10^6$ cells were transplanted into 4-5 recipients. No leukemias were found after transplantation and a one-year observation period in the transplanted mice.

Reverse transcriptase activity was determined in protein isolates of spleen cells of mice presented with hematopoietic malignancy as described (40). Mouse spleen cells were precipitated by centrifugation (1600 rpm, 5 minutes) and resuspended in 200 ml protease inhibitor cocktail (Boehringer). The protein content was determined and samples were stored at –20°C. 10 mg protein was incubated with 6 ng BMV template RNA (Promega), 10 nmol dNTP, 200 nmol $MgCl_2$, 1.25 U AmpliTaq Gold (Applied Biosystems), 4 U RNAseOUT recombinant ribonuclease inhibitor (Invitrogen) , 15 pmol of each primer and 5 pmol probe (Eurogentec, Seraing, Belgium) and 150 ng activated calf thymus DNA (Sigma).

The product of reverse transcription of BMV RNA was amplified in an ABI Prism 7900 Sequence Detection System (Applied Biosystems) using real-time PCR (forward primer (5')TCTTGAGTTAGACCACAACGTTCCT(3'), reverse primer (5')TGCGCTTGTCTCTGTGTGAGA(3') and a 5'FAM (6-carboxyfluorescein) and 3'TAMRA (6-carboxytetramethyl-rhodamine) labeled probe (5')TCTGCTCGAGGAGAGCCCTGTTCC(3'). PCR conditions were 30 minutes at 48°C followed by 40 cycles of 1 minute 94°C, 30 seconds 60°C and 30 seconds of 72°C and a final 10 minutes of 72°C.

Protein samples from AM12 SF91 EGFP retrovirus producer cell lines were used as a positive control. Superscript II reverse transcriptase was used to calibrate the reaction in a range of 10^{-1} to 10^{-8} units reverse transcriptase. All tested samples had less than $4x10^{-6}$ (range $6.8x10^{-9}$ to $3.2x10^{-6}$) units reverse transcriptase activity, compared to $5x10^{-5}$ units for the virus producer cell line.

Integration site analysis

We analyzed the genome – retrovirus boundaries as previously described using LAM-PCR (41). In short, 500 ng DNA was amplified linearly for 100 cycles using a biotinylated primer (Eurogentec) designed against the virus LTR, after which the products were captured on streptavidin coated beads (Kilobase binder kit, Dynal, Norway) and a second DNA strand was synthesized using Klenow polymerase and hexanucleotides (Roche). Subsequently, the products were digested using Tsp509I or HpyCH4 IV (New England Biolabs). A complementary linker cassette was ligated (Fastlink Ligase, Epicentre technologies) and nested PCR with primers designed against the virus LTR and the linker cassette was performed. Primers used included:
Tsp Linker cassette1: 5' – GACCCGGGAGATCTGAATTCAGTGGCACAGCAGTTAGG-3'
Tsp Linker cassette2: 5'- AATTCCTAACTGCTGTGCCACTGAATTCAGATC-3'
Hpy Linker cassette1: 5'-GACCCGGGAGATCTGAATTCAGTGGCACAGCAGTTACG-3'
Hpy Linker cassette2: 5'-CGCCTAACTGCTGTGCCACTGAATTCAGATC-3'
LTRI: 5'-[biotin]TGCTTACCACAGATATCCTG-3'
LTRII: 5'-GACCTTGATCTGAACTTCTC-3'
LTRIII: 5'-[biotin]TTCCATGCCTTGCAAAATGGC-3'
LC1:5'-GACCCGGGAGATCTGAATTC-3'
LC2:5'-GATCTGAATTCAGTGGCACAG-3'

Integration sites were retrieved from spleen and BM DNA samples obtained from 42 primary transplanted mice or secondary recipients, as well as 16 mice that did not show signs of disease after long-term follow up (range: 227 to 645 days). A LAM-PCR based method, modified to be less sensitive so that predominant clones in the polyclonal sample could be detected more accurately, was also employed to analyze predominant clones in the leukemia samples. In this analysis 10 ng input DNA was used and an additional product clean up using biotinylated primers and streptavidin coated magnetic beads (Dynal) was performed.

LAM-PCR products were cloned into pCR4-TOPO (Invitrogen) vectors, after which bacteria were transformed. Single colonies were plucked, grown and sequenced (GATC, Konstanz, Germany) using standard procedures. Only when both LTR and linker sequence were present in the sequence, the sequence was accepted. Sequences of good quality (sequences meeting Phred quality) were then trimmed for vector, linker and virus LTR sequences. The remaining genomic sequences were analyzed for repeat

sequences using RepeatMask which removes the repeat sequences and aligned using and TF Target Mapper[12] which performs BLAST alignment with the repeat trimmed sequences. Only sequences longer than 27 bp were aligned. Virus integration locations were then determined by the best matching unambiguous BLAST alignment. BLAST alignments that resulted in multiple alignments with equal E-values were discarded. Subsequently, the nearest gene to each virus integration or the genes within 100 kbp up- or downstream were identified (VIS gene) by comparing the virus insertion site to the transcription start sites in a database obtained from Ensembl Biomart[13].

LSK cell gene expression measurements

Mouse lineage$^-$ Sca-1$^+$ c-Kit$^+$ (LSK) cells were isolated from C57BL6 mouse BM using mouse lineage cell depletion (cocktail of CD5, CD45R (B220), CD11b, Gr-1, 7-4 and Ter-119 antibodies) and c-Kit selection (CD117 monoclonal antibody) kits according to the manufacturers protocol Miltenyi Biotec) and magnetic activated cell sorting (AutoMACS, Miltenyi Biotech) followed by flow cytometric cell selection of the CD117- FITC and Sca-1-PE double positive cells (FACS DiVa, BD Biosciences, San Jose, USA) to obtain LSK cells. Total RNA from the isolated LSK cells (>95% purity, n=2) was obtained using RNeasy columns according to the manufacturers protocol (Qiagen, Hilden, Germany). RNA quality was assessed using the Agilent 2100 BioAnalyzer (Agilent Technologies, Palo Alto, CA), directly after sorting (Day 0) or after maintaining the cells in serum-free liquid culture medium in the presence of mouse SCF, mouse TPO and Flt3L for 2 days (Day 2), similar to the transduction procedure as described. Linear amplification of the RNA was performed and the RNA was hybridized to Affymetrix 430 2.0 microarray. For this, 100 ng of total RNA from LSK cells was used in the GeneChip Eukaryotic Small Sample Target Labeling Assay Version II (Affymetrix) to generate biotinylated cRNA. Eleven µg of cRNA was fragmented for 35 min at 95˚C. 10 µg of fragmented cRNA was then hybridized to mouse 430 2.0 microarray (Affymetrix) for 16 hours at 45˚C followed by washing, staining and scanning at 570 nm, according to standard methods as described earlier. For each gene (described by Gene Symbol) present on the micro array, the highest expression was determined. The genes were then sorted based on expression level and divided into 10 expression level categories. In every expression level category, the number of VIS genes present was reported. To determine whether virus integrations occur more frequently near genes that are higher expressed in LSK cells, a Cochran – Armitage test for trend was performed.

A statistical approach was used to determine the frequency of integration sites within ±10 kbp, ±50 kbp, ±100 kbp of genes or transcription start sites. For all genes present

12 http://tftargetmapper.erasmusmc.nl

13 http://www.ensembl.org/biomart/martview

in Ensembl Biomart, the locations were determined and for each gene the presence of a virus integration within the given distance was determined. A χ^2 test for goodness of fit was performed to determine whether integrations occur more frequently within the given distance from genes. The calculations were then repeated for genes with expression exceeding threshold levels (300, 1000, 3000 and 10000 arbitrary expression units, respectively).

Annotation Analysis

Genes correlated with replication competent retrovirus induced leukemias were obtained from RTCGD (mm7) selecting all retrovirus models and all leukemia phenotypes, which resulted in 1130 genes. Co-occurrence of genes closest to retrovirus integrations and RTCGD genes was analyzed.

Using data from the Mouse genome information mammalian phenotype browser, which lists known phenotypes for existing knockout mice, the occurrence of genes within 100 kbp upstream or downstream of virus integration sites in 33 phenotype categories was determined. Ingenuity Pathway Analysis was used to identify gene networks.

Functional enrichment of lineage$^-$ Sca-1$^+$c-kit$^+$ immature hematopoietic cells

To gain insight into the level of enrichment of functional hematopoietic progenitors in the isolated LSK cells, spleen-colony forming units (CFU-S) and marrow repopulating ability (MRA) assays were performed. The LSK cells were 50-fold enriched in CFU-S (4.7 CFU-S per 3x10^4 cells in BM and 238.7 CFU-S per 3x10^4 cells in LSK) and a 104 and 118-fold enrichments were found in CFU-C and BFU-E numbers, respectively, in LSK cells compared to whole BM cells.

Micro array validation.

To validate the expressions found by the Affymetrix microarray analysis, we performed RQ-PCR using 13 (tested 20, 7 probe sets did not yield expression values) primer/probeset pairs of the 135 VIS genes, which correlated with the expression values found by microarray. The primer/probes sets were obtained from Applied Biosystem (Probes used were: *Itm2b, 1300001I01Rik, AI462493, Ly86, AQ050020, 3110045G13Rik, Evi1, 2300003P22Rik, 6620401K05Rik, 3110023E09Rik, 1700026G02Rik, 1810037K07Rik, 0610039P13Rik*).

RESULTS

Mouse bone marrow cell transduction

The transduction efficiency of the mouse BM cells ranged from 23.9% – 28.6% (IRES-EGFP) and 20.3% - 24.4% (wtStat5-IRES-EGFP). Transplantation experiments in the

α-thalassemic mouse model showed that this result translates *in vivo*. Transplantation of 10^6 transduced mouse BM cells (transduced with either stStat5-IRES-EGFP or the control EGFP construct) results in high levels of chimerism of 90% – 100%. An initial difference in chimerism level was not observed using either construct. The MOI chosen was 1 to have an average integration of 1 per cell.

At 117, 342, 202, 206, 245 and 342 days after the initial transplantation, 37, 34, 10, 10 and 35 mice, respectively, were killed and the bone marrow was retransplanted into secondary recipients (37, 69, 10 and 10 recipients, respectively). These mice were then observed until ethical guidelines required the experiments to be terminated and the mice subjected to a complete autopsy. In the primary recipients we observed 3 malignancies. In the secondary transplantations, 21 malignancies were observed (42).

Integration site analysis

Virus integration sites (VIS) were identified by LAM-PCR and subsequent sequencing of the virus/genome boundaries. This was done in DNA samples from peripheral blood (PB) primary and secondary recipient mice that were healthy at the time (227 to 645 days after transduction) of analysis (174 integrations) and from mice that displayed signs of hematopoietic malignancies (204 integrations). LAM-PCR, modified to be less sensitive, was also performed to determine the most abundant clone present in the samples of mice with malignancies (127 integrations). In total, 505 sequences were obtained, of which known genomic repeat sequences were removed to allow for more specific blast alignment. 397 sequences could be aligned to the mouse genome. Alignments showed that 39% of the integrations were within 10 kbp of a transcription start site (TSS) and 83% within 50 kbp. Evi1, a well-known target for retrovirus integration, was a target for integration on 11 unique locations. Other genes frequently targeted (≥2) are listed in Table 1.

Retrovirus integrations occur near genes expressed in mouse hematopoietic stem cells

The correlation between retrovirus integration sites in long term reconstituting hematopoietic cells, found in our mice, and gene expression levels in LSK cells, was determined by analyzing the expression level of genes closest to the retrovirus integration site within 100 kbp. This resulted in 135 genes of which 117 could be identified on the Affymetrix mouse 430 2.0 micro array. All genes present on the micro array chip were assigned to 10 categories according to the highest expression level measured for each gene as shown in figure 1a. The expression levels of the 117 VIS genes within these categories reveals that retrovirus integration preferentially occurs near highly expressed genes. This preference is significant in both datasets in which the freshly isolated LSK cells ($p=9.55x10^{-9}$) and the (day 2) cytokine stimulated LSK cells ($p=2.47x10^{-7}$) are considered. No obvious differences between the number of genes present in the expression

Table 1: Genes with multiple viral vector integration and their functions (retrieved from Genecards and Entrez Gene). RTCGD genes are indicated in bold, common insertions in RTCGD are indicated with *.

Gene	Number of times retrieved	Function
Evi1*	11	Oncogene
Otud7b	6	Deubiquitination activity
Tln2	4	Cell adhesion
Fcrla	4	B-cell differentiation and lymphomogenesis
Cd4	4	T-cell surface glycoprotein
Usp1	3	Negative regulator of DNA damage repair
Ttl	3	Post-translational addition of tyrosine to alpha-tubulin
Hist1h2ah	3	Wrap and compact DNA into chromatin
Tmprss4	2	Membrane bound protease
Tbpl1	2	Initiation of transcription
Evi5	2	Regulator of cell cycle progression
Pold3	2	DNA replication and repair
Myo1b	2	Cell migration
Klk4	2	Serine protease, implicated in carcinogenesis
Kif2a	2	Microtubule dependent motor
Ivns1abp*	2	Unknown
Incenp	2	Regulator of mitosis
Galtnt11	2	Oligosaccharide biosynthesis
Dgka	2	Attenuates protein kinase C activity
Cct5	2	Assists protein folding
Chid1	2	Carbohydrate catabolism
2610019F03Rik	2	Unknown
Hvcn1	2	Acute production of ROS

level categories using gene expression of the non-stimulated day 0 and day 2-stimulated LSK cells are found.

The distribution of insertions over the range of observed gene expression values (Figure 1b) shows in addition, that although insertions occur at higher expression levels in general, this does not lead to a clustering of the insertions that were retrieved multiple times uniquely at the highest expression levels, indicating that gene specific functions might cause the insertions to cluster near these genes.

The probability of integration is strongly correlated with the expression levels of the surrounding genes

Since it is unclear whether the gene nearest to the retrovirus integration site or an entire locus determines the preference for insertion, all genes within 100 kbp upstream or downstream of a retrovirus integration were taken into account. The RNA expression levels of the total number of genes with a retrovirus integration within the indicated distance of 10, 50, and 100 kbp were determined. Both distances to the transcription start site (TSS) as to the gene itself were calculated. (Table 2). A χ^2 test for goodness of

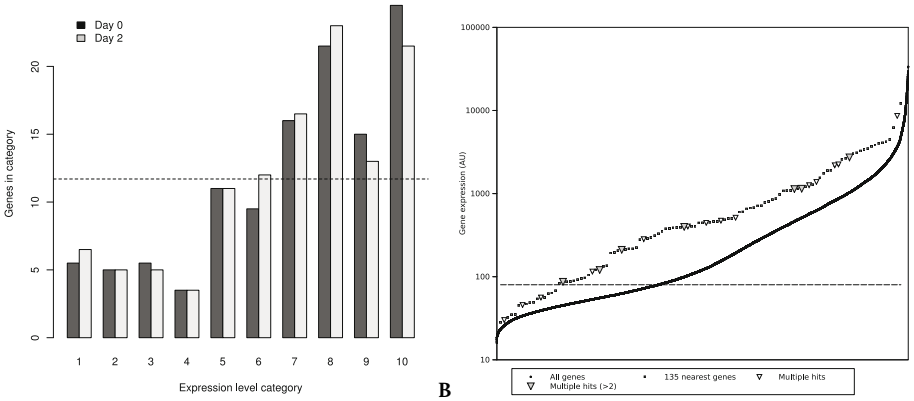


Figure 1: (A) Assigning genes closest to virus integration sites to LSK cell expression level categories. RNA expression values were determined directly after selection of LSK cells or after 2 days of culturing in the presence of mouse SCF, mouse TPO and Flt3L. Expression values for each gene were determined using the highest expression available for each gene symbol after which the gene symbols were sorted for expression and divided into 10 equal sized categories. The number of VIS genes in each category is depicted. The dotted line indicates the number of genes expected in each category assuming uniform random distribution. A Cochran – Armitage test for trend was used to determine whether virus integration occurs more frequently near highly expressed genes and p-values indicating a highly significant trend are shown for both freshly isolated LSK cell gene expressions (Day 0: p=9.55x10^{-9}) and for 2-day stimulated LSK cells (Day 2: p=2.47x10^{-7}). (B) Gene expression values for unstimulated LSK cells showing all genes present on the microarray in black, the 135 RVIS genes in grey and the multiple integrations as inverted triangles. Common RVIS with 2 hits are depicted in white and hits with more than 2 hits are depicted in grey. The dashed line indicates the threshold for expressed genes, as determined by the MAS 5.0 algorithm.

fit, indicating increased likelihood of insertion inside the indicated regions, yields more significant p-values for TSS and genes at higher expression levels and for larger distance taken into consideration (10 kbp, 50 kbp and 100 kbp). For the highest threshold on expression (10000 units) and the smallest distance considered (10 kbp), this test yields non-significant p-values, because only very few genes exceeding the highest threshold have integrations within the indicated distances.

In compliance with the analysis of nearest genes (Figure 1a), these results show that retrovirus integration is not only correlated with the expression of the nearest gene, but also with the average expression of all genes within 100 kbp of the integration site.

Genes upregulated by growth factor stimulation do not display a higher frequency of retroviral landing than other expressed genes

The RNA expression data allowed evaluation of the effect of cytokine stimulation as used in the gene transfer protocol on the integration of the retrovirus. The differences in gene expression between freshly isolated LSK (day 0) and 2-day SCF, Flt3L and TPO stimulation were analyzed (Figure 2a), which identified 780 probesets describing 632 genes that are differentially expressed. Of the 135 genes closest to the VIS, 11 (8.1%) were

Table 2: Integrations occur more often near highly expressed genes. Distances to all Ensembl genes and their transcription start sites (TSS) were calculated. The number of genes exceeding the expression threshold with a virus integration within the indicated distance was reported. The ratio of genes meeting these criteria was divided by the total number of virus integrations. This ratio was then tested against the ratio of the length represented by the genes exceeding the threshold and the indicated surrounding area and the total mouse genome size, using a χ^2 test for goodness of fit. *p*-values indicate increased likelihood of insertion inside the region determined by the expression threshold and the distance from TSS or gene.

Area	Expression				
	>100	>300	>1000	>3000	>10000
Gene	2.49×10^{-1}	3.80×10^{-1}	**5.97×10^{-3}**	7.51×10^{-2}	**9.41×10^{-3}**
Gene ± 10 kpb	2.95×10^{-1}	**2.00×10^{-4}**	**1.50×10^{-10}**	**3.48×10^{-5}**	3.62×10^{-1}
Gene ± 50 kpb	-	-	**3.10×10^{-3}**	**6.85×10^{-13}**	**3.24×10^{-6}**
Gene ± 100 kpb	-	-	-	**5.04×10^{-5}**	**1.69×10^{-3}**
TSS ± 10 kpb	1.40×10^{-1}	**8.54×10^{-5}**	**8.68×10^{-10}**	7.91×10^{-2}	9.11×10^{-1}
TSS ± 50 kpb	-	-	**8.08×10^{-4}**	**1.63×10^{-13}**	**1.02×10^{-15}**
TSS ± 100 kpb	-	-	-	**6.75×10^{-6}**	**7.64×10^{-4}**

differentially expressed, but did not exceed the 4-fold expression difference threshold (Figure 2b). Only one of these genes was differentially up-regulated, the other 10 were expressed at lower levels as a result of the cytokine stimulation. Expanding the analysis to the 632 genes present on the micro array within 100 kbp upstream or downstream of the virus integration, 26 (4.1%) are differentially regulated of which 5 are up-regulated and 21 are down-regulated.

Gene expression analysis of both nearest genes and genes within 100 kbp upstream or downstream showed that cytokine treatment as performed in the transduction protocol did not result in higher incidence of virus integration in cytokine – upregulated genes. We can therefore conclude that the relation to insertion near expressed genes in this study is not altered by the cytokine treatment of the target cells.

Global functional analysis of the genes surrounding integration sites

As described (43,44) the possible deregulatory effects on expression of genes near the retrovirus integration site by the promoter region of the transgene are not limited to the genes closest to the virus integration site. To elucidate the what possible effect the insertions might bring about in nearby genes, the function of both the nearest genes and the genes within 100 kbp of the retrovirus integration was studied.

To this end, the function of these genes and their known involvement in leukemogenesis were analyzed by screening against 3 well known databases, i.e., the retrovirus tagged cancer database (RTCGD; genes involved in leukemogenesis), the Mouse Genome Informatics (MGI) phenotype browser (function of genes in tumor phenotype, immunologic phenotype and hematopoietic phenotype) and Ingenuity Pathway Analysis (network analysis). Figure 3 shows that RTCGD genes, categorized into 10 expression levels

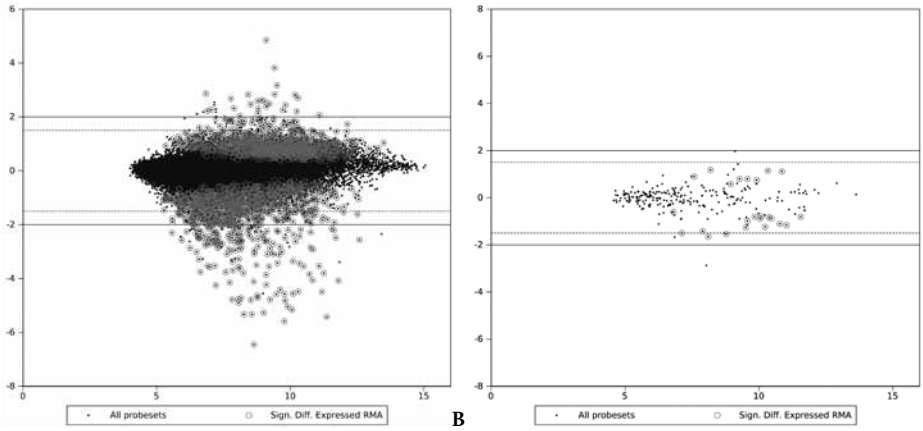


Figure 2: Differentially expressed probesets after mSCF, mTPO and Flt3L stimulation. Using log2 fold change >2 and Šidák step-down corrected p-values <0.05 (indicated with red circles) as a threshold, 151 genes were differentially expressed after cytokine stimulation. 20 genes were upregulated, and 131 genes were downregulated. (B) None of the probesets describing the 135 RIS genes were differentially expressed, although small significant differences for 4 upregulated and 13 downregulated genes were observed.

and assigned to either the day 0 micro array data set or the day 2 data set, are mostly expressed in the higher level categories and show a similar behavior as the insertion sites retrieved in this study. Moreover, when analyzing twenty-four of the 117 VIS (nearest genes) that were found to be present in the RTCGD database, it was obvious that these genes are also found in the highest expression categories (Figure 3). To confirm this notion, we performed a χ^2 test for trend on the fraction of RTCGD VIS genes of the total RTCGD genes. This test showed significantly (p=0.024) more RTCGD VIS in the higher expression bins than were observed in the entire RTCGD dataset. Secondly, functional information was obtained by determining the presence of the genes of both data sets in the Mouse Genome Informatics (MGI) phenotype browser[14]. Twenty-six genes (of the 117 VIS) are found in transgenic animal models with tumor, immunological or hematological phenotype (MGI mammalian phenotype browser) (Supplementary Figure 5).

Of the 135 genes nearest to integration sites 71 are annotated in Ingenuity Pathway Analysis. Of these genes, 62 cluster in 5 networks with hematological system development and function (16 genes), immune response (14 genes), cellular assembly and organization (13 genes), cell-to-cell signaling and interaction (11 genes), immune and lymphatic system development and function (11) genes as the five highest functions. Of the 632 genes within 100 kbp upstream or downstream of the VIS 256 genes could be used to generate 18 functionally related networks, the 5 gene functions with the largest number of genes involved were cancer (13 genes), cellular movement (9 genes), cellular assembly and organization (8 genes), hematological disease (8 genes) and hematological system

14 http://www.informatics.jax.org/searches/MP_form.shtml

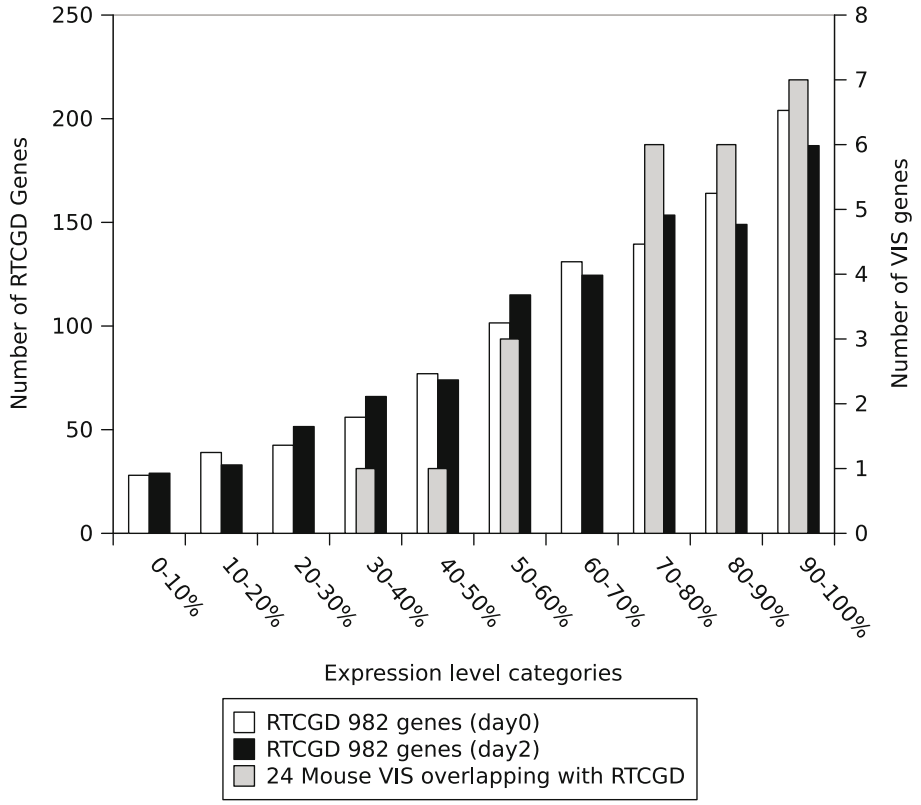


Figure 3: Genes present in the retrovirus-tagged cancer gene database (RTCGD, 52) assigned to freshly isolated and 2-day cytokine stimulated LSK cell gene expression level categories. 1130 genes present in RTCGD were assigned to LSK expression level categories (blue bars, day0 and purple bars, day2). The grey bars indicate the RTCGD genes also present in the 24 genes closest to retrovirus integrations.

development and function (7 genes). Presence of genes with immune system and hematological functions in these networks, as well as leukemia associated genes (RTCGD) and genes with immunological, hematological and tumor phenotypes in the MGI analysis suggest that at least part of the genes either next to or within 100 kbp of a virus integration are involved in hematopoietic reconstitution and development of leukemias in mice. A closer look into the gene expression values obtained from the LSK microarrays after Ingenuity annotation in addition showed that with increasing expression, the number of unique genes associated with immunological and hematopoietic functionality increased. (Figure 4a) With increasing expression, the fraction of genes associated with diseases in the hematopoietic and immunological functions also increased, until it reached 72% and 74% for hematopoietic and immunological functions respectively. Additionally general cell functions such as Cell cycle and protein synthesis increase with expression level in LSK cells. When looking into metabolism associated genes, no such increase is observed

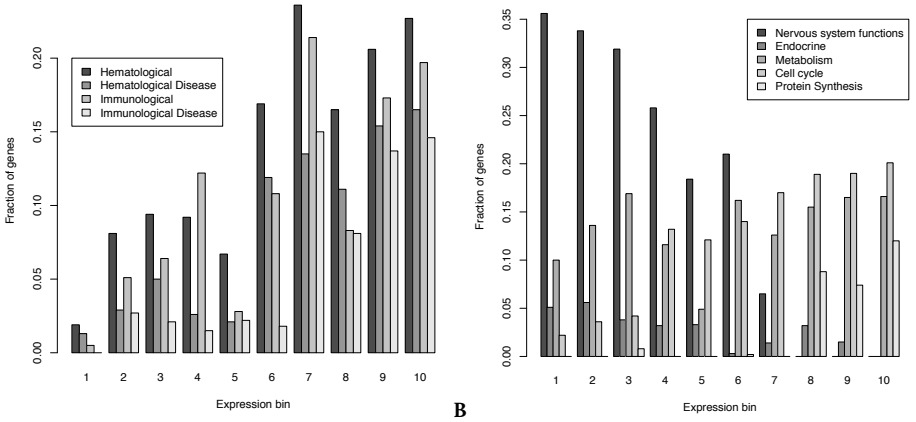


Figure 4: Hematopoietic and immunological annotations of the genes in different expression level categories. Gene expression values obtained from LSK cells were divided into ten equal sized categories. The number of unique genes annotated in the Ingenuity Pathway Analysis with (A) hematopoietic or immunological functions are indicated as well as the number of unique genes annotated with hematopoietic or immunological disease or (B) the neurological and endocrine functions as well as metabolic, cell cycle and protein synthesis functions. The fraction of annotated functions in the bin with the specific function is indicated.

(Figure 4b) and for endocrine functionality and nervous system development functions, the number of genes is negatively correlated with LSK gene expression.

Comparison of the murine integration sites with those in clinical trials for X-linked SCID

For studies on the safety of retrovirus as a gene therapy vector in mouse models it is important that the model reflects the human situation. The mouse virus integrations found in this mouse study were compared to those found in two human X-SCID studies, where retrovirus was used to express the *IL2RG* gene. 11 of the 125 human homologues to the 135 mouse nearest genes to retrovirus integrations (7.4%) were also present in 481 nearest genes to the retrovirus integration sites identified in the Paris XSCID trial, whereas 6 of the 125 human homologues to 135 mouse VIS genes (5.2%) were present in 523 nearest genes the London XSCID trial. (Table 3) Although the actual number of overlapping genes is low, Fisher exact tests to determine whether these numbers could be obtained by uniform random integration of the virus showed significance for both comparisons. (Fisher exact two sided p-values, Paris trial: p=0.0028, London trial: p=0.008753). Interestingly, when comparing the overlap with a pyrosequencing experiment (Chapter 4) performed on samples obtained in the London XSCID trial, one fifth of the insertions retrieved in our mouse model co-occurred.

Interestingly, *LMO2*, the gene associated with the T-cell proliferations that occurred in 4 patients in the Paris XSCID trial, was not found in this mouse dataset. Finding such large overlaps between insertion studies using similar viral vectors, but performed in

Table 3: Overlap between integration retrieved in the London and Paris XSCID trials and the mouse integration sites retrieved in this study. The human homologues of the mouse genes were retrieved using NCBI Homologene.

	Closest mouse genes (125 genes)	Hits with locus (640 genes)
London XSCID (523 genes)	7 (5.2%)	12 (1.9%)
Paris XSCID (481 genes)	10 (7.4%)	12 (1.9%)
All non-454 XSCID (926 genes)	13 (8.9%)	21 (3.2%)
London XSCID 454 (2089 genes)	25 (20%)	46 (7.1%)

different species, shows that virus insertion pattern obtained with similar viral backbones is related to the transduced cell type, more than the species.

DISCUSSION

Consistent with previous reports, this study demonstrates that retrovirus preferentially integrates in or near highly expressed genes in immature hematopoietic cells of the mouse and that integration is correlated with the level of expression of the nearest gene and the locus 100 kbp upstream and downstream of that gene. As much as 83% of the integrations were found near expressed genes, with a highly significant overrepresentation of proto-oncogenes, genes involved in general cell functions and genes involved in hematopoietic and immune functions. Comparison with human hematopoietic stem cell integration sites displayed a strongly conserved pattern, predicting a considerable overlap of integration in individual patients and among independent clinical trials. Gene expression has previously been shown to direct the integration of retrovirus to transcription start sites and lentivirus to in-gene regions in cell lines (HeLa, 293T, IMR90 and supT1 (6,45,46), as well as human PBMC(6) and $CD34^{+}CD38^{-}$ BM cells (47), but other mechanism may also be involved, such as the weak preference for specific nucleotide sequences (48,49)), fragile sites (50) and open chromatin configurations (46,49,51). This study shows that virus integration in the long term repopulating cells, which are target cells in many gene therapy protocols, is for a large proportion directed by the gene expression levels in these cells.

The use of replicating gamma-retroviruses raises difficulties when analyzing the function of the insertion sites, because malignancies observed in mice infected with these viruses typically show a large number of insertions in the affected tissue. Single clones therefore will also have several insertions. To address difficulties in analysis of true functional insertions, mathematical approaches such as clustering of insertion sites (17) and the identification of common insertion sites (12) have been proposed. Instead, we aimed at a low number of insertions per cell to ensure only 1 or 2 virus insertions occurred per cell. The initial transductions lead to 23.9-28.6% transduction in the EGFP group and 20.3-24.4% in the wtStat5 group. This is within the range that

results predominantly in 1 insertion per cell, thus allowing us to focus on the function of a single insertion rather than a host of insertions in a single cell.

A comprehensive description of the insertion profile should focus on the description of insertions using both the closest genes and the surrounding locus, since the influence of the viral promoter is not uniquely restricted to the closest gene (43,44). In our analysis, we analyzed both datasets, obtained from normal and malignant tissues and retrieved 23 insertions that occurred more than once in a dataset of 135 unique insertions. Thirteen of these insertions had been associated with hematopoietic malignancies before (52).

Gene expression has previously been shown to direct integration of gamma-retrovirus to transcription start sites and lentivirus to in-gene regions of transcribed genes in cell lines (6) as well as in human PBMC (29) and $CD34^{+}CD38^{-}$ BM cells, but other mechanisms might also be involved. Insertions also favor fragile sites and open chromatin and have been shown to have a weak preference for specific nucleotide sequences. This study shows that virus integration in the long term repopulating cells, which are the targets in hematopoietic gene therapy protocols since their transduction is expected to lead to a lifetime of gene correction, is in a large proportion directed by the gene expression level in the host cell. Not only the expression of the closest genes dictates the insertion behavior, but the entire locus is draws insertions when highly expressed. An exception to this pattern is represented by the common insertion sites, which unexpectedly do not cluster at highly expressed genes uniquely, pointing at other mechanisms such as accelerated cell growth resulting in clonal dominance. Similarly, the RVIS genes retrieved in this study that also occur in the RTCGD, show an even more pronounced tendency to insert near highly expressed genes, also hinting at mechanisms other than just high expression. On the other hand, when analyzing the hematopoietic and immunological function annotation in LSK gene expression bins we observed a clear increase in the number of genes associated with these functions with higher expression in LSK cells. This effect was even more pronounced when focusing on genes associated with diseases in these functions. It is tempting to speculate that the increase in insertions in RTCGD genes, the relation between expression and RTCGD gene presence (figure 3) and the functions of the highly expressed genes in the transduced cells that repopulated the mice point to cell type specific vulnerability of the hematopoietic stem cells for transduction with gamma-retroviral vectors. Indeed, the most commonly found insertion in this study, *Evi1*, is known to accelerate cell growth (53) and also commonly occurs in an *in vitro* assay for immortalization of mouse bone marrow cells (30). There is however, a drawback in the use of RTCGD for gamma-retrovirus insertion site annotation. The data in RTCGD was generated from experiments using replication competent gamma-retroviruses, and many of the genes listed in it can be considered

bystanders to genes that actually act as oncogenes. Finding a gene in RTCGD therefore does not necessarily mean that such a gene is related to leukemia directly.

Next to the accessibility of the gene due to open chromatin structure, other mechanisms for the preferred integration must be involved. It might be possible that these lowly expressed genes are present in an area of open chromatin and that the expression of genes is tightly regulated. Alternatively, a small proportion of the virus integrations might be driven by an unidentified process or might be truly random.

In mouse bone marrow one in 3-5x10^4 cells is expected to give rise to repopulation (54). We isolated bone marrow cells by Percoll density centrifugation, which increase the repopulating cell content by a factor four. The transplantations ranged from 10^4 to 10^6 cells per mouse (1 to 100 repopulating cells per mouse), on which basis it is assumed that hematopoiesis from transduced cells in the transplanted mice originates from a limited number of stem cells. We related the gamma-retroviral vector insertions in these cells to gene expression levels in LSK cells and showed a preference for expressed genes. The hematopoietic stem cell is pluripotent and, similarly to embryonic stem cells (55), have a characteristic open chromatin formation (56). Upon differentiation, H3K4 methylation, a marker of active chromatin regions, has been shown to decrease (56), inactivating parts of the genome. Not all genes are expressed in HSC, however, but the open chromatin structure in these cells could prime genes for expression. Studies in clonal succession in of human hematopoietic cells in mice (57) have shown that only a small part of the hematopoietic cell population is contributing to hematopoiesis, similar to what has been shown in earlier studies in mice (54), in cats (58) and in rhesus monkeys (59, 60). McKenzie (57) therefore proposed that a stochastic process might be involved in the decision to stay dormant or become activated (also modeled by Roeder (61)). This poses a difficulty for gene expression measurements in HSC populations: if the phenotype of the cells does not distinguish its behavior, but rather a stochastic process, gene expression measurements on a bulk population might not be informative for the repopulating cells, but only on the entire population (which might contain a certain proportion of cells that will become activated)

The retroviral transduction protocol, which consists of growth factor stimulation to allow efficient transduction of the target cells, has been considered a potential cause for the occurrence of insertions in malignant loci, i.e. those of which expression is stimulated by the supplied growth factors (62). In our experiments, such an effect was not observed, since only very few genes in the next to insertions showed significant differences in expression after cytokine treatment. This is in line with earlier analysis performed using the human XSCID integrations and their target cell expression (25) and leads to the conclusion that cytokine stimulation does not appear to be a prominent determinant of potentially dangerous insertion patterns.

Functional analysis of the gene targeted gene loci can be performed by using different available sources for annotation, such as gene ontology, functional pathways and phenotype databases. The coverage of these databases is not complete, because not all transcripts have been assigned a function. Further, genes might have different functions in different tissues or might not have been correctly described. Since half of the genes in our dataset have functions that can clearly be associated with a meaningful function in the hematopoietic cell, genes lacking annotation or which annotations seem irrelevant to the tissue, might be relevant to hematopoiesis as well. This specifically holds for the common insertion sites observed, since one would expect that many of these genes have a function in hematopoiesis that leads to their independent selection in several transductions. These common insertion sites, as well as highly expressed genes in hematopoietic cells might therefore be interesting targets for further functional research.

An insertion profile study such as the one presented here aims to be predictive for the behavior of gamma-retroviral integration patterns in human cells. Obvious differences between mouse and human such as life span of the individual, metabolic rate (63) and even functional and expression differences between species might indeed present hurdles for the translation of mouse data to the human situation. However, since similar patterns of integration were found in the individual XSCID trial data sets (24,25) and a highly significant one fifth of our limited murine dataset integrations were identical in the nearest gene, mouse models should be highly predictive for human integration patterns.

The insertion pattern of gamma-retroviral vectors favors genes that have a function in hematopoiesis, which might well be explained by their evolutionary development as oncoviruses. The target specificity for genes with hematopoietic function, or genes that are expressed in the hematopoietic system makes these vectors suitable for transduction of hematopoietic cells, but carries the risk of deregulation of these genes. Further improvements in vector design (64,65) have shown that the self- inactivating design with an internal promoter, or the addition of insulator elements (66) will reduce genotoxicity of gamma-retroviral vectors. A further option for improvement consists of the transduction of hematopoietic stem cells under such cell culture conditions that allow proliferation and transduction, but maintain engraftment potential of the transplant. Such protocols would only require a limited number of cells, which might be pre-screened for possibly harmful insertions. Another avenue for safety improvement entails limiting the number of transplanted target cells by further purification. LSK cells in mice and $CD34^{+}CD38^{-}$ cells in human cannot currently be transduced efficiently with gamma-retroviral vectors, but lentiviral vectors, which transduce the non-proliferating cells as efficiently as proliferating cells do not have this limitation. Considering the existence of improved vector backbones and the integration patterns of gamma-retroviral vectors, LTR driven gamma-retroviral vectors, however efficient, are no longer recommended for therapeutic use in hematopoietic stem cells. Improved

transduction protocols and optimized vector backbones, both lentiviral and retroviral, need to replace the first generation of therapeutic vectors.

Acknowledgments

The authors would like to thank Edwin de Haas, Department of Immunology, ErasmusMC Rotterdam for cell sorting, as well as the Experimental animal facility at ErasmusMC Rotterdam for biotechnical support. TF targetmapper was designed by Sebastiaan Horsman, Department of Bioinformatics, ErasmusMC Rotterdam. The authors acknowledge Drs. F.J.T. Staal and K. Pike-Overzet, Dept. Immunology, Erasmus University Medical Center, for assistance in the UCB array analyses.

REFERENCES

1. Hacein-Bey-Abina S, Von Kalle C, Schmidt M, et al. Lmo2-associated clonal t cell proliferation in two patients after gene therapy for scid-x1. Science. 2003;302:415-9.
2. Hacein-Bey-Abina S, Garrigue A, Wang GP, et al. Insertional oncogenesis in 4 patients after retrovirus-mediated gene therapy of scid-x1. J. Clin. Invest. 2008;118:3132-3142.
3. Howe SJ, Mansour MR, Schwarzwaelder K, et al. Insertional mutagenesis combined with acquired somatic mutations causes leukemogenesis following gene therapy of scid-x1 patients. J Clin Invest. 2008;118:3143–3150.
4. Li Z, Dullmann J, Schiedlmeier B, et al. Murine leukemia induced by retroviral gene marking. Science. 2002;296:497.
5. Montini E, Cesana D, Schmidt M, et al. Hematopoietic stem cell gene transfer in a tumor-prone mouse model uncovers low genotoxicity of lentiviral vector integration. Nat Biotechnol. 2006;24:687-96.
6. Mitchell RS, Beitzel BF, Schroder AR, et al. Retroviral dna integration: aslv, hiv, and mlv show distinct target site preferences. PLoS Biol. 2004;2:E234.
7. Baum C, Kustikova O, Modlich U, Li Z, Fehse B. Mutagenesis and oncogenesis by chromosomal insertion of gene transfer vectors. Hum Gene Ther. 2006;17:253-63.
8. Themis M, Waddington SN, Schmidt M, et al. Oncogenesis following delivery of a nonprimate lentiviral gene therapy vector to fetal and neonatal mice. Mol Ther. 2005;12:763-71.
9. Seggewiss R, Pittaluga S, Adler RL, et al. Acute myeloid leukemia is associated with retroviral gene transfer to hematopoietic progenitor cells in a rhesus macaque. Blood. 2006;107:3865-7.
10. Moloney JB. Biological studies on a lymphoid-leukemia virus extracted from sarcoma 37. i. origin and introductory investigations. J Natl Cancer Inst. 1960;24:933-51.
11. Friend C. Transplantation immunity and the suppression of spleen colony formation by immunization with murine leukemia virus preparation (friend). Int. J. Cancer. 1968;3:523-529.
12. Suzuki T, Shen H, Akagi K, et al. New genes involved in cancer identified by retroviral tagging. Nat Genet. 2002;32:166-74.
13. Erkeland SJ, Verhaak RG, Valk PJ, et al. Significance of murine retroviral mutagenesis for identification of disease genes in human acute myeloid leukemia. Cancer Res. 2006;66:622-6.
14. Joosten M, Vankan-Berkhoudt Y, Tas M, et al. Large-scale identification of novel potential disease loci in mouse leukemia applying an improved strategy for cloning common virus integration sites. Oncogene. 2002;21:7247-55.
15. Mikkers H, Allen J, Knipscheer P, et al. High-throughput retroviral tagging to identify components of specific signaling pathways in cancer. Nat Genet. 2002;32:153-9.

16. Lund AH, Turner G, Trubetskoy A, et al. Genome-wide retroviral insertional tagging of genes involved in cancer in cdkn2a-deficient mice. Nat Genet. 2002;32:160-5.
17. de Ridder J, Kool J, Uren A, et al. Co-occurrence analysis of insertional mutagenesis data reveals cooperating oncogenes. Bioinformatics. 2007;23:i133-41.
18. Modlich U, Kustikova OS, Schmidt M, et al. Leukemias following retroviral transfer of multidrug resistance 1 (mdr1) are driven by combinatorial insertional mutagenesis. Blood. 2005;105:4235-46.
19. Pike-Overzet K, de Ridder D, Weerkamp F, et al. Ectopic retroviral expression of lmo2, but not il2r-gamma, blocks human t-cell development from cd34+ cells: implications for leukemogenesis in gene therapy. Leukemia. 2007;21:754-63.
20. Hematti P, Hong BK, Ferguson C, et al. Distinct genomic integration of mlv and siv vectors in primate hematopoietic stem and progenitor cells. PLoS Biol. 2004;2:e423.
21. Kustikova OS, Schiedlmeier B, Brugman MH, et al. Cell-intrinsic and vector-related properties cooperate to determine the incidence and consequences of insertional mutagenesis. Mol. Ther. 2009;.
22. Cattoglio C, Facchini G, Sartori D, et al. Hot spots of retroviral integration in human cd34+ hematopoietic cells. Blood. 2007;110:1770-8.
23. Ott MG, Schmidt M, Schwarzwaelder K, et al. Correction of x-linked chronic granulomatous disease by gene therapy, augmented by insertional activation of mds1-evi1, prdm16 or setbp1. Nat Med. 2006;12:401-9.
24. Deichmann A, Hacein-Bey-Abina S, Schmidt M, et al. Vector integration is nonrandom and clustered and influences the fate of lymphopoiesis in scid-x1 gene therapy. J Clin Invest. 2007;117:2225-32.
25. Schwarzwaelder K, Howe SJ, Schmidt M, et al. Gammaretrovirus-mediated correction of scid-x1 is associated with skewed vector integration site distribution *in vivo*. J Clin Invest. 2007;117:2241-9.
26. Maxfield LF, Fraize CD, Coffin JM. Relationship between retroviral dna-integration-site selection and host cell transcription. Proc Natl Acad Sci U S A. 2005;102:1436-41.
27. Stocking C, Bergholz U, Friel J, et al. Distinct classes of factor-independent mutants can be isolated after retroviral mutagenesis of a human myeloid stem cell line. Growth Factors. 1993;8:197-209.
28. Laufs S, Gentner B, Nagy KZ, et al. Retroviral vector integration occurs in preferred genomic targets of human bone marrow-repopulating cells. Blood. 2003;101:2191-8.
29. Laufs S, Nagy KZ, Giordano FA, et al. 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;10:874-81.
30. Modlich U, Bohne J, Schmidt M, et al. Cell-culture assays reveal the importance of retroviral vector design for insertional genotoxicity. Blood. 2006;108:2545-53.
31. Bokhoven M, Stephen SL, Knight S, et al. Insertional gene activation by lentiviral and gammaretroviral vectors. J. Virol. 2009;83:283-294.
32. Smith LG, Weissman IL, Heimfeld S. Clonal analysis of hematopoietic stem-cell differentiation *in vivo*. Proc Natl Acad Sci U S A. 1991;88:2788-92.
33. van Hennik PB, Verstegen MM, Bierhuizen MF, et al. Highly efficient transduction of the green fluorescent protein gene in human umbilical cord blood stem cells capable of cobblestone formation in long-term cultures and multilineage engraftment of immunodeficient mice. Blood. 1998;92:4013-22.
34. Wagemaker G, Visser TP. Erythropoietin-independent regeneration of erythroid progenitor cells following multiple injections of hydroxyurea. Cell Tissue Kinet. 1980;13:505-17.
35. Wagemaker G. Selective multiplication of hematopoietic stem cells for bone marrow transplantation in mice and rhesus monkeys. Transplant Proc. 1987;19:2721-5.
36. Bierhuizen MF, Westerman Y, Visser TP, et al. Enhanced green fluorescent protein as selectable marker of retroviral-mediated gene transfer in immature hematopoietic bone marrow cells. Blood. 1997;90:3304-15.
37. Wognum AW, Visser TP, Peters K, Bierhuizen MF, Wagemaker G. Stimulation of mouse bone marrow cells with kit ligand, flt3 ligand, and thrombopoietin leads to efficient retrovirus-mediated gene transfer to stem cells, whereas interleukin 3 and interleukin 11 reduce transduction of short- and long-term repopulating cells. Hum Gene Ther. 2000;11:2129-41.

38. Moritz T, Dutt P, Xiao X, et al. Fibronectin improves transduction of reconstituting hematopoietic stem cells by retroviral vectors: evidence of direct viral binding to chymotryptic carboxy-terminal fragments. Blood. 1996;88:855-62.
39. Markowitz D, Goff S, Bank A. Construction and use of a safe and efficient amphotropic packaging cell line. Virology. 1988;167:400-406.
40. Lovatt A, Black J, Galbraith D, et al. High throughput detection of retrovirus-associated reverse transcriptase using an improved fluorescent product enhanced reverse transcriptase assay and its comparison to conventional detection methods. J. Virol. Methods. 1999;82:185-200.
41. Schmidt M, Hoffmann G, Wissler M, et al. Detection and direct genomic sequencing of multiple rare unknown flanking dna in highly complex samples. Hum Gene Ther. 2001;12:743-9.
42. Brugman MH, Visser TP, Arshad SP, et al. Characteristics of gammaretrovirus integration-related leukemias in mice. Submitted for publication..
43. Kustikova OS, Wahlers A, Kuhlcke K, et al. Dose finding with retroviral vectors: correlation of retroviral vector copy numbers in single cells with gene transfer efficiency in a cell population. Blood. 2003;102:3934-7.
44. Bartholomew C, Ihle JN. Retroviral insertions 90 kilobases proximal to the evi-1 myeloid transforming gene activate transcription from the normal promoter. Mol Cell Biol. 1991;11:1820-8.
45. Wu X, Li Y, Crise B, Burgess SM. Transcription start regions in the human genome are favored targets for mlv integration. Science. 2003;300:1749-51.
46. Schroder AR, Shinn P, Chen H, et al. Hiv-1 integration in the human genome favors active genes and local hotspots. Cell. 2002;110:521-9.
47. Wagner W, Laufs S, Blake J, et al. Retroviral integration sites correlate with expressed genes in hematopoietic stem cells. Stem Cells. 2005;23:1050-8.
48. Holman AG, Coffin JM. Symmetrical base preferences surrounding hiv-1, avian sarcoma/leukosis virus, and murine leukemia virus integration sites. Proc Natl Acad Sci U S A. 2005;102:6103-7.
49. Carteau S, Hoffmann C, Bushman F. Chromosome structure and human immunodeficiency virus type 1 cdna integration: centromeric alphoid repeats are a disfavored target. J Virol. 1998;72:4005-14.
50. Bester AC, Schwartz M, Schmidt M, et al. Fragile sites are preferential targets for integrations of mlv vectors in gene therapy. Gene Ther. 2006;.
51. Rohdewohld H, Weiher H, Reik W, Jaenisch R, Breindl M. Retrovirus integration and chromatin structure: moloney murine leukemia proviral integration sites map near dnase i-hypersensitive sites. J Virol. 1987;61:336-43.
52. Akagi K, Suzuki T, Stephens RM, Jenkins NA, Copeland NG. Rtcgd: retroviral tagged cancer gene database. Nucleic Acids Res. 2004;32:D523-7.
53. Yuasa H, Oike Y, Iwama A, et al. Oncogenic transcription factor evi1 regulates hematopoietic stem cell proliferation through gata-2 expression. Embo J. 2005;24:1976-87.
54. Micklem HS, Lennon JE, Ansell JD, Gray RA. Numbers and dispersion of repopulating hematopoietic cell clones in radiation chimeras as functions of injected cell dose. Exp. Hematol. 1987;15:251-257.
55. Efroni S, Duttagupta R, Cheng J, et al. Global transcription in pluripotent embryonic stem cells. Cell Stem Cell. 2008;2:437-447.
56. Chung YS, Kim HJ, Kim T, et al. Undifferentiated hematopoietic cells are characterized by a genome-wide under-methylation dip around the transcription start-site and a hierarchical epigenetic plasticity. Blood. 2009;.
57. McKenzie JL, Gan OI, Doedens M, Wang JCY, Dick JE. Individual stem cells with highly variable proliferation and self-renewal properties comprise the human hematopoietic stem cell compartment. Nat. Immunol. 2006;7:1225-1233.
58. Abkowitz JL, Linenberger ML, Newton MA, et al. Evidence for the maintenance of hematopoiesis in a large animal by the sequential activation of stem-cell clones. Proc. Natl. Acad. Sci. U.S.A. 1990;87:9062-9066.

59. Schmidt M, Zickler P, Hoffmann G, et al. Polyclonal long-term repopulating stem cell clones in a primate model. Blood. 2002;100:2737-2743.
60. Kuramoto K, Follman D, Hematti P, et al. The impact of low-dose busulfan on clonal dynamics in nonhuman primates. Blood. 2004;104:1273-1280.
61. Roeder I, Horn K, Sieburg H, et al. Characterization and quantification of clonal heterogeneity among hematopoietic stem cells: a model-based approach. Blood. 2008;112:4874-4883.
62. Kohn DB, Sadelain M, Glorioso JC. Occurrence of leukaemia following gene therapy of x-linked scid. Nat Rev Cancer. 2003;3:477-88.
63. Rangarajan A, Weinberg RA. Opinion: comparative biology of mouse versus human cells: modelling human cancer in mice. Nat Rev Cancer. 2003;3:952-9.
64. Zychlinski D, Schambach A, Modlich U, et al. Physiological promoters reduce the genotoxic risk of integrating gene vectors. Mol Ther. 2008;16:718-25.
65. Modlich U, Navarro S, Zychlinski D, et al. Insertional transformation of hematopoietic cells by self-inactivating lentiviral and gammaretroviral vectors. Mol. Ther. 2009;.
66. Esnault G, Majocchi S, Martinet D, et al. Transcription factor ctf1 acts as a chromatin domain boundary that shields human telomeric genes from silencing. Mol. Cell. Biol. 2009;29:2409-2418.

SUPPLEMENTARY DATA

	0-10%	10-20%	20-30%	30-40%	40-50%	50-60%	60-70%	70-80%	80-90%	90-100%
										Item2b
										Cct5
										Pcbp1
										Rbx1
								F11r		*Ptprc*
								AI462493		***Scap2***
								Zfp217		***Dgkz***
								1300001I01Rik		*Anxa1*
								Cd97	***Kif2a***	*C1qr1*
							B4galt5	***Serpini1***	*Ccl4*	***Ptpre***
							Syde2	2310007H09Rik	***Evi1***	*Zfpn1a1*
							Prdm16	0610039P13Rik	*Il10rb*	***Iqgap1***
							Pkn2	Galnt11	*Pold3*	***Slc2a3***
							Ccdc21	*Tbpl1*	Fmnl2	*Anp32e*
							1810037K07Rik	*Dgka*	***Pip5k1c***	***Usp1***
							Wrn	Fgd5	Tars	*Ier3*
					Tmprss4	**Tbn**	***Tpm1***	*Cog6*	Rusc1	***Incenp***
					Pros1	*D15Ertd366e*	Fcrlm1	Syne1	***Shmt1***	***Angpt1***
				Elovl7	**Ttl**	A730041O15Rik	*Cd200*	Otud7b	3110045G13Rik	*Ivnsabp*
		Sgip1		Trim36	***Sry***	BC067068	Plxdc2	Parp8	Cxcl10	*Rasa3*
	Ypel2	6330406I15Rik	Klhl1	***Tlr1***	Echdc2	Wdfy1	6620401K05Rik	*Cap1*	*Smndc1*	*Ly86*
	Abat	*Myo1b*	2610018I05Rik	**Tln2**	***P2ry2***	*Hivep2*	AW050020	Zc3h5	Spred2	***Nrip1***
	Klk4	Dnaic1	2610019F03Rik	1700026G02Rik	Mela	*Ripk2*	St3gal2	***Rab5c***	*Arid1b*	Arhgap25
	1810030J14Rik	4930485B16Rik	Syt13	2300003P22Rik	***Atp1a2***	B830013J05Rik	3110023E09Rik	Rin3	***Ptk2b***	*Nfkb1*
	0-10%	10-20%	20-30%	30-40%	40-50%	50-60%	60-70%	70-80%	80-90%	90-100%
RTCGD genes	0	0	0	1	1	3	0	6	6	7
MGI phenotype annotation	0	0	0	2	4	1	2	3	5	9
Ingenuity focus genes	0	1	0	1	4	4	4	10	11	22

Supplementary Figure: Genes closest to virus integration sites were assigned to LSK (dayo) gene expression level categories and RTCGD genes, MGI phenotype annotations (for tumor phenotype, immunologic phenotype and hematopoietic phenotype) and Ingenuity focus genes are indicated (grey, bold and italic, respectively). The order in each category represents the expression levels of the genes. The bottom table shows the number of genes with the indicated functional annotation for each expression level category.

CHAPTER

4

Integration sites in engrafted cells cluster within a limited repertoire of genes after retrovirus vector gene therapy

Adapted from: Annette Deichmann[*,1] Martijn H. Brugman[*,2,3] Cynthia C. Bartholomae[*,1] Kerstin Schwarzwaelder[1], Monique M.A. Verstegen[2], Steven J. Howe[4], Anne Arens[1,5], Marion G. Ott[6], Dieter Hoelzer[6], Reinhard Seger[7], Manuel Grez[8], Salima Hacein-Bey Abina[9,10], Marina Cavazzana-Calvo[9,10], Alain Fischer[9,11], Anna Paruzynski[1], Richard Gabriel[1], Hanno Glimm[1], Ulrich Abel[1,12], Claudia Cattoglio[13], Fulvio Mavilio[13,14], Barbara Cassani[15], Alessandro Aiuti[15,16], Cynthia E. Dunbar[17], Christopher Baum[3], H. Bobby Gaspar[4,18], Adrian J. Thrasher[4,18], Christof von Kalle[§,1], Manfred Schmidt[#,1] and Gerard Wagemaker[#,2]. Integration sites in engrafted cells cluster within a limited repertoire of genes after retrovirus vector gene therapy. *Submitted for publication.*

[1] *Department of Translational Oncology, National Center for Tumor Diseases and German Cancer Research Center (DKFZ), Heidelberg, Germany*

[2] *Department of Hematology, Erasmus Medical Center, Rotterdam, The Netherlands*

[3] *Department of Experimental Hematology, Hannover Medical School, Hannover, Germany*

[4] *Molecular Immunology Unit, Institute of Child Health, University College, London, United Kingdom*

[5] *Core Facility of Proteomics and Genomics, German Cancer Research Center, Heidelberg, Germany*

[6] *Department of Hematology/Oncology, University Hospital, Frankfurt, Germany*

[7] *Division of Immunology/Hematology, University Children's Hospital, Zürich, Switzerland*

[8] *Institute for Biomedical Research, Georg-Speyer-Haus, Frankfurt, Germany*

[9] *INSERM, Unit 768, Hôpital Necker and Faculté de Médecine Université René Descartes Paris V, Paris, France;*

[10] *Département de Biothérapies, Hôpital Necker, Paris, France*

[11] *Unité d'Immunologie et d'Hématologie Pédiatriques, Hôpital Necker-Enfants Malades, Paris, France*

[12] *Department of Medical Biometry, University of Heidelberg and Tumor Center Heidelberg/ Mannheim, Heidelberg, Germany*

[13] *IIT unit of Molecular Neuroscience, Instituto Scientifico H. San Raffaele, Milan, Italy*

[14] *Department of Biomedical Sciences, University of Modena and Reggio Emilia, Modena, Italy*

[15] *San Raffaele Telethon Institute for Gene Therapy (HSR- TIGET), Milano, Italy*

[16] *Università di Roma Tor Vergata, Rome, Italy*

[17] *Hematology Branch, National Heart, Lung and Blood Institute, Bethesda, MD 20892, USA*

[18] *Department of Clinical Immunology, Great Ormond Street Hospital NHS Trust, London, United Kingdom*

ABSTRACT

Vector associated side effects in clinical gene therapy have provided new insights into the molecular mechanisms of hematopoietic regulation *in vivo*. Surprisingly, many retrovirus insertion sites (RIS) present in engrafted cells have been found to cluster non-randomly in close association with specific genes. Our data demonstrated that these genes directly influence the *in vivo* fate of hematopoietic cell clones. Analysis of RIS thus far has been limited to each particular study. Here, we studied more than 7000 RIS across multiple species and studies. More than 40% of all insertions found in engrafted gene-modified cells were located in few common integration loci representing less than 0.36% of the genome.

Gene classification analyses displayed significant over-representation of genes associated with hematopoietic functions and relevance for cell growth and survival *in vivo*. Association was not detectable before transplantation, indicating that engraftment influences clonal selection. The similarity of RIS distributions indicates that vector insertion in repopulating cells clusters in predictable patterns. Based on this analysis, RIS analyses of preclinical *in vitro* and murine *in vivo* studies as well as prospective analyses of vector insertion repertoires in clinical trials should be excellent tools for modeling and predicting the distribution and biological impact of insertional behavior.

INTRODUCTION

Integrating gamma-retroviral vectors have the demonstrated ability to successfully treat life-threatening diseases, as convincingly shown by correction of the genetic defect and amelioration of clinical manifestations in X-linked and *ADA*-deficient severe combined immunodeficiencies (SCID), and in chronic granulomatous disease (X-CGD) (1-5). Unfortunately, 5 (6,7) of 20 treated X-SCID patients developed clonal acute T cell lymphoproliferative disorders and 2 of 4 treated CGD patients developed myelodysplasia (personal communication). In all cases, these complications were causally linked to insertional activation of proto-oncogenes, most strikingly *LMO2* in X-SCID, and *MDS1/EVI1* in X-CGD. As a result, there has been recent intense investigation of vector integration patterns in an attempt to understand the vector, target cell and patient variables that produce an increased risk of genotoxicity, and to design safer vectors and clinical protocols.

Integrated standard murine gamma-retroviral proviruses contain transgene cDNA flanked by strong retrovirus promoter and enhancer elements that frequently activate surrounding gene sequences, potentially resulting in dysregulated adjacent gene activation. Analysis of vector integration patterns in the SCID trials clearly demonstrated non-random and potentially detrimental integration effects (8). Clinical and experimental studies have further uncovered evidence of insertional influences ranging from subtle clonal selection to *in vitro* immortalization of myeloid cells (9) and clonal dominance (3,10,11) and subsequent leukemia in mice, non-human primates and human trial participants (6,12-14). We hypothesize that biological consequences of insertional mutagenesis are much more frequent than predicted, affecting neighboring genes in many transduced repopulating cell clones. Comparative analysis in large numbers of insertion sites should enable investigators to assess probable vector mutagenic effects prior to or in the absence of overt clonal dominance in animal models and clinical trials.

To clarify the scope of this problem and to uncover trends not apparent in smaller data sets, in the present report we analyze retrovirus integration site (RIS) profiles from 5 different clinical gene therapy studies and 3 preclinical models *in toto*, and in comparison with each other (3,8,15-20). Genome integration sites were analyzed with identical bioinformatics tools and aligned to the identical human or animal genome using NCBI BLAST tools. Species and study specific features were defined in relation to (i) genomic distribution of RIS before and after transplantation, (ii) relevance of common integration sites, (ii) vector targeted genes and their classifications using Gene Ontology and Ingenuity databases, and (iv) analogy and predictive potential of preclinical models for clinical applications.

MATERIAL AND METHODS

Sequencing and integration site analyses

LAM-PCR, LM-PCR, bacterial cloning and Sanger sequencing were performed as previously described (21,22). The RIS datasets included were from a CML gene marking study (15), two SCID-X1 clinical trials (8,16), a CGD clinical trial (3) an ADA-SCID clinical trial (19), a pre-transplantation dataset from human $CD34^+$ cells (20), a set of murine studies (17) and a rhesus macaque study (18). To verify data obtained using shotgun cloning of LAM-PCR products in bacteria, direct high throughput pyrosequencing of a single clinical trial was performed and compared to the results of the conventional cloning method. Comparisons were conducted using different cell fractions from all patients of the X-SCID, London, UK trial at variable time points. The RIS were determined by direct 454 sequencing (Roche) of LAM amplification products (23). To analyze different samples in parallel in one sequencing run, a fusion primer containing one of 24 distinct barcodes was used to amplify the conventional LAM-PCR product (24-26). The LTR-fusion primers and linker fusion primers (MWG Biotech, Ebersberg) were designed as recommended by Roche for amplicon sequencing and used at concentrations of 5.6 pmol.

30-100 ng of purified LAM-PCR products were used as template for fusion primer PCR. PCR was performed by initial denaturation 2 minutes at 95°C followed by 11 cycles of 95°C for 45 seconds, 60°C for 45 seconds and 72°C for 1 minute, terminated by a final extension of 5 minutes at 72°C. 5µl of PCR product were visualized on a 2% agarose gel, and DNA concentrations of each PCR product were quantified (Nanodrop Technologies). The barcoded PCR products were pooled and sequenced with the 454 GS FLX platform (Roche). In the context of the European 6th framework project (CONSERT), we have developed an integration site database (LAM-PCR Database) that can store, reanalyze and update the location and associated features of retroviral vector integrants. The software trims, aligns, locates and annotates the genomic human or murine RIS. The sequences obtained were analyzed by uploading the reads as a FastA format to the LAM-PCR Database[15]. We have used NCBI BLAST for the alignment of the sequences on human (Build 35) or mouse (Build 37) genome assembly. The RIS of the primate study have only been aligned to the human genome. For the analysis of the SCID sequences obtained by 454 (Roche) high throughput sequencing we used the LAM-PCR Database with NCBI genome Build 36.2. The sequences are available in the database. User name and password should be requested from the corresponding author. The BLAST results for each sequence are given in supplementary Table 1.

15 https://consert.gatc-biotech.com/lampcr/index.html

Definition of CIS

The determination of CIS has been performed as previously described (27). In brief, we have measured the distance between individual integrants independently of being located in or outside of gene coding regions. 2 retroviral integration sites (RIS) form a CIS if they fell within a 30 kbp window. We call such an integration region CIS of 2nd order. A CIS of 3rd order bears 3 RIS in a 50 kbp window and a CIS of 4th order 4 RIS in a 100 kbp window. For CIS of 5th order or higher we assumed a window of 200 kbp. These definitions imply that a CIS of higher order always contains at least one CIS of lower order.

Gene Ontology analysis (GO) and Ingenuity Pathway Analysis (IPA)

To classify vector targeted genes according to GO terms and/or their interactions, for each RIS we classified the closest RefSeq genes interrupted by the vector or within 10 kbp of the RIS. GO analysis was performed using the publicly available NIH-DAVID Bioinformatics resources[16] (28,29) and IPA analysis using a license for the IPA bioinformatics application, version 7.5[17]. In these analyses, each gene was scored only once, even if multiple RIS were located within the gene or within 10 kbp of the gene.

Biostatistics

Computer simulations were used to derive datasets of 10,000 synthetic random integration events for each analysis. Comparisons between the observed and expected properties of experimental versus random RIS were made. To assess the randomness in the occurrence of common integration sites (CIS), we applied mathematical models of CIS formation accounting for the number of RIS, size of the genome, known insertion preferences and other parameters as previously described (27).

The comparison of the number of CIS before transplant and after transplant was analyzed using a modified Monte Carlo approach which adjusts for the differences in the number of integration sites in the two datasets. In brief, let npre and npost be the numbers of RIS before and after transplantation, and let RISpre and RISpost be the set of the exact positions of these RIS, respectively. In our analysis it turned out that npre < npost. Random samples of size npre were drawn repeatedly (10000 simulation runs) from RISpost, and for each of these samples the numbers of CIS of order 2,...,8 were counted. This yielded simulated distributions (n=10000) of the number of CIS (of each order), with which the observed numbers of CIS in RISpre were then compared to obtain p-values. This comparison was not biased by the differences in the sample sizes npre, npost because the random samples drawn from RISpost contained the same numbers of RIS as RISpre.

16 http://david.niaid.nih.gov/david/ease.htm

17 www.ingenuity.com

The overrepresentation of gene categories affected by an insertion was examined using the output of the NIH-DAVID Bioinformatic resources software. Over-represented gene categories were determined by Fisher's exact test, which compares the proportion of genes having at least one RIS and belonging to a particular category, with the proportion of all genes that fall into this category. The results were adjusted for multiple testing using the Bonferroni correction of the p-values provided by the software. Note that this analysis presupposes that under the null hypothesis all genes have the same probability of being affected by a RIS. This is only an approximation, given that even in the special case of uniform distribution of the RIS this probability depends on the length of the genes.

Gene expression profiling

Umbilical cord blood (UCB) $CD34^+$ cells from 6 donors were isolated and pooled into 2 batches. They served as samples for RNA expression analysis. RNA was isolated using TriReagent (Sigma) following the manufacturer's protocol. The mRNA expression levels were determined using Affymetrix U133 Plus 2.0 arrays and normalized as described previously (30). The normalized microarray values were sorted upwardly on expression and divided into 10 equally sized expression level categories (0–9). The presence of the gene closest to a vector integration site as identified by LAM-PCR analysis was determined in each expression category. For every gene symbol, the highest expression value of the corresponding probesets was used to determine the expression of that gene.

RESULTS

Distribution of retroviral insertion sites (RIS) among the different studies

To compare the RIS of different studies with each other referring to the same annotation of the human or mouse genome, all sequences were imported as raw FASTA formatted sequence data. Out of 3863 exactly mappable RIS, 1316 and 2547 RIS were derived from preclinical and clinical samples, respectively. 2711 RIS were determined in mature circulating blood cells or bone marrow samples after transplantation (1984 in clinical samples) and 1152 were derived from $CD34^+$ cells present at the end of transduction, before transplantation (563 in clinical samples). 49% (1323 of 2711) of all RIS of post-transplantation samples and 53% (616 of 1152) of all pre-transplantation samples were located in the gene coding region of a RefSeq gene. When including the 10 kbp DNA region surrounding RefSeq genes that is probably equally relevant for insertional activation, we found roughly 3/4th (73%; 1979 of 2711) of all RIS in post-transplantation samples and 74% (850 of 1152) of pre-transplantation samples in or near a RefSeq gene. As expected from prior studies, a third of these integration sites were located within 10 kbp around the transcription start site of the gene (Table 1).

Table 1. Distribution of retroviral integration sites (RIS) detected in murine, primate, human preclinical and clinical studies

Results are shown separately for each individual study and summarized. The number of mappable RIS identified before and after transplantation in each study is given as absolute numbers. The proportion of RIS located in different genomic regions is given as percentage of the total number. Only Refseq genes are considered. CML, chronic myeloid leukemia; X-SCID, X-linked severe combined immunodeficiency; F, French X-SCID trial; UK, United Kingdom X-SCID trial; ADA-SCID, adenosine deaminase deficient severe combined immunodeficiency; I, Italian ADA-SCID trial; CGD, chronic granulomatose disease trial; $CD34^+$, human retroviral transduced $CD34^+$ cells; Primate, non-human primate study; Mouse, murine study; Total; summary of all analyzed trials; post, post-transplantation; pre, pre-transplantation samples which contain $CD34^+$ cells at the end of transduction, immediately before reinfusion; TSS, transcription start sites; kb, kilo base pairs; +/-, up-/downstream.

	CML	SCID F	SCID UK	CGD	ADA-SCID I	$CD34^+$	Primate	Mouse	Total
Mappable RIS	40	554	560	722	671	589	305	422	3863
RIS pre	-	96	265	-	202	589	-	-	1152
RIS post	40	458	295	722	469	-	305	422	2711
RIS in gene, post	58%	42%	46%	57%	43%	-	55%	45%	49%
RIS in gene, pre	-	57%	52%	-	50%	55%	-	-	53%
RIS gene region, post	73%	68%	70%	77%	71%	-	73%	73%	73%
RIS gene region, pre	-	75%	76%	-	68%	75%	-	-	74%
RIS TSS post	23%	33%	33%	29%	39%	-	27%	39%	34%
RIS TSS pre	-	28%	37%	-	28%	33%	-	-	33%

Presence of common integration sites (CIS) in human/primate pre- and post-transplantation samples

Comparative analysis of the 3441 RIS of the human/primate dataset including pre- and post-transplantation samples revealed that 45% (1547 RIS) were part of a CIS in at least one instance compared to 6.5% expected under a uniform random distribution of the RIS ($P<10^{-5}$). The proportion of RIS involved in CIS post-transplant (37%, 839 of 2289) was significantly higher than the corresponding proportion pre-transplant (20%, 232 of 1152) even after adjusting for the difference in RIS numbers between the pre- and post-transplant samples ($P<10^{-5}$). The degree of integration site clustering in the human/primate post-transplant samples showed a further substantial difference to the pre-transplant cells. Of all RIS in CIS post-transplant, 41% (340 of 839) were involved in a CIS of 4th or higher order whereas only 11% (25 of 232) of all RIS involved in CIS from pre-transplantation samples were located in CIS of 4th or higher order even after adjusting for the difference in RIS numbers between the pre- and post-transplant samples ($P<10^{-4}$). Most CIS are affected by retroviral integration not only in other individual patients, but in more than one study (Table 2). Even when eliminating the very redundant CIS at the EVI1/MDS1 locus from the CGD trial, we observed that 32% (724 of 2174) of all post-transplantation RIS were involved in CIS versus 20% of all pre- transplantation RIS ($P<10^{-4}$).

Table 2. Common integration sites (CIS) identified in post-transplantation samples of murine, primate and clinical studies.
A gene region was classified as a common integration site (CIS); CIS of 5th, 6th, 7th or higher order if 5, 6, 7, or more RIS were located in a 200 kb window independent of the study, in which the RIS were detected. The number of retroviral integration sites (RIS) contributing to a specific CIS, the corresponding study and the gene closest to the CIS are indicated. CIS locus, gene next to the CIS; CML, chronic myeloid leukemia; X-SCID, X-linked severe combined immunodeficiency; F, French X- SCID trial; UK, United Kingdom X-SCID trial; ADA-SCID, adenosine deaminase deficient severe combined immunodeficiency, Italian ADA-SCID trial; Primate, primate study; Mouse, murine study; numbers in brackets, murine integration sites near the same gene, due to different database versions direct comparison with human CIS is not possible.

	CIS Locus	CML	XSCID F	XSCID UK	CGD	ADA SCID	Primate	Mouse
CIS 5th order	PSMA6		2		1	2		
	LOC152225		2		2		1	
	BACH2		1		3		1	
	DYRK1A		1		1	3		
	ZNF217		2	1		2		
	PTPRC		3	1	1			
	ESRRBL1		3		1		1	
	ANGPT1		3		1		1	(1)
	LYL1		1		2	1	1	
	GSN		2		2		1	
	THUMPD1		2	1	1		1	
CIS 6th order	RUNX1		3	1		1	1	
	C14orf4		4		1	1		
	MN1				5		1	
	PDE4B		3			2	1	
	STAT3		3	1	2			
	BLC2L1		2	3				
CIS 7th order	BCL2		2	1		3	1	
	RBM34		2	2	2	1		
	MRPL36P1		3	2	1	1		
CIS higher than 7th order	PRDM16				36			(1)
	MDS1/EVI1		2		79	1	7	19
	LMO2		5		2	5	1	
	CCDN2		9			3		
	SETBP1	1			7		2	

Genes most frequently involved in CIS

In pre-transplantation samples, only one CIS of 5th order could be found around the gene FLJ10597, a zinc finger protein with unknown function. 4 of these RIS derived from 14 days expanded pre-transplantation cells and only one from cells directly after transduction. The other 22 RefSeq genes located in a CIS region of ≥ 5th order were found exclusively in post-transplantation samples. For 18 of these genes, evidence for direct involvement in tumor formation is available from previous mutagenesis studies (31). Of the 8 genes comprising CIS locations of ≥ 7th order (*BCL2*, *RBM34*, *PCBP1*,

PRDM16, *MDS1/EVI1*, *LMO2*, *CCND2* and *SETBP1*) all were known as cancer promoting genes. Of these 8 genes, 5 have been associated with overt clonal expansion in clinical trials (3,7) (Table 2). All human and non-human primate *in vivo* studies showed insertions affecting at least 3 CIS of 7th or higher order per 300 insertions studied.

Overrepresentation of distinct gene categories

We used gene ontology (GO) analyses to identify overrepresented functional gene categories within the RIS datasets, The combined GO analysis using all 9 available post-transplantation datasets yielded highly significantly overrepresented specific gene categories by Fisher's Exact test after Bonferroni correction, including regulation of cellular processes, protein kinase activities and regulation of cell death (Table 3). By contrast, GO analysis of the individual pre-transplantation datasets did not detect significantly overrepresented gene categories.

Network analysis using Ingenuity Pathway Analysis (IPA)

To define specific physiological functions and networks included in the RIS datasets, we performed comparative Ingenuity Pathway Analyses (IPA). Ingenuity analysis also offers further insight into gene sets associated with specific diseases. In pre- and post-transplantation samples, RIS were mainly found in genes involved in hematological system development and functions. Hematopoiesis related gene classes were the only physiological category significantly overrepresented in engrafted cells, i.e. post-transplantation samples

(Figure 1A).

Overrepresented "molecular function" gene categories in post-transplantation samples were "gene expression", "molecular growth and proliferation", "cell death", "cell cycle" and other cellular growth related categories. In CIS genes, over representation of these categories was most significant. The most significant disease gene categories in the RIS dataset were "cancer" followed by "immunological disease", "hematological disease" and "connective tissue disorders".

Gene expression profiling of genes next to the integration site

Highly expressed genes have been found to be preferential targets for both gamma-retroviral and lentiviral RIS in various *in vitro* and *in vivo* models (32,33). We carried out gene expression array analysis on umbilical cord blood (UCB) $CD34^+$ cells, based on the hypothesis that cord blood $CD34^+$ cells were likely to have a similar gene expression profile to the bone marrow $CD34^+$ target cells in children with X-SCID, CGD or other gene therapy target diseases. Obtaining sufficient $CD34^+$ cells from actual patients with these diseases or even from age-matched normal children to perform gene expression profiling is not feasible. We compared the full dataset of RIS within or located within

Table 3. Gene Ontology (GO) analysis
Genes of all studies with an insertion within the gene or the neighboring 10 kb were classified according to the biological process or the molecular function in which their proteins are involved. Results are divided in functional groups and classified regarding the enrichment score. System, indicates if the result is related with a biological process or the molecular function of the protein; Level, determines the specificity and coverage of the result. A low number indicates high coverage and low specificity and a high number indicates low coverage and high specificity; Gene category, specifies the gene class, in which the RIS were located; Count, number of genes of the corresponding gene class with an insertion hit; significance-values determined by Bonferroni adjusted Fisher's exact test.

System	Level	Gene Category	Count	p-value
Functional Group 1		**Enrichment Score 10.34**		
Biological Process	2	Regulation of cellular process	267	$1.4x10^{-10}$
	2	Regulation of physiological process	257	$5.2x10^{-9}$
	3	Regulation of cellular physiological process	250	$1.7x10^{-7}$
Functional Group 2		**Enrichment Score 6.67**		
Biological Process	4	Positive regulation of cellular physiological process	53	$5.0x10^{-5}$
	3	Positive regulation of cellular process	59	$4.4x10^{-5}$
	3	Positive regulation of physiological process	53	$9.8x10^{-5}$
	2	Positive regulation of biological process	64	$2.9x10^{-5}$
Functional Group 3		**Enrichment Score 6.34**		
Biological Process	4	Phosphotransferase activity, alcohol group as acceptor	79	$1.7x10^{-4}$
	4	Kinase activity	95	$1.0x10^{-4}$
	3	Transferase activity, transferring phosphorus-containing groups	101	$6.0x10^{-4}$
Functional Group 4		**Enrichment Score 6.18**		
Biological Process	2	Death	61	$2.3x10^{-5}$
	3	Cell death	61	$1.1x10^{-4}$
	4	Programmed cell death	59	$3.2x10^{-4}$
	5	Apoptosis	58	$2.5x10^{-6}$
Functional Group 5		**Enrichment Score 5.42**		
Biological Process	4	Regulation of programmed cell death	42	$1.1x10^{-3}$
	5	Regulation of programmed cell death	42	$3.4x10^{-3}$
	5	Regulation of apoptosis	41	$7.7x10^{-3}$
Functional Group 6		**Enrichment Score 4.68**		
Biological Process	3	Negative regulation of cellular process	68	$6.7x10^{-4}$
	2	Negative regulation of biological process	70	$3.4x10^{-4}$
	4	Negative regulation of cellular physiological process	58	$4.0x10^{-2}$
	3	Negative regulation of physiological process	58	$5.5x10^{-2}$

10 kbp of the retroviral integration site to gene expression of these sites in the UCB samples. Retrovirus integrations were found over-represented in genes expressed at high but not the highest expression levels. Genes located in CIS regions are expressed at a higher level than the average of all RIS, but genes involved in the CIS of highest order were expressed less than the average of the other CIS categories (Figure 2).

Predictive value of non-human preclinical data for clinical studies

To study potential overlaps between human and mouse integration sites and to determine possible CIS in homologous human and murine genomic regions, shared between both datasets, we translated the mouse gene names into human gene names with aid

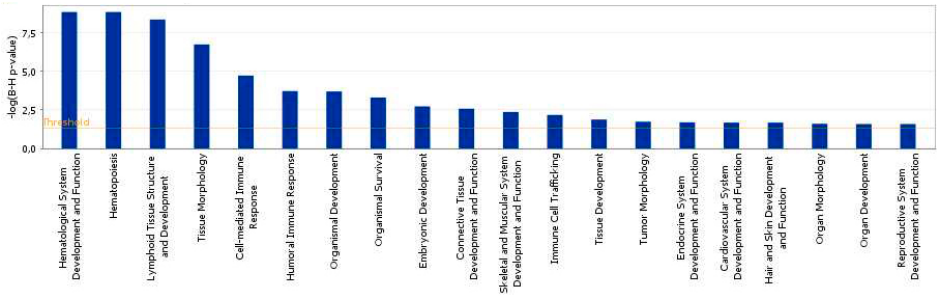


Figure 1. Ingenuity pathway analysis (IPA). Genes of post-transplantation samples of all studies with an insertion within the gene or in the neighboring 10kb were classified according to physiological function. The x-axis indicates the category to which the analyzed genes contribute. We show only the significant categories. . For each analyzed gene group the statistical significance (Bonferroni corrected Fisher's exact test) of overrepresented genes in a pathway is given on the y-axis.

of the NCBI Matchminer (34) and the clone ID converter (35). 104 RIS (25%) of murine integrations were located within a 100 kbp region of human and non-human primate genes (pre and post) harboring RIS. When we determined the CIS separately in the mouse dataset we could show that 22% (92 RIS) of all integration sites were located in CIS. 19 RIS in mice could be identified in the EVI1 gene locus (supplement table 2). Of all mouse RIS, 44 (10%) were located in or near a gene of a human/non-human primate CIS. Insertions from primate models were looked up directly in the human database to allow the direct comparison of the integration sites between human and primates. This is possible due to the high degree of identity between these two genomes. The recent publication of the rhesus genome also now allows direct mapping of RIS to the actual species genome, however for comparisons with human RIS, use of the human genomic sequence allows direct comparisons and better annotations of genes and other genomic features, which is much less fully developed in the rhesus draft genomic sequence. The comparison of retroviral integration sites between non-human preclinical and human clinical studies showed a large overlap. 77 RIS (25%) of the primate study could be localized in human CIS including EVI1 as the only CIS locus ≥ 5th order. In primates, 7 RIS were located in this genomic region.

Human *in vivo* insertion inventories with pyrosequencing technology

To investigate whether standard LAM-PCR analysis followed by shotgun cloning of PCR products in bacteria identifies the full spectrum of RIS efficiently, we reanalyzed materials from the British XSCID study with new high throughput pyrosequencing technology on samples derived from all 11 patients at different time points after transplantation. Sequencing reactions were carried out in 5 different runs with the 454 GS FLX platform (Roche). In

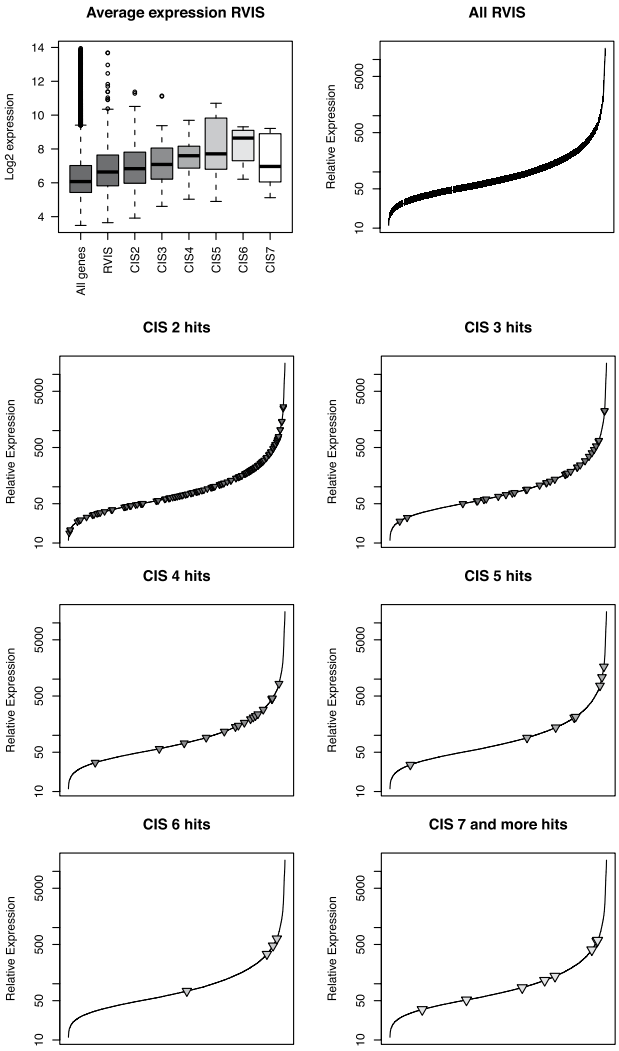


Figure 2. Comparison of CIS occurrence and gene expression in human target cells. (A) The log2 expression values of all genes on the U133 plus 2.0 microarray chip compared to the genes with a retroviral integration site (RIS) within 10 kb and to the genes involved in a common integration site (CIS) of different orders. (B) The expression values of the genes with an insertion in or in the neighboring 10kb identified in post-transplant samples from all studies are plotted. The relative expression level of genes with nearby RIS is indicated as blue vertical lines on a line representing the relative expressions for all genes retrieved from the U133plus2 microarray. The x-axis indicates the genes present on the U133plus2 microarray chip. The y-axis determines the relative expression level. (C-H) The gene expression values of genes identified as CIS are plotted and shown separately according to number of hits. The expression level of the genes involved in CIS is indicated as blue triangles on a line representing the relative expression of all genes retrieved from the U133plus2 microarray. The x-axis indicates the genes present on the U133 plus 2.0 microarray chip. The y-axis indicates the relative expression level. All genes, expression level of all genes on the U133 plus 2.0 microarray chip; CIS 2, 3, 4, 5, 6, or 7, common integration site of 2nd, 3rd, 4th, 5th, 6th, or 7th order; CIS 2, 3, 4, 5, 6, 7 hits, CIS of 2nd, 3rd, 4th, 5th, 6th, or 7th order.

total we collected 224,457 reads. After aligning the sequences against the human genome assembly (Build 36.2) via NCBI BLAST and removing short sequences, sequences with incomplete or without LTR sequences and double sequences, we obtained 3625 mappable RIS. The RIS distribution per patient, at specific time intervals relative to transplantation and in specific cell lineages is shown in Table 4. Of these RIS, 483 (13%) could be detected

Table 4. Pyrosequencing results of the British X-SCID study. Shown are the number of sequence reads obtained, and the unique integration sites then identified on each sample. The samples differ in days after transplantation and analyzed cell type.

Patient	Days after transplantation	Cell fraction	Total 454 reads	Unique RIS per sample
1	430	CD3	7754	158
	588	CD3	14114	522
	2001	Neutro	5698	10
	2001	PBMCs	12438	550
2	292	CD3	9525	12
	404	CD3	5727	40
3	159	CD3	5005	44
	259	CD3	8180	83
	488	CD3	2228	66
	793	CD3	5246	44
	1581	PBMCs	11112	123
4	357	CD3	2332	20
	1693	PBMCs	12269	267
5	49	PBMCs	5694	11
6	236	CD3	5440	61
	1091	CD19	1414	1
	1091	CD3	9573	9
	1091	CD56	1902	3
	1091	Gran	2054	0
	1091	Mono	65	2
7	84	CD3	2527	2
8	180	CD3	7199	19
	362	CD3	3159	10
	539	CD3	2555	51
	717	PBMCs	10533	206
	Post chemo	PBMCs	7054	600
9	89	PBMCs	9910	203
	166	CD3	5107	25
	166	CD19	269	1
	166	Gran	322	0
	166	Mono	514	0
	545	PBMCs	13816	307
10	215	CD19	191	1
	215	CD3	2385	11
	215	CD56	8870	12
	215	Gran	515	3
	215	Mono	77	0
	441	PBMCs	8154	105
11	1219	PBMCs	13530	43
Total			224457	3625

Table 5. Presence of cancer genes near integration sites. The sample "all RIS" contains all closest genes in the 10 kb surrounding of a retroviral integration site (RIS). CGC: Cancer gene census; this database of the Cancer gene project contains 384 human cancer genes; RTCGD: Retrovirus Tagged Cancer Gene Database; this database contains 3381 mouse genes affected by retroviral integration.

Sample	Genes in dataset	CGC genes found	% of CGC genes	RTCGD genes found	% of RTCGD genes
All RIS	1628	63	4%	450	28%
All CIS	505	34	7%	192	38%
CIS 2nd order	292	10	3%	92	32%
CIS 3rd order	88	5	6%	31	35%
CIS 4th order	44	6	14%	23	52%
CIS 5th order	26	3	12%	14	54%
CIS 6th order	24	2	8%	12	50%
CIS ≥ 7th order	31	8	26%	20	65%

at more than one time point or in more than one lineage. If we subtract these double RIS we identified 3142 unique RIS (supplement table 3). We compared the 3142 pyrosequenced RIS with the 295 RIS identified using Sanger sequencing on post-transplantation samples samples from the same patients. 77 RIS (26%) identified by Sanger sequencing were also identified via 454 pyrosequencing technology, suggesting that the pyrosequencing approach yields similar results to Sanger sequencing. As not all of the original LAM-PCR products could be resequenced by 454 because of lack of material, we could not carry out a formal comparison of the relative efficiency of each method. However, the large number of RIS identified via high throughput sequencing impressively demonstrates the extent of the clonal inventory. 404 RIS (13%) of the pyrosequencing samples and 230 (15%) RIS involved in CIS could be detected more than once at different time intervals.

Concerning the occurrence of common integration sites in these pyrosequenced samples of one trial, we detected 1560 RIS (50%) in such CIS compared to 187 (6%) expected under a uniform distribution ($P<10^{-5}$). If we examined only the RIS located in CIS, we found 710 (46%) in CIS ≥ 4th order, 534 (34%) in CIS ≥ 5th order and 404 (26%) in CIS ≥ 6th order. Among the 505 genes involved in CIS, 34 (7%) are listed in the human cancer gene census database[18] and 192 (38%) in the mouse Retrovirus Tagged Cancer Gene Database (RTCGD) (36). Higher order CIS have a higher percentage of genes listed in these two cancer gene databases. In CIS ≥7th order, 8 (26%) of the 31 genes are listed in the CGC and 20 (65%) in the RTCGD database (Table 5). The genes involved in CIS ≥ 7th order are listed in table 6. 20 of the 31 genes were CIS genes also in the comparative analysis, and 10 also occurred in the British Sanger sequenced samples (Table 6). Ingenuity analysis of the pyrosequencing samples confirms the results of the comparative meta-analysis. Gene expression profiles of the closest genes show comparable results to the combination analysis of clinical trials (supplement Figure 1). In the higher order CIS we still see some spread into lower and higher expression levels. CIS were not uniformly associated with high expression levels.

18 http://www.sanger.ac.uk/genetics/CGP/Census/

Table 6. Common integration sites (CIS) of 454 X-SCID samples and comparison with Sanger CIS. A gene region was classified as CIS of 7th, 8th or 9th order if 7, 8 or 9 RIS were located in a 200 kb window. Shown are all genes of the pyrosequencing analysis of the British X-SCID samples in CIS regions composed of 7 or more RIS. The column "CIS in comparative analysis" shows, in which order of the comparative study the gene has been found, the last column shows how many RIS contributing to these CIS were found in the original British X-SCID analysis performed via Sanger sequencing.

CIS order	Gene	CIS order in comparative analysis	X-SCID UK Sanger sequencing
7th	PTPRC	5th	1 RIS
	ANKRD44		
	MAML3		
	PIM1	2nd	1 RIS
	ANGPT1	5th	
	MRVI1		
	ZNRF1	2nd	
	PRKCBP1	4th	3 RIS
	PLCB4	2nd	
	RUNX1	6th	1 RIS
	C20orf94		
8th	ENSA		
	MRPL36P1	7th	1 RIS
	FOXP1	2nd	
	STK10	2nd	1 RIS
	TRIO		
	NINJ2		
9th	LOC283551	3rd	1 RIS
≥10th	TOMM20		
	MLLT3	4th	1 RIS
	HMGA2	2nd	1 RIS
	BCL2L1	4th	2 RIS
	MDS1	≥9th	
	LMO2	≥9th	
	ETV6	2nd	
	ZNF217	2nd	
	PACSIN2		
	BTN3A1		
	PSMA6	5th	
	GPR97		
	PRKCB1	2nd	

DISCUSSION

Because of the limited number of subjects in clinical phase I trials, and the small animal hosts of preclinical gene therapy models, single trials can only furnish small clone numbers and limited statistical evaluation on how intensively the integration site preferences of retrovirus vectors and their clonal selection *in vivo* influences functional integrity and

fate of the targeted cells. Differences in annotations of the mammalian genomes used for mapping, differences in parameters used to define and map "true" integrations and differences in bioinformatic and statistical approaches to data analysis have thus far prevented true comparative evaluation of separate studies. We reasoned that comparing and pooling of data from all available RIS datasets of important clinical and preclinical gene therapy trials should allow to significantly extend the insight into the integration preferences of retroviral vectors, and relate the likelihood of clonal dominance to the types of insertions that can be found. We performed a comparative bioinformatical analysis that overcomes technical limitations by realigning the raw sequence data of all relevant preclinical and clinical large gamma-retroviral RIS datasets resulting from transduction of primitive hematopoietic stem and progenitor cells (HSPC) to an identical build of the species' genome. The combined view of the resulting large information set encompasses data from 24 individual patients, 1 normal donor, 142 mice, 23 primates and 1 *in vitro* set of transduced cells with a combined content of 3863 insertions. Retrovirus vector insertions were not at all randomly distributed, and had far more influence on the biological fate of engrafted cells than had initially been anticipated.

Across all datasets, the affinity of the viral insertion repertoire to genes was quite surprising. When analyzing both human and non-human primate datasets obtained from post-transplant samples, we found that approximately 75% of RIS were located in or within 10 kbp of RefSeq genes. A similar frequency of integrants located in and/or near RefSeq genes also held true in pre-transplant samples, reflecting a very high affinity of gamma-retroviruses for these genomic regions (37). Common insertion sites (CIS) affecting the same genomic region in different cells of the same or different individuals provide a simple but highly effective statistical tool to demonstrate non-randomness in the distribution of insertions. CIS allow to determine possible sites of functional deregulation in a background of sites without such effects (27,37,38). With a larger group of insertions, the increased likelihood of detecting those integrations closer to each other has to be distinguished from true CIS.

Event with this statistical correction, CIS analysis unveils several very strong biological consequences of vector biology. The frequency of CIS was much higher than can be statistically anticipated based on the gene preference of MLV derived vectors alone. Our analysis of the pooled datasets indicates that more than 40% of gamma-retroviral RIS in engrafted hematopoietic cells are in common genomic regions. CIS regions were predominantly identical between independent studies, even when conducted in different species, with their overall target area representing a very small fraction of the entire genome. The majority of frequently affected ("higher order") CIS genes are known to promote cellular transformation from the RTCGD and CGP cancer gene databases (36), and have been found activated by gamma-retroviral insertion in preclinical (14,39,40) and clinical studies (3,6,41).

In pretransplantation samples, the proportion of 2nd order CIS was more than 10-fold higher than the expected random value calculated. Assuming that it is unlikely that short term transduction culture selects growth-impacting RIS in only 72-96 hours (39), the number of pretransplantation CIS may point to a preference for specific gene regions already at the time of transduction, which could be due to differences in the accessibility of genes to vector integration or due to the presence of transcription factors that facilitate integration into these specific genomic areas (20). Regulatory viral LTR elements present in the pre-integration complex can bind a variety of transcription factors (42-44) and may preferentially target vector integration to specific elements present in gene coding regions, regulatory regions and near transcription start sites.

When we scored CIS occurrence *in vivo* over time, the development of overrepresentation in particular CIS made evident that the presence of a vector insert in CIS strongly influences clonal expansion. After engraftment, the number of CIS of 2nd order increases to more than 35-fold higher than expected. For CIS of >4th order, the observed value is more than one- million-fold higher compared to the expected value. Much speculation has been published that highly overrepresented insertion loci might be related to a particular underlying clinical condition, a vector design feature or transgene effects (3,8,13,16). Our data do not fully support this hypothesis. We found that prominent CIS (i.e. of 5th to 7th orders) are targeted by gamma-retroviral integration in many data sets, indicating that the mechanism of selection is not restricted to particular constellations in single trials. However, some of these studies might share characteristics relevant for the interpretation of this finding. These include a selective advantage resulting from transgene expression at different levels, different types of immunodeficiency that influence bone marrow and immune function, and specific vector elements.

If CIS are a statistical indicator of clonal *in vivo* selection, we hypothesized, ranking the most frequently encountered CIS should predict which loci most likely produce clinically symptomatic manifestations of insertional mutagenesis (36). Strikingly, the top five most frequent CIS are exactly those loci associated with clinically-relevant adverse insertional side effects that have occurred in clinical retrovirus gene therapy: *LMO2*, *MDS1/EVI1*, *PRDM16*, *SETBP1* and *CCND2*. Moreover, *MDS1/EVI1* is definitely a preferred integration region in mice and primates. These frequent CIS also show the most pronounced difference in occurrence before and after transplantation. Ongoing clonal selection after engraftment demonstrates the power of biological effects resulting from vector integration. Not a single of the higher order CIS detectable after engraftment *in vivo* affected a RIS detectable in the pre-transplant samples, indicating a strong disparity between the biology of *in vitro* integration preferences in a predominantly non-engrafting progenitor cell population and the biology underlying insertion, engraftment and then selection of *in vivo* repopulating stem cell compartment.

While the development of a leukemia depends on secondary events, which are not yet well understood (12,40,45), our data aptly demonstrate that functional effects of mutagenesis in the CIS gene regions as a first event is not related to single, happenstance insertions into particular codons. Insertion locations in a particular CIS often differ by tens of thousands of base pairs. Therefore, their biological effects are not likely caused by direct mutagenic effects on the primary sequence, but rather result from changing the resident gene's activity by enhancer interference, trans-splicing or incapacitation of the original gene's transcript. For a stratification of CIS according to gene function, involved gene loci were assessed by Gene Ontology (GO) analysis and Ingenuity Pathway Analysis. GO analysis revealed significantly overrepresented gene categories in post-, but not in pre-transplant samples. These genes involved in regulation of cellular processes or cell death. The proteins are likely to involve kinase- or transferase activity. This difference between pre- and post-transplant samples is further corroboration that the biological function of the gene adjacent to the integration locus is responsible for cell fate and *in vivo* selection. Ingenuity analysis revealed that vectors integrate in genes involved in physiological functions of the hematopoietic system, a preferred integration already evident before transplantation. Therefore, this feature is likely characteristic of the retroviral integration process related to the gene expression pattern present in the hematopoietic target cells at transduction (46).

Integration in expressed genes was indeed highly overrepresented when comparing expressed genes in umbilical cord blood (UCB) $CD34^+$ cells to RIS in our datasets. The expression levels of CIS genes were not necessarily in the highest categories. In non-transduced target cells, genes of the most frequent CIS (i.e. of highest orders) expressed even less than other CIS genes, emphasizing that despite pre-existing gene expression, additional gene activation by insertion could indeed still be a decisive selection determinant in CIS.

Our comparative analysis elucidates a surprising degree of primary and selected retroviral integration preferences. Depending on the function of the activated genes, *in vivo* selection of clones following engraftment is substantial. These observations are very important for further gene therapy trials with retroviral and other insertional vectors. Obviously, this feature may positively influence the therapeutic success of gene therapy by favoring vector-containing clones over non-transduced cells, but may come at the expense of an increased incidence of progression to abnormal clonality, overt myelodysplasia or leukemia.

Quantitative determination of CIS is a valuable tool for the identification of the genomic context in which adverse events are more likely to occur (39). Prior to the clinical utilization of new vectors, it appears both necessary and feasible to assess safety with a large scale RIS analysis from preclinical *in vitro* and *in vivo* models using high throughput screening of transduced cells. The complete congruence between the most

frequent and the most troublesome CIS shows that prospective insertion site repertoire studies would allow detection of potentially risky sites and predicting the likelihood of clonal changes related to vector activity before these become symptomatic as serious adverse events. An acceleration of the integration site analysis procedure could in future result in a prospective analysis of the genomic integration site distribution by identifying relevant loci before transplantation of the gene modified cells in clinical trials. Avoiding the reinfusion of clones with insertions in the most frequent CIS locations might be all it would take to render the likelihood of insertional side effects very low. CIS might also help to define genes whose temporary expression can foster engraftment and expansion of stem cells in a clinical setting.

Acknowledgments

The authors acknowledge Drs. F.J.T. Staal and K. Pike-Overzet, Dept. Immunology, Erasmus University Medical Center, for assistance in the UCB array analyses.

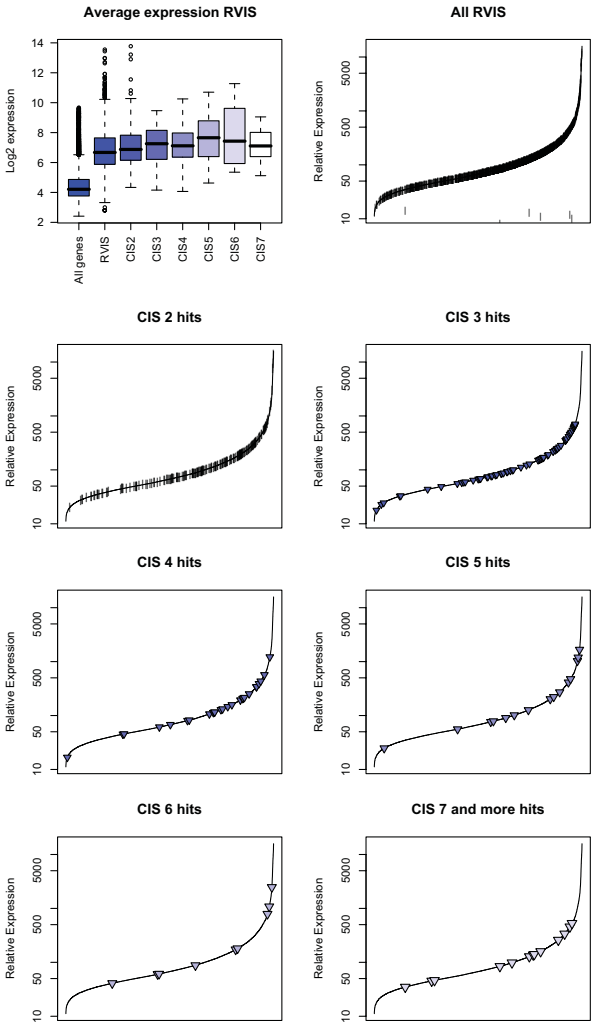
REFERENCES

1. Cavazzana-Calvo M, Hacein-Bey S, de Saint Basile G, et al. Gene therapy of human severe combined immunodeficiency (SCID)-X1 disease. Science. 2000;288:669-672.
2. Gaspar HB, Parsley KL, Howe S, et al. Gene therapy of X-linked severe combined immunodeficiency by use of a pseudotyped gammaretroviral vector. Lancet. 2004;364:2181-2187.
3. Ott MG, Schmidt M, Schwarzwaelder K, et al. Correction of X-linked chronic granulomatous disease by gene therapy, augmented by insertional activation of MDS1-EVI1, PRDM16 or SETBP1. Nat Med. 2006;12:401-409.
4. Seger R SU, Reichenbach J, Notheis G, Wintergerst U, et al. Immediate clinical benefit, but variable long-term correction of X-linked CGD by gene therapy in children. Human Gene Therapy. 2008;19:1097.
5. Aiuti A, Cattaneo F, Galimberti S, et al. Gene therapy for immunodeficiency due to adenosine deaminase deficiency. N Engl J Med. 2009;360:447-458.
6. Hacein-Bey-Abina S, von Kalle C, Schmidt M, et al. A serious adverse event after successful gene therapy for X-linked severe combined immunodeficiency. N Engl J Med. 2003;348:255-256.
7. Hacein-Bey-Abina S, Von Kalle C, Schmidt M, et al. LMO2-associated clonal T cell proliferation in two patients after gene therapy for SCID-X1. Science. 2003;302:415-419.
8. Deichmann A, Hacein-Bey-Abina S, Schmidt M, et al. Vector integration is nonrandom and clustered and influences the fate of lymphopoiesis in SCID-X1 gene therapy. J Clin Invest. 2007;117:2225-2232.
9. Du Y, Jenkins NA, Copeland NG. Insertional mutagenesis identifies genes that promote the immortalization of primary bone marrow progenitor cells. Blood. 2005;106:3932-3939.
10. Calmels B, Ferguson C, Laukkanen MO, et al. Recurrent retroviral vector integration at the Mds1/Evi1 locus in nonhuman primate hematopoietic cells. Blood. 2005;106:2530-2533.
11. Kustikova O, Fehse B, Modlich U, et al. Clonal dominance of hematopoietic stem cells triggered by retroviral gene marking. Science. 2005;308:1171-1174.
12. Li Z, Dullmann J, Schiedlmeier B, et al. Murine leukemia induced by retroviral gene marking. Science. 2002;296:497.
13. Woods NB, Bottero V, Schmidt M, von Kalle C, Verma IM. Gene therapy: therapeutic gene causing lymphoma. Nature. 2006;440:1123.

14. Seggewiss R, Pittaluga S, Adler RL, et al. Acute myeloid leukemia is associated with retroviral gene transfer to hematopoietic progenitor cells in a rhesus macaque. Blood. 2006;107:3865-3867.
15. Glimm H, Schmidt M, Fischer M, et al. Efficient marking of human cells with rapid but transient repopulating activity in autografted recipients. Blood. 2005;106:893-898.
16. Schwarzwaelder K, Howe SJ, Schmidt M, et al. Gammaretrovirus-mediated correction of SCID-X1 is associated with skewed vector integration site distribution *in vivo*. J Clin Invest. 2007;117:2241-2249.
17. Kustikova OS, Geiger H, Li Z, et al. Retroviral vector insertion sites associated with dominant hematopoietic clones mark "stemness" pathways. Blood. 2007;109:1897-1907.
18. Hematti P, Hong BK, Ferguson C, et al. Distinct genomic integration of MLV and SIV vectors in primate hematopoietic stem and progenitor cells. PLoS Biol. 2004;2:e423.
19. Aiuti A, Cassani B, Andolfi G, et al. Multilineage hematopoietic reconstitution without clonal selection in ADA-SCID patients treated with stem cell gene therapy. J Clin Invest. 2007;117:2233-2240.
20. Cattoglio C, Facchini G, Sartori D, et al. Hot spots of retroviral integration in human CD34+ hematopoietic cells. Blood. 2007;110:1770-1778.
21. Kustikova OS BC, Fehse B. Retroviral integration site analysis in hematopoietic stem cells. Methods Mol Biol. 2008;430:255-267.
22. Schmidt M SK, Bartholomae C, Zaoui K, Ball C, Pilz I, Braun S, Glimm H, von Kalle C. High-resolution insertion-site analysis by linear amplification-mediated PCR (LAM-PCR). Nat Methods. 2007;4:1051-1057.
23. Margulies M, Egholm M, Altman WE, et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature. 2005;437:376-380.
24. Parameswaran P, Jalili R, Tao L, et al. A pyrosequencing-tailored nucleotide barcode design unveils opportunities for large-scale sample multiplexing. Nucleic Acids Res. 2007;35:e130.
25. Meyer M, Stenzel U, Hofreiter M. Parallel tagged sequencing on the 454 platform. Nat Protoc. 2008;3:267-278.
26. Bushman FD. Retroviral integration and human gene therapy. J Clin Invest. 2007;117:2083-2086.
27. Abel U, Deichmann A, Bartholomae C, et al. Real-time definition of non-randomness in the distribution of genomic events. PLoS ONE. 2007;2:e570.
28. Dennis G, Jr., Sherman BT, Hosack DA, et al. DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol. 2003;4:P3.
29. Hosack DA, Dennis G, Jr., Sherman BT, Lane HC, Lempicki RA. Identifying biological themes within lists of genes with EASE. Genome Biol. 2003;4:R70.
30. Dik WA, Pike-Overzet K, Weerkamp F, et al. New insights on human T cell development by quantitative T cell receptor gene rearrangement studies and gene expression profiling. J Exp Med. 2005;201:1715-1723.
31. Wu X, Luke BT, Burgess SM. Redefining the common insertion site. Virology. 2006;344:292-295.
32. 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.
33. Recchia A, Bonini C, Magnani Z, 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 U S A. 2006;103:1457-1462.
34. Bussey KJ, Kane D, Sunshine M, et al. MatchMiner: a tool for batch navigation among gene and gene product identifiers. Genome Biol. 2003;4:R27.
35. Alibes A, Yankilevich P, Canada A, Diaz-Uriarte R. IDconverter and IDClight: conversion and annotation of gene and protein IDs. BMC Bioinformatics. 2007;8:9.
36. Akagi K, Suzuki T, Stephens RM, Jenkins NA, Copeland NG. RTCGD: retroviral tagged cancer gene database. Vol. 32; 2004:D523-527.
37. de Ridder J, Uren A, Kool J, Reinders M, Wessels L. Detecting statistically significant common insertion sites in retroviral insertional mutagenesis screens. PLoS Comput Biol. 2006;2:e166.
38. Fehse B, Roeder I. Insertional mutagenesis and clonal dominance: biological and statistical considerations. Gene Ther. 2008;15:143-153.

39. Modlich U, Kustikova OS, Schmidt M, et al. Leukemias following retroviral transfer of multidrug resistance 1 (MDR1) are driven by combinatorial insertional mutagenesis. Blood. 2005;105:4235-4246.
40. Baum C, Dullmann J, Li Z, et al. Side effects of retroviral gene transfer into hematopoietic stem cells. Blood. 2003;101:2099-2114.
41. Howe SJ, Mansour MR, Schwarzwaelder K, et al. Insertional mutagenesis combined with acquired somatic mutations causes leukemogenesis following gene therapy of SCID-X1 patients. J Clin Invest. 2008;118:3143-3150.
42. Ariumi Y, Serhan F, Turelli P, Telenti A, Trono D. The integrase interactor 1 (INI1) proteins facilitate Tat-mediated human immunodeficiency virus type 1 transcription. Retrovirology. 2006;3:47.
43. Frith MC, Fu Y, Yu L, Chen JF, Hansen U, Weng Z. Detection of functional DNA motifs via statistical over-representation. Nucleic Acids Res. 2004;32:1372-1381.
44. Felice B, Cattoglio C, Cittaro D, et al. Transcription factor binding sites are genetic determinants of retroviral integration in the human genome. PLoS ONE. 2009;4:e4571.
45. Gilliland DG, Tallman MS. Focus on acute leukemias. Cancer Cell. 2002;1:417-420.
46. Mitchell RS, Beitzel BF, Schroder AR, et al. Retroviral DNA integration: ASLV, HIV, and MLV show distinct target site preferences. PLoS Biol. 2004;2:E234.

SUPPLEMENTARY DATA



Supplement Figure 1: Comparison of CIS occurrence and gene expression in the British X-SCID study after pyrosequencing. (A) The log2 expression values of all genes on the U133 plus 2.0 microarray chip compared to the genes with a retroviral integration site (RIS) within 10 kb and the genes involved in a common integration site (CIS) of different orders. (B) The expression values of the genes with an insertion in or in the neighboring 10kb identified in post transplantation samples from all studies are plotted. The relative expression level of genes with nearby RIS is indicated as blue vertical lines on a line representing the relative expressions for all genes retrieved from the U133plus2.0 microarray. The x-axis indicates the genes present on the U133 plus 2.0 microarray chip. The y-axis determines the relative expression level. (C-H) The gene expression values of genes identified as CIS are plotted and shown separately according to number of hits. The expression level of the genes involved in CIS of a certain order is indicated as blue triangles on a line representing the relative expressions for all genes retrieved from the U133 plus 2.0 microarray. The x-axis indicates the genes present on the U133 plus 2.0 microarray chip. The y-axis indicates the relative expression level. All genes, expression level of all genes on the U133 plus 2.0 microarray chip; CIS 2, 3, 4, 5, 6, or 7 and more, common integration site of 2nd, 3rd, 4th, 5th, 6th, or >7th order.

Supplement Table 1: Table of insertion sites (RIS) of Sanger sequenced samples. Given are the insertion sites and their location to the next Refseq gene. NCBI RIS locus: Exact gamma-retroviral integration site; TSS: Transcription start site; 5' TSS [bp]: Distance to the TSS in bp if the RIS is located outside the gene at its 5 prime end; 3' TSS (in gene) [bp]: Distance to the TSS in bp if the RIS is located within the gene; 3' Gene [bp]: Distance to the Transcription stop site if the RIS is located outside the gene at its 3' end; Time: Time point of sample collection in weeks (w), days (d) or years (y), after transplantation (post) or before transplantation (pre). (Due to its size, supplemental table 1 is provided digitally)

Supplement Table 2: Identical integration regions between mouse and human/primate dataset. Given are the mouse insertion sites which correspond to one or more RIS in areas up to 100 kb around the same gene of the human / non-human primate dataset. NCBI RIS locus: Exact gamma-retroviral integration site; Identical region with human/primate: Indicated is whether the gene is found only once in the human/non-human primate dataset (RIS) or to which kind of CIS it belongs. pre: RIS from pre-transplantation samples; post: RIS from post-transplantation samples. (Due to its size, supplemental table 2 is provided digitally)

Supplement Table 3: Table of insertion sites (RIS) of 454 sequenced SCID UK samples. Given are the insertion sites and their location to the next Refseq gene. NCBI RIS locus: Exact gamma-retroviral integration site; TSS: Transcription start site; 5' TSS [bp]: Distance to the TSS in bp if the RIS is located outside the gene at its 5 prime end; 3' TSS (in gene) [bp]: Distance to the TSS in bp if the RIS is located within the gene; 3' Gene [bp]: Distance to the Transcription stop site if the RIS is located outside the gene at its 3' end; Source: Cell samples for the determination of RIS; CD3: early T-cells; Neutros: neutrophil granulocytes; PBMC: peripheral blood mononuclear cells; CD19: B lymphocytes; CD56: T cells and natural killer cells; Mono: mononuclear cells, Gran: granulocytes; Time: Timepoint of sample collection in days or after chemotherapy (post-chemo). (Due to its size, supplemental table 3 is provided digitally)

CHAPTER 5

Characteristics of gamma-retrovirus integration-related leukemias in mice

Adapted from: Martijn H. Brugman[1,2], Trudy P. Visser[1], Shazia P. Arshad[1], Carla Oerlemans-Bergs[1], Monique M.A. Verstegen[1] and Gerard Wagemaker[1], Characteristics of gamma-retrovirus integration-related leukemias in mice. *Submitted for publication.*

[1]Department of Hematology, Erasmus MC, Rotterdam, The Netherlands
[2]Present address: department of Experimental Hematology, Hannover Medical School, Hannover, Germany

ABSTRACT

Gamma-retroviral vectors have been used successfully to deliver transgenes into hematopoietic stem cells in gene therapy trials for X-linked SCID and ADA-SCID, but efficacy in X-SCID also resulted in a relatively high frequency of insertional mutagenesis resulting in leukemia. The frequency of retroviral malignant transformation based on cell line studies has been estimated at approximately $2.2x10^{-7}$ per transduced cell. To examine the *in vivo* incidence, we transduced BALB/c donor bone marrow cells with either wtStat5-IRES-EGFP or IRES-EGFP amphotropic gamma-retroviral vectors with an MOI of 1 and monitored irradiated α-thalassemic BALB/c mice transplanted with the transduced cells for up to 342 days after which bone marrow was transplanted into secondary recipients. In 313 mice, leukemia developed with a median latency time of 383 days post primary transplantation. Immunophenotyping identified 8 myeloid, 10 B cell and 3 T cell leukemias, and 3 of a less well defined phenotype. There was no significant difference in leukemia incidence or phenotype between the recipients of wtStat5 IRES-EGFP and control IRES-EGFP transduced cells. Overall frequency after retransplantation was $1.5x10^{-6}$ per transduced cell, considerably higher than previously predicted, whereas primary transplantations yielded a frequency of $1.9x10^{-7}$. Integration analysis revealed no specific relation between insertions retrieved from mice with malignancies compared to the entire dataset of insertions.

INTRODUCTION

Gamma-retroviruses integrate their genome stably into host cells and gene therapy vectors based on these viruses can be used to integrate therapeutic genes resulting in persisting presence of a transgene. This is however not without risk, since replication competent gamma-retroviruses are known to cause leukemias in mice (1,2) and replicating viruses have been used to elucidate genes that might cause leukemia when hit by a virus insertion (3-6). The viral vectors derived from these viruses retain this property, as has been demonstrated by leukemia development in mice (7,8) and in macaques (9).

These retroviral vectors have been used in clinical trials to express adenosine deaminase (10,11) or IL2RG (12,13) in patients that carry dysfunctional copies of these genes and suffer from severe combined immunodeficiencies. In otherwise very successful clinical trials for X-linked SCID performed in Paris and London, 5 children transplanted with replication-defective retrovirus containing the *IL2RG* gene, developed leukemia (14,15).

Although insertional oncogenesis by retroviral integration is predicted to be rather rare, for gamma-retroviral insertions on a per cell basis $2x10^{-7}$ (16) to 10^{-9} for lentiviral vectors (17), in a clinical setting, 5-20 million (12) transduced hematopoietic progenitor cells are transplanted. Using the incidences and transplanted cell numbers above, the number of potentially leukemic clones can estimated to be as large as 4-5.

Detailed analyses of the effect of expressing IL2RG on the HSC that occurred in the wake of the leukemia that occurred in the clinical trials showed that the onset of leukemia could very well be initiated by the signaling molecule IL2RG itself (18,19) although this notion was debated by others (20-22).

To analyze the potential risk of leukemogenesis after retrovirus-mediated gene transfer in more detail, we (23) and others (8, 24-27) analyzed retrovirus integration patterns in several experimental settings. The study described here was initiated to investigate the frequency of malignant transformation in a low dose viral vector transduction setting in mice, mimicking the preferred clinical setting of having low numbers of integrated vector copies per cell and was designed to consist of a primary transplantation followed by a retransplantation to allow a long term follow up using wild-type BALB/c mice.

MATERIAL & METHODS

Experimental setting / background

The mice described in this study originate from 4 independent experiments in which the safety and efficiency of retroviral mediated gene transfer of a signaling protein. This study used wtStat5 as a transgene and EGFP as control. The retroviral vectors used were pLZRS-wtStat5-IRES-EGFP and pLZRS-IRES-EGFP (28). To analyze the clonal capacity

of the engrafted cells, cohorts of primary transplanted mice were killed and retransplanted into secondary recipients at several time intervals after transplantation to extend increase proliferative pressure on the transplanted cells and to extend the observation time of the experiments. Secondary recipients were transplanted at 117, 202, 206, 245 and 342 days after the primary transplantation and, cohort sizes were 37, 34,10,10 and 35 mice, respectively, Table 1). At the time the mice started to develop leukemia-like features (high white blood cell counts, enlarged spleen, low platelet counts), full analysis of the phenotype and virus integration pattern was performed in addition to a complete obduction and routine histological examination. Hematopoietic malignancies were found in 3 primary transplanted mice as well as in 21 mice carrying transgene and 9 mice that did not carry transgene after secondary transplantation. Bone marrow of these leukemic mice was retransplanted in sublethally irradiated recipients (6 Gy) to confirm the leukemias and to obtain purified clones of leukemic cells for further analysis.

Table 1. Overview of the four experiments comparing wtStat5-IRES-EGFP and IRES-EGFP expression in mice. In experiment 1, 2 and 3 escalating doses of transduced cells were transplanted. In experiment 4 recipient mice were irradiated with 1, 2 or 3 Gy before receiving the transplant. The total number of transplanted cell as well as the number of transgene positive cells, as measured by FACS on the day of transplantation is indicated for each experimental group.

			Mice transplanted per cell dose						
Group	Experiment	Total no. of animals transplanted	10^6	$3x10^5$	10^5	$3x10^4$	10^4	Transplanted cells	Transgene$^+$ cells in this group
EGFP	1	24	5	4	5	5	5	$6.9x10^6$	$1.65x10^6$
wtStat5	1	25	5	5	5	5	5	$7.2x10^6$	$1.46x10^6$
Control	1	39	0	0	13	13	13	$1.82x10^6$	0
EGFP	2	24	4	5	5	5	5	$6.2x10^6$	$1.77x10^6$
wtStat5	2	24	4	5	5	5	5	$6.2x10^6$	$1.51x10^6$
Control	2	29	0	0	9	10	10	$1.3x10^6$	0
EGFP	3	25	5	5	5	5	5	$2.2x10^6$	$5.86x10^5$
wtStat5	3	25	5	5	5	5	5	$7.2x10^6$	$2.04x10^6$

			Mice transplanted per radiation Dose (10^6 cells each, $5x10^5$ for controls)					
			0 Gy	1 Gy	2 Gy	3 Gy	Transplanted cells	Transgene$^+$ cells in this group
EGFP	4	20	5	5	5	5	$2x10^7$	$3.9x10^6$
wtStat5	4	20	5	5	5	5	$2x10^7$	$2.8x10^6$
Control	4	19	5	5	5	4	$9.5x10^6$	0

	Total number of mice	Total number of cells	Total number of transgene$^+$ cells
EGFP	93	$35.3x10^6$	$7.9x10^6$
wtStat5	94	$40.6x10^6$	$7.9x10^6$
Control	87	$12.6x10^6$	
Total	274	$88.5x10^6$	$15.7x10^6$

Animals and cells

Specific pathogen free (SPF), BALB/c mice, 8-12 weeks of age, were bred and housed under SPF conditions at the Experimental Animal Facility of Erasmus University (Rotterdam, The Netherlands) in conformity with legal regulations, which include approval by an ethical committee under Dutch law. BM of male mice was isolated by Percoll gradient centrifugation (Amersham Biosciences) and transduced using amphotropic (MLV) Phoenix pLZRS-IRES-EGFP or pLZRS-wtStat5-IRES-EGFP (MOI ~1) on retronectin coated dished as described before (29) with an average transduction efficiency of 23% EGFP+ cells. These cells were then transplanted in different doses (10^4, $3x10^4$, 10^5, $3x10^5$ and 10^6) into 274 (93 in the EGFP group, 94 in the wtStat5 group and 87 no-virus controls) female, α-thalassemic mice, that had been irradiated with 6 Gy total body irradiation (TBI) from a ^{137}Cs source (Gammacell, Canada) before. A total of $8x10^6$ transgene positive cells were transplanted in both groups. Monthly, the level of donor-chimerism was measured in the peripheral blood (PB) using the red blood cell size as a parameter (α-thalassemic mice have on average smaller erythrocytes compared to healthy BALB/c mice) (30). Furthermore, the level of EGFP in the erythrocytes, thrombocytes and leukocytes was determined by flow cytometry (FACSCalibur, BD Biosciences) and clonogenic culture assays (CFU-C and BFU-E). Blood was obtained using retro-orbital puncture in isoflurane anesthetized mice, collected in EDTA tubes for small volumes (BD Bioscience, San Jose, CA) and analyzed using a blood cell counter (ABC-VET, ABX Diagnostics, Montpellier, France).

Mice were considered leukemic when they were moribund, showed increased (>$25x10^6$ WBC/ml) white blood cell counts or thrombocytopenia or leukopenia and when retransplanted bone marrow or spleen cells gave rise to a leukemia in the secondary recipients.

Analysis of hematological malignancies

When mice were moribund, they were euthanized in conformity to the ethical guidelines of the Animal Experiment committee and Animal Welfare officer at ErasmusMC. At the day of sacrifice BM and spleen cells were analyzed for the presence of donor-derived EGFP-positive cells in erythroid, leukocytic and thrombocytic lineages by flow cytometry. For phenotyping the peripheral blood, bone marrow and spleen cell isolates we used Ter119, CD4, CD8, CD11b, Gr-1, B220 c-Kit and Sca-1 antibodies (All BD Biosciences) and clonogenic culture assays (CFU-C and BFU-E). When transgene expression could not be determined, the presence of the virus LTR was determined by PCR, using primers designed against the LTR of the virus backbone (LTR_LZRS_L,

5'-CCAAAGCGGATATCTGTGGT and LTR_LZRS_R: 5'-AAGGCACAGGGT-CATTTCAG-3') using an initial incubation at 95°C, followed by 35 cycles of 95°C for 1 min., 60°C for 45 sec and 72°C for 1 min, resulting in a 227 bp product when the virus LTR is present in the sample. In 3 cases, no expression of transgene was measured by flow cytometry, but the LTR sequences could be determined by PCR, possibly showing that a virus with a deleted genome was inserted or that the transgene was deleted after integration.

Hematological abnormalities found were confirmed by secondary or tertiary transplantations. To this end, 10^6 BM and/or spleen cells from the leukemic mice were transplanted into 5 6 Gy irradiated α-thalassemic recipients. The mice generally became moribund at 1 to 3 weeks after transplantation and were sacrificed. PB, spleen and BM cells were analyzed as described.

Integration analysis

We analyzed the genome – retrovirus boundaries as previously described (31) using LAM-PCR. In short, 500 ng DNA was amplified linearly for 100 cycles using a biotinylated primer (Eurogentec) designed against the virus LTR, after which the products were digested using Tsp509I or HpyCH4 IV (New England Biolabs). A complementary linker cassette was ligated (Fastlink Ligase, Epicentre technologies) and nested PCR with primers designed against the virus LTR and the linker cassette was performed.

A LAM-PCR based method, modified to be less sensitive so that predominant clones in the polyclonal sample could be detected more accurately, was also employed to analyze predominant clones in the leukemia samples. In this analysis 10 ng input DNA was used and an additional product clean-up using biotinylated primers and streptavidin coated magnetic beads (Kilobase binder kit, Dynal) was performed.

Integration annotation

Virus – genome boundaries were sequenced by ligation of the LAM-PCR amplicons into pCR4-TOPO plasmids (Invitrogen). One Shot® Competent E. coli (Invitrogen) were transformed with these plasmids and after overnight growth single colonies were picked and their plasmids sequenced. Each LAM-PCR product was oversampled at least 3 times, to make sure that the dominant bands were recovered. The bacterial clones were sequenced at GATC (Konstanz, Germany). The resulting sequences were trimmed for linker, plasmid and LTR sequences. Sequences with evidence for LTR and linker sequences were considered of good quality and were masked for rodent and simple repeat sequences (using Repeatmasker[19]) aligned to the mouse genome (Build 36) using tools available in TFTargetmapper (Department of Bioinformatics, ErasmusMC, Rotter-

19 http://repeatmasker.org

dam, The Netherlands[20]). Genes within 100 kbp upstream or downstream of the virus integration as well as the genes closest to the virus integration were identified.

RCR analysis

The initial virus preparations were tested for the absence of replication competent retrovirus (RCR) according to Markowitz32. To do so, NIH 3T3 cells transduced with the virus stocks used in the animal experiments were cultured in DMEM +10% FCS for 14 days, passaging cells every 3 days. Fourteen days after transduction supernatant of the transduced cells was collected, passed through a 0.45 mm filter and added to $5x10^5$ NIH 3T3 in the presence of 4 mg/ml polybrene for 24 hours, after which cells were cultured in DMEM + 10% FCS. When the cells reached confluency, they were trypsinized and analyzed for GFP expression by flow cytometry.

To show absence of RCR after in mice that presented with hematopoietic malignancies, spleen cells of three mice that were found with hematopoietic malignancies were 20 Gy irradiated and $3x10^6$ cells were transplanted into 4-5 recipients. No leukemias were found after transplantation and a one-year observation period in the transplanted mice, demonstrating absence of replication competent leukemia-inducing virus.

Reverse transcriptase activity was determined in protein isolates of spleen cells of mice presented with hematopoietic malignancy as described (33). Mouse spleen cells were precipitated by centrifugation (1600 rpm, 5 minutes) and resuspended in 200 ml protease inhibitor cocktail (Boehringer). The protein content was determined and samples were stored at −20°C. 10mg protein was incubated with 6 ng BMV template RNA (Promega), 10 nmol dNTP, 200 nmol $MgCl_2$, 1.25 U AmpliTaq Gold (Applied Biosystems), 4 U RNAseOUT recombinant ribonuclease inhibitor (Invitrogen) , 15 pmol of each primer and 5 pmol probe (Eurogentec) and 150 ng activated calf thymus DNA (Sigma).

The product of reverse transcription of BMV RNA was amplified in an ABI Prism 7900 Sequence Detection System (Applied Biosystems) using real-time PCR (forward primer (5')TCTTGAGTTAGACCACAACGTTCCT(3'), reverse primer (5')TGCGCTTGTCTCTGTGTGAGA(3') and a 5'FAM (6-carboxyfluorescein) and 3'TAMRA (6-carboxytetramethyl-rhodamine) labeled probe (5')TCTGCTCGAGGAGAGCCCTGTTCC(3'). PCR conditions were 30 minutes at 48°C followed by 40 cycles of 1 minute 94°C, 30 seconds 60°C and 30 seconds of 72°C and a final 10 minutes of 72°C.

Protein samples from AM12 SF91 EGFP retrovirus producer cell lines were used as a positive control. Superscript II reverse transcriptase was used to calibrate the reaction in a range of 10^{-1} to 10^{-8} units reverse transcriptase. All tested samples had less than $4x10^{-6}$ (range $6.8x10^{-9}$ to $3.2x10^{-6}$) units reverse transcriptase activity, compared to $5x10^{-5}$ units for the virus producer cell line.

20 http://tftargetmapper.erasmusmc.nl

RESULTS

Provirus associated hematological malignancies in long-term follow up of recipient mice.

Three hematological malignancies were observed in the primary transplanted animals, consistent with a minimum frequency of malignant transformation of $1.9x10^{-7}$ per transduced cell. (Table 2). To investigate the occurrence malignancies in more detail, BM of the primary recipients was transplanted into a total of 126 secondary recipients in total, equivalent to approximately 90% of the initial transgene positive cells transplanted into primary recipients. In the secondary recipients, we observed 21 mice with hematological malignancies (12 EGFP, 9 wtStat5) carrying the transgene and 9 with hematological malignancies in which the transgene was absent (Table 2). First, we set out to determine whether the expression of wtStat5 resulted in faster progression to malignancy or an increase in the number of cases. A survival analysis did not reveal significant differences in survival time between wtStat5 IRES-EGFP or IRES-EGFP (Figure 1A. log rank test, p=0.767). A similar result was obtained when analyzing the incidence of hematological malignancies in the wtStat5-IRES-EGFP or IRES-EGFP groups, where the number of cases was not significantly different between the groups (p=0.672 (Fisher Exact 2-sided), and neither was the phenotype distribution of the malignancies (Table 3, p=0.518 (Chi square 2-sided). Representative micrographs of PB, liver and kidney are shown in Figure 2. We therefore concluded that the additional expression of wtStat5 in these experiments did not result in differences in the occurrence of hematological malignancies and combined the data from both groups for further analysis. The median survival time of the mice that presented with hematopoietic malignancies was 378 days (range 105-556 days) after transduction.

In total we observed 33 malignancies. In 9 leukemias a provirus could not be detected, demonstrating that in these experiments a 2.7-fold excess leukemia incidence over background should be attributed to retroviral integration. In both wtStat5-IRES-EGFP group and the IRES-EGFP group, $8x10^{6}$ transduced cells were transplanted (Table 1), which leads to an estimated overall frequency of malignant transformation of $1.5x10^{-6}$ transduced cell (Table 2, wtStat5 $1.38x10^{-6}$, EGFP $1.63x10^{-6}$).

Three of 187 primary transplanted mice with transgene developed malignancies, whereas the majority of events was observed only after secondary transplantation (Figure 1B).

Integration site analysis shows mono- and oligoclonal expansion

Retrovirus integration sites were identified by amplification and sequencing of the virus-genome boundaries by LAM-PCR. The sequences were inspected for presence of both linker cassette and LTR primer presence after which they were aligned to the mouse

Table 2. Summary of the hematological parameters measured in the mice observed with hematopoietic malignancies. Flow cytometry phenotypes are presented as percentage of cell differentiation antigen positive cells in either GFP$^+$ or GFP$^-$ populations.

ID	Mouse	Group	Days post transduction	Provirus determined by	Phenotype	WBC (10^6/ml)	Immunophenotype in peripheral blood (%)							
							GFP+				GFP-			
							CD4+	CD8+	CD11b+	B220+	CD4+	CD8+	CD11b+	B220+
1	2ST106	wtStat5	105	FACS	M	18.5					NA			
2	2ST107	wtStat5	189	FACS	T	261.2					NA			
3	1EG211	EGFP	373	PCR	B	27.6-177	2.4	0.3	9.7	1.9	3.2	1.8	1.9	45.8
4	2EG209	EGFP	556	FACS	B	104.6	0.7	3.0	29.8	52.5	2.3	1.8	7.2	10.4
5	2EG211	EGFP	347	FACS	B	12.2	3.7	5.8	26.9	18.5	1.3	2.8	1.5	4.7
6	2EG212	EGFP	378	FACS	B/T	NA	6.6	25.6	7.4	48.9	0.6	1.2	0.2	4.2
7	2EG213	EGFP	347	FACS	M	50.6	1.6	1.4	79.6	10.9	3.6	1.4	1.5	8.3
8	2EG214	EGFP	388	FACS	M	348.8	2.6	20.2	17.7	86.0	0.3	0.6	0.3	2.7
9	2EG216	EGFP	301	FACS in RTX	T	173	0.2	0.0	0.0	0.4	93.0	87.6	0.0	0.1
10	2EG217	EGFP	378	FACS	B	NA	2.4	10.4	39.4	74.0	0.7	0.3	0.1	1.8
11	2EG218	EGFP	413	FACS	M	265	1.0	2.4	70.4	20.3	6.0	2.3	4.4	11.7
12	2EG237	EGFP	527	FACS	M	83	0.7	0.8	75.5	7.5	2.0	1.0	5.0	6.9
13	2EG238	EGFP	474	FACS	B	399	2.5	5.4	16.0	79.7	1.7	2.3	2.1	3.0
14	4EG204	EGFP	298	FACS	M	141.6	0.6	0.3	88.2	3.1	4.8	0.5	1.8	0.1
15	4EG205	EGFP	298	FACS	M	181.6	0.0	0.3	83.5	3.2	4.7	0.9	2.3	1.0
16	2ST204	wtStat5	378	FACS	B	NA	0.2	0.2	2.9	4.2	0.7	1.2	2.2	71.7
17	2ST205	wtStat5	449	FACS	M(B)	199	12.3	3.4	66.1	54.1	3.1	2.1	2.4	4.3
18	2ST221	wtStat5	449	FACS	M(B)	23.1	1.8	0.8	32.9	18.9	14.2	4.4	7.1	22.5
19	2ST222	wtStat5	356	FACS in RTX	T	13.2	1.3	0.8	6.6	0.2	3.7	15.5	5.6	22.2
20	2ST223	wtStat5	534	FACS	B	80.8	0.5	0.3	10.9	66.2	2.8	2.8	1.0	9.1
21	4ST201	wtStat5	410	FACS	B	150	0.7	1.8	65.1	39.9	0.6	0.5	2.6	2.9
22	4ST202	wtStat5	426	FACS	B	230	0.3	0.9	2.7	60.8	0.4	0.4	0.2	1.8
23	4ST203	wtStat5	371	FACS	B	68.7	1.0	1.4	10.0	64.2	4.6	2.1	8.9	8.2
24	4ST205	wtStat5	445	FACS	B/M	110	4.2	7.7	35.8	71.7	0.8	1.3	0.2	3.0

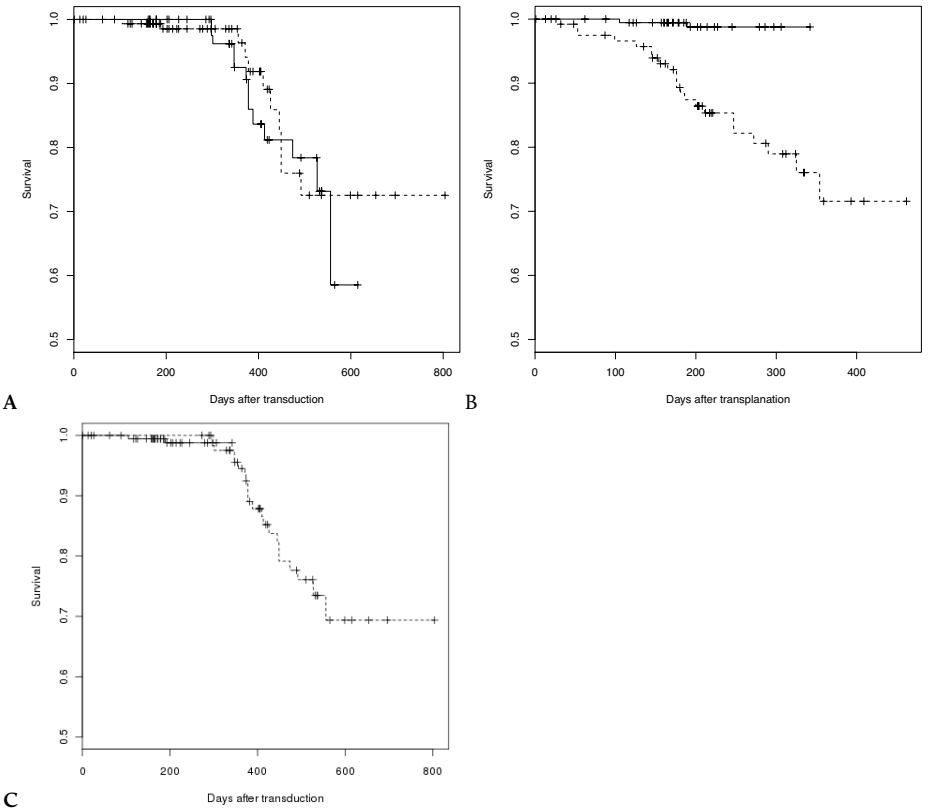


Figure 1. Survival analysis of mice transplanted with transduced cells with wtStat5-IRES-EGFP or IRES-EGFP. Plus signs indicate censored values without hematopoietic malignancy. (A) Survival after transduction, IRES EGFP shown as solid line, wtStat5-IRES-EGFP shown as a dashed line. (B) Survival after transplanation. The solid line indicates primary transplant recipients and the dashed line indicates the secondary transplant recipients. (C) Survival after transduction, compared between primary and secondary transplant recipients.

Table 3. Summary of the observed phenotypes in the mice with hematopoietic malignancies, indicating the numbers of mice in each group and the estimated incidence of leukemia. Incidence is reported as leukemias/total amount GFP positive cells.

	Group	Phenotype	Number of mice	Incidence
n=313	EGFP (n=162)	Myeloid	6	1.63×10^{-6}
		B-cell	5	
		T-cell	1	
		Mixed/Unclear	1	
	WtStat5 (n=151)	Myeloid	2	1.38×10^{-6}
		B-cell	5	
		T-cell	2	
		Mixed/Unclear	2	
		Total	24	1.50×10^{-6}
n=92	Untransduced	B-cell	1	
		Myeloid	1	
	Transduced	Provirus negative	9	

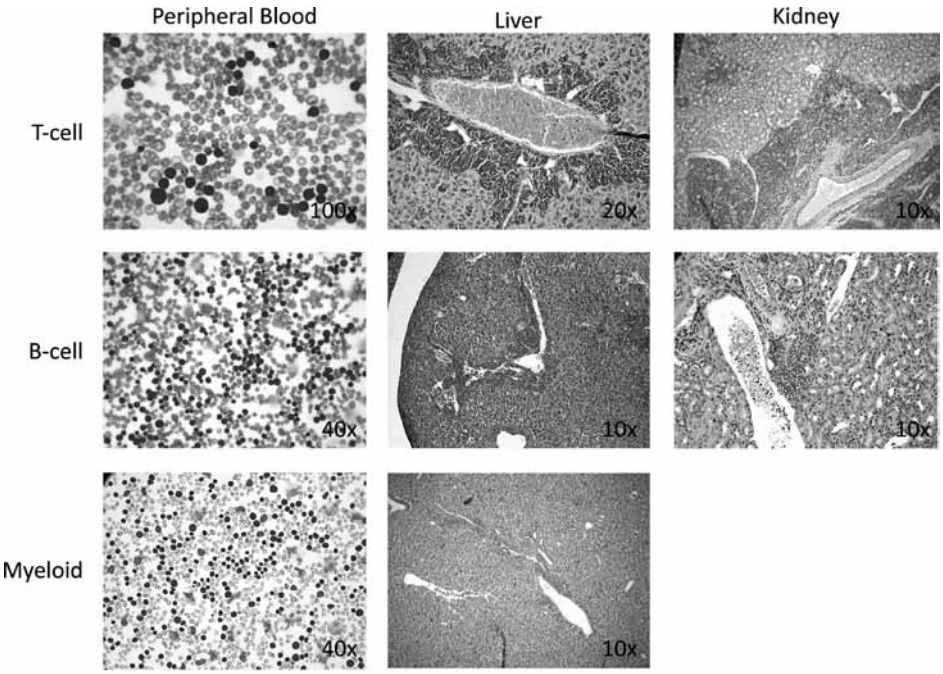


Figure 2. Histological sections of peripheral blood, liver and kidney, for T, B and myeloid malignancies. Liver and kidney infiltrates are clearly visible. Liver and kidney sections are stained with hematoxylin and eosin. Peripheral blood smears are stained with May Giemsa Grünwald.

genome using TFTargetmapper on database build m35 (April 2006). Only unambiguous BLAST alignments were considered virus insertion sites. Since the transductions were performed with MOI=1, we expected to obtain a low number of integrations per cell and predominantly clones carrying only one insertion. Since we could not exclude that the LAM-PCR procedure also amplified clones that did not contribute to the malignancy, we retransplanted the malignant cells in limited cell dose to dilute contaminating non-malignant cells and confirm the malignancy. In another approach, we reduced the amount input DNA for the LAM-PCR procedure to 10 ng to reduce the sensitivity of the method. The integration site analysis, together with the phenotype of the malignant clone in the peripheral blood/bone marrow, suggest that in most of the malignancies observed only monoclonal or oligoclonal integration was present.

Common insertions in *Evi1* and RTCGD genes

In addition to the survival and leukemia occurrence described in this paper, we analyzed how gamma-retroviral insertions are influenced by gene expression of the target locus and whether the functions of the genes in the vicinity of the virus played a role. Since we observed hematopoietic malignancies in this study, we analyzed whether the insertions retrieved from mice that developed malignancies occurred more frequently in RTCGD

genes, tumor suppressor genes or oncogenes. From the 135 insertion sites retrieved from all mice observed in these experiments, 88 insertion sites were also found in the diseased mice. Oncogenes and tumor suppressor genes were retrieved from Entrez Gene, and the comparison showed that in both datasets, *Rab5c* was the only annotated oncogene, while *Zfpn1a1* was the only annotated tumor suppressor gene. When analyzing the RTCGD, we observed 34 of the 88 insertion sites from diseased mice in the RTCGD, which is not significantly different from the 46 insertions out of the 135 sites in the entire dataset. We did observe insertions in *Evi1* in several different mice (Table 4).

RCTGD genes or dominant clones are not causing faster progression

The over-representation of RTCGD genes in this dataset raises the question of the involvement of these genes in leukemia progression. We therefore analyzed the time at which the malignancies occurred and compared those with the RTCGD genes that were hit. (Figure 3a), in addition, we compared time and insertions that we identified using LAM-PCR with limited sensitivity. (Figure 3b). In the figure 3a, it is obvious that *Evi1* was observed most frequently, but except for *Rab5c*, none of the RTCGD genes are uniquely associated with malignancies that occur earlier that the median survival (Figure 1C). The same holds for the insertions that were retrieved as dominant insertions using LAM-PCR with reduced sensitivity. (Figure 3b), although there was a significant higher amount of RTCGD genes found in the dominant clones than was observed in the entire dataset (Fisher exact test p=0.01)

DISCUSSION

In the present study, we set out to analyze the effect of gamma-retroviral insertions in mouse bone marrow cells after transplantation. We compared vectors expressing wtStat5-IRES-GFP and IRES-GFP control vectors and did not observe any difference in the frequency of malignancies and phenotype distribution between the groups, showing that over expression of wtStat5 does not lead to an increase in malignancies. The frequency of malignancies with 3 of 187 mice after primary transplantation and an additional 21 of 126 mice after secondary transplantation is considerably higher than 11 out of 92 mice in mice that were untreated and served as a background for spontaneous leukemia occurrence in the BALB/c background. The observed median latency 378 days after transduction, which points to the necessity of long term observation in studies that employ mice with a wild-type background. In studies in the *Cdnk2a*$^{-/-}$ mouse model, the latency of ~225 days is obviously shorter, but due to the high background of malignancies the differences between vectors, especially those with reduced oncogenic potential, might be difficult to analyze. The low background of malignancies in the

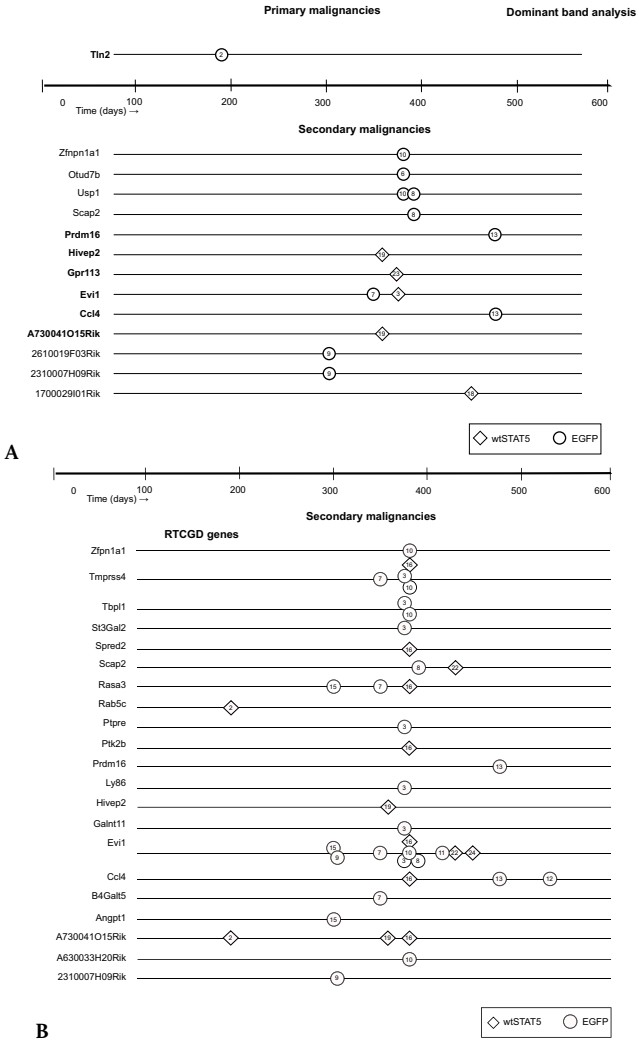


Figure 3. Insertion site occurrence over time. Insertion sites for mice observed with malignancies in wtStat5-IRES-EGFP (diamonds) or IRES-EGFP group (circles). Mouse numbers (Table 2) are indicated in the symbols. (A) RTCGD genes observed in the mice with malignancies. (B) Prominent bands sequenced using LAM-PCR with reduced sensitivity. RTCGD genes are indicated in bold.

current study and the large number of transduced cells and mice involved allow us for the first time to reliably estimate frequency and latency of malignancies induced by replication defective gamma-retroviral insertion.

Looking closer into the insertion sites retrieved from the mice that developed leukemias shows that a direct causal relation between insertions and leukemogenesis is very difficult to establish beyond doubt, even though we aimed to integrate only limited numbers of virus copies per cells. Analyzing dominant amplicons showed that the

Table 4. Common insertions observed in the mice with hematopoietic malignancies. Unique insertions occurring near the same gene are indicated with their insertion site ID (ISID).

Integration sites	No. of integrations observed
ISID002 (no associated gene)	12
ISID119 (no associated gene)	9
Asprv1	7
Tmprss4, Rasa3, ISID017, ISID060	5
Evi1 (ISID013), Ccl4, ISID065, Elovl7	4
Evi1 (ISID001), NP_032627.1, Q3TBQ1, Scap2, Akap13, Cct5, Otud7b, Evi1(ISID089), Ttl	3
Evi1(ISID022), Otud7b(ISID047), Evi1(ISID061), Evi1(ISID069), Tmprss4 (ISID073), ENSMUST86520, Evi1(ISID090), ISID091, Klhl1, Ccl4, ISID115, Ypel2, Fcrlm1	2
286 others	1

clones carrying these insertions are not prone to faster progression to leukemia, similar to insertions in RTCGD genes.

In the current study, we did not find an increase or difference in lineage distribution between the wtStat5b transduced mice and the GFP control group. This is in contrast to studies in mice in which wtStat5 was over-expressed in the lymphoid compartment, which readily developed lymphomas (34). The expression and activation of Stat5 has been widely studied, since its constitutive activation in AML is caused by a FLT-ITD mutation (35,36). Furthermore, the absence of Stat5 and Stat1 reduces leukemia growth in cells over-expressing both MN1 and HOXA9 (37). The cellular effects of constitutive active Stat5 have been shown not only to occur in the nucleus, but also in the cytoplasm, where the activated Stat5 can interfere PI3K signaling (38). Earlier studies were performed where constitutive active Stat5 was expressed in $CD34^+$ LSK or $CD34^-$ LSK cells, which showed that caStat5, when transduced into $CD34^-$LSK, the more immature stem cell population of the mouse, resulted in a myeloproliferative disease, whereas this was not observed with $CD34^+$LSK host cells (39). Taken together, Stat5 seems to act as a facilitator for leukemia development. The lack of effect in our study is interesting, since it shows that either the occurrence of a Stat5 specific collaborative event is not very likely or that the effect of gamma-retroviral insertion is more likely to result in leukemia.

We investigated the leukemogenic effect of a pLZRS based gamma-retroviral vector that transferred wtStat5 IRES EGFP or IRES EGFP into mouse BM cells. Since the expression of wtStat5 did not lead to an increase in the number of hematopoietic malignancies or the distribution of the malignancies observed, all data were pooled for further analysis. In our experiments analyzing 274 mice in the primary transplant cohort and 126 mice in the secondary cohort, we observed a 2.7 fold increase in the number of hematopoietic malignancies compared to the control group. We calculated that in the current study 1 in 700,000 transduced cells finally lead to a hematopoietic malignancy. The mean survival time of the mice carrying these malignancies was 383

days after transduction. This is expectedly longer than observed in retroviral leukemogenesis studies in *Cdnk2a*-/- mice (40) where a mean survival of ~225 days after transduction was observed for the gamma-retroviral vector tested and much slower than leukemogenesis induced by replication competent gamma-retroviruses. It is also longer than observed before in C57Bl6 mice (194 days, 8 cases) (8). The shorter latency in the C57Bl6 mice could be attributed to the relatively high (>5) copy number used, contrasting our present study. The low MOI used in our study resulted in a low number of integrations per cell, which probably requires secondary events to create a leukemic phenotype. The existence of such secondary events has been shown in cells of a XSCID patient that received gene therapy (15). In the current model the secondary transplantation might provide additional proliferative pressure (41), which would explain the higher frequency observed. The frequency of the observed malignancies ($1.5x10^{-6}$ malignancies/transplanted transgene positive cell) is about ten times higher than the mutation frequency of gamma-retrovirus in TF-1 cells ($2.2x10^{-7}$) (16). It seems that *in vitro* studies of mutation frequency can be compared within the model, but do not necessarily predict the actual mutation frequency in mouse transplantation models.

High numbers of integrated vectors might lead to collaborative events that eventually might bring about malignancies (42). To adequately assess how often a single insertion would lead to a malignancy, it is important to study vector insertions in a low copy number setting. Our study therefore used MOI 1 to limit the number of insertions per cell. In the *Cdnk2a*$^{-/-}$ mouse model, the vector copy number was between 2 and 4, but the shorter latency can readily explained by the effect of the knock-out phenotype, since the mock transduced cells only survive 25 days longer on average than the mice with malignancies brought about by the viral vector insertions. In the study presented here, the longer observation time might be helpful when analyzing optimized vector backbones such as self-inactivating gamma-retroviruses or lentiviruses. The accelerated onset of malignancies in the *Cdnk2a*$^{-/-}$ model only allows analysis of very strong effects and milder effects of the integration of optimized backbones or lentiviruses will need even higher copy numbers (43). In clinical gene therapy protocols, such high vector copy numbers are unlikely.

Insertion sites are often related to genes involved in human cancer or previously retrieved gamma-retroviral insertions (44). The existence of single dominant clones in a sample is also considered a potential risk. We investigated whether mice carrying insertion in RTCGD genes showed shorter survival than the average. Except for the insertion in *Rab5c*, which was found in after 189 days (mouse 2), the insertions in RTCGD genes did not seem to lead to faster progression to leukemia. By limiting the amount of DNA analyzed, we reduced the sensitivity of LAM-PCR to be able to detect dominant clones. The insertions detected using this approach, did also not show faster progression to leukemia. While the appearance of dominant clones over time might

leads to unfavorable results (45), it can also be a benign manifestation of clones with increased hematological fitness (23).

The *in vitro* studies investigating the leukemogenic effect of gamma-retroviral vector designs and the effect of internal promoters in the backbones are very effective and time efficient (17, 46-48), but are primarily dependent on the deregulation of a specific site such as *Evi1* or *Prdm16* (48) or *Ghr* (17), whereas a mouse transplantation study with long term observation allows the opportunity to analyze a much wider range of insertions, as demonstrated by this study. Mouse transplantation experiments do also allow the analysis of different target cell types, such as more primitive hematopoietic stem cell fractions (23) or lymphoid cells (49). Mouse transplantation studies are therefore a meaningful addition to safety testing after initial assessment of the deregulatory effect of the virus in an *in vitro* setting.

REFERENCES

1. Friend C. Transplantation immunity and the suppression of spleen colony formation by immunization with murine leukemia virus preparation (friend). Int. J. Cancer. 1968;3:523-529.
2. Moloney JB. Biological studies on a lymphoid-leukemia virus extracted from sarcoma 37. i. origin and introductory investigations. J Natl Cancer Inst. 1960;24:933-51.
3. Justice MJ, Morse HC3, Jenkins NA, Copeland NG. Identification of evi-3, a novel common site of retroviral integration in mouse akxd b-cell lymphomas. J Virol. 1994;68:1293-300.
4. Suzuki T, Shen H, Akagi K, et al. New genes involved in cancer identified by retroviral tagging. Nat Genet. 2002;32:166-74.
5. Joosten M, Vankan-Berkhoudt Y, Tas M, et al. Large-scale identification of novel potential disease loci in mouse leukemia applying an improved strategy for cloning common virus integration sites. Oncogene. 2002;21:7247-55.
6. Erkeland SJ, Verhaak RG, Valk PJ, et al. Significance of murine retroviral mutagenesis for identification of disease genes in human acute myeloid leukemia. Cancer Res. 2006;66:622-6.
7. Li Z, Dullmann J, Schiedlmeier B, et al. Murine leukemia induced by retroviral gene marking. Science. 2002;296:497.
8. Modlich U, Kustikova OS, Schmidt M, et al. Leukemias following retroviral transfer of multidrug resistance 1 (mdr1) are driven by combinatorial insertional mutagenesis. Blood. 2005;105:4235-46.
9. Seggewiss R, Pittaluga S, Adler RL, et al. Acute myeloid leukemia is associated with retroviral gene transfer to hematopoietic progenitor cells in a rhesus macaque. Blood. 2006;107:3865-7.
10. Aiuti A, Slavin S, Aker M, et al. Correction of ada-scid by stem cell gene therapy combined with nonmyeloablative conditioning. Science. 2002;296:2410-3.
11. Kohn DB, Hershfield MS, Carbonaro D, et al. T lymphocytes with a normal ada gene accumulate after transplantation of transduced autologous umbilical cord blood cd34+ cells in ada-deficient scid neonates. Nat. Med. 1998;4:775-780.
12. Hacein-Bey-Abina S, Le Deist F, Carlier F, et al. Sustained correction of x-linked severe combined immunodeficiency by *ex vivo* gene therapy. N Engl J Med. 2002;346:1185-93.
13. Gaspar HB, Parsley KL, Howe S, et al. Gene therapy of x-linked severe combined immunodeficiency by use of a pseudotyped gammaretroviral vector. Lancet. 2004;364:2181-7.
14. Hacein-Bey-Abina S, Garrigue A, Wang GP, et al. Insertional oncogenesis in 4 patients after retrovirus-mediated gene therapy of scid-x1. J. Clin. Invest. 2008;118:3132-3142.

15. Howe SJ, Mansour MR, Schwarzwaelder K, et al. Insertional mutagenesis combined with acquired somatic mutations causes leukemogenesis following gene therapy of scid-x1 patients. J Clin Invest. 2008;118:3143–3150.
16. Stocking C, Bergholz U, Friel J, et al. Distinct classes of factor-independent mutants can be isolated after retroviral mutagenesis of a human myeloid stem cell line. Growth Factors. 1993;8:197-209.
17. Bokhoven M, Stephen SL, Knight S, et al. Insertional gene activation by lentiviral and gammaretroviral vectors. J. Virol. 2009;83:283-294.
18. Woods NB, Bottero V, Schmidt M, von Kalle C, Verma IM. Gene therapy: therapeutic gene causing lymphoma. Nature. 2006;440:1123.
19. Shou Y, Ma Z, Lu T, Sorrentino BP. Unique risk factors for insertional mutagenesis in a mouse model of xscid gene therapy. Proc Natl Acad Sci U S A. 2006;103:11730-5.
20. Pike-Overzet K, de Ridder D, Weerkamp F, et al. Gene therapy: is il2rg oncogenic in t-cell development? Nature. 2006;443:E5.
21. Thrasher AJ, Gaspar HB, Baum C, et al. Gene therapy: x-scid transgene leukaemogenicity. Nature. 2006;443:E5-6; discussion E6-7.
22. Pike-Overzet K, de Ridder D, Weerkamp F, et al. Ectopic retroviral expression of lmo2, but not il2r-gamma, blocks human t-cell development from cd34+ cells: implications for leukemogenesis in gene therapy. Leukemia. 2007;21:754-63.
23. Kustikova OS, Geiger H, Li Z, et al. Retroviral vector insertion sites associated with dominant hematopoietic clones mark "stemness" pathways. Blood. 2007;109:1897-907.
24. Recchia A, Bonini C, Magnani Z, 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 U S A. 2006;103:1457-62.
25. Hematti P, Hong BK, Ferguson C, et al. Distinct genomic integration of mlv and siv vectors in primate hematopoietic stem and progenitor cells. PLoS Biol. 2004;2:e423.
26. Mitchell RS, Beitzel BF, Schroder AR, et al. Retroviral dna integration: aslv, hiv, and mlv show distinct target site preferences. PLoS Biol. 2004;2:E234.
27. Laufs S, Nagy KZ, Giordano FA, et al. 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;10:874-81.
28. Brugman MH, Pike-Overzet K, Schmidt M, et al. Retroviral vector integrations relate to hematopoietic stem cell gene expression patterns. Submitted for publication..
29. van Hennik PB, Verstegen MM, Bierhuizen MF, et al. Highly efficient transduction of the green fluorescent protein gene in human umbilical cord blood stem cells capable of cobblestone formation in long-term cultures and multilineage engraftment of immunodeficient mice. Blood. 1998;92:4013-22.
30. van den Bos C, Kieboom D, van der Sluijs JP, et al. Selective advantage of normal erythrocyte production after bone marrow transplantation of alpha-thalassemic mice. Exp Hematol. 1994;22:441-6.
31. Schmidt M, Hoffmann G, Wissler M, et al. Detection and direct genomic sequencing of multiple rare unknown flanking dna in highly complex samples. Hum Gene Ther. 2001;12:743-9.
32. Markowitz D, Goff S, Bank A. Construction and use of a safe and efficient amphotropic packaging cell line. Virology. 1988;167:400-406.
33. Lovatt A, Black J, Galbraith D, et al. High throughput detection of retrovirus-associated reverse transcriptase using an improved fluorescent product enhanced reverse transcriptase assay and its comparison to conventional detection methods. J. Virol. Methods. 1999;82:185-200.
34. Kelly JA, Spolski R, Kovanen PE, et al. Stat5 synergizes with t cell receptor/antigen stimulation in the development of lymphoblastic lymphoma. J. Exp. Med. 2003;198:79-89.
35. Hayakawa F, Towatari M, Kiyoi H, et al. Tandem-duplicated flt3 constitutively activates stat5 and map kinase and introduces autonomous cell growth in il-3-dependent cell lines. Oncogene. 2000;19:624-631.
36. Mizuki M, Fenski R, Halfter H, et al. Flt3 mutations from patients with acute myeloid leukemia induce transformation of 32d cells mediated by the ras and stat5 pathways. Blood. 2000;96:3907-3914.

37. Heuser M, Sly LM, Argiropoulos B, et al. Modelling the functional heterogeneity of leukemia stem cells: role of stat5 in leukemia stem cell self-renewal. Blood. 2009;.
38. Harir N, Pecquet C, Kerenyi M, et al. Constitutive activation of stat5 promotes its cytoplasmic localization and association with pi3-kinase in myeloid leukemias. Blood. 2007;109:1678-1686.
39. Kato Y, Iwama A, Tadokoro Y, et al. Selective activation of stat5 unveils its role in stem cell self-renewal in normal and leukemic hematopoiesis. J. Exp. Med. 2005;202:169-179.
40. Montini E, Cesana D, Schmidt M, et al. Hematopoietic stem cell gene transfer in a tumor-prone mouse model uncovers low genotoxicity of lentiviral vector integration. Nat Biotechnol. 2006;24:687-96.
41. Baum C, Dullmann J, Li Z, et al. Side effects of retroviral gene transfer into hematopoietic stem cells. Blood. 2003;101:2099-114.
42. Hanahan D, Weinberg RA. The hallmarks of cancer. Cell. 2000;100:57-70.
43. Montini E, Cesana D, Schmidt M, et al. The genotoxic potential of retroviral vectors is strongly modulated by vector design and integration site selection in a mouse model of hsc gene therapy. J. Clin. Invest. 2009;119:964-975.
44. Akagi K, Suzuki T, Stephens RM, Jenkins NA, Copeland NG. Rtcgd: retroviral tagged cancer gene database. Nucleic Acids Res. 2004;32:D523-7.
45. Ott MG, Schmidt M, Schwarzwaelder K, et al. Correction of x-linked chronic granulomatous disease by gene therapy, augmented by insertional activation of mds1-evi1, prdm16 or setbp1. Nat Med. 2006;12:401-9.
46. Modlich U, Bohne J, Schmidt M, et al. Cell-culture assays reveal the importance of retroviral vector design for insertional genotoxicity. Blood. 2006;108:2545-53.
47. Zychlinski D, Schambach A, Modlich U, et al. Physiological promoters reduce the genotoxic risk of integrating gene vectors. Mol Ther. 2008;16:718-25.
48. Modlich U, Navarro S, Zychlinski D, et al. Insertional transformation of hematopoietic cells by self-inactivating lentiviral and gammaretroviral vectors. Mol. Ther. 2009;.
49. Newrzela S, Cornils K, Li Z, et al. Resistance of mature t cells to oncogene transformation. Blood. 2008;112:2278-2286.

CHAPTER

6

An automated online tool for virus integration site annotation

Adapted from: Martijn Brugman[1,3*], Marshall Huston[1*], Sebastiaan Horsman[2], Andrew Stubbs[2], Peter van der Spek[2] and Gerard Wagemaker[1**], An automated online tool for virus integration site prediction and annotation. *Submitted for publication.*

[1] Erasmus University Medical Center, Department of Hematology, Rotterdam, the Netherlands.
[2] Erasmus University Medical Center, Department of Bioinformatics, Rotterdam, the Netherlands
[3] Present address: Hannover Medical School, Department of Experimental Hematology, Hannover, Germany
[*]These authors contributed equally to this study.

ABSTRACT

Retroviral vector mediated gene transfer allows for therapy of monogenic diseases by introducing therapeutic genes into hematopoietic stem cells (HSC).This method of treatment was demonstrated in animal models for a variety of inherited diseases and in trials for human severe combined immune deficiency (SCID). The risks of therapeutic gene integration include aberrant expression of a neighboring gene, resulting at low frequencies (10^{-7}-10^{-6}/transduced cell) in oncogenesis. Mechanisms governing insertional mutagenesis are subject of intensive ongoing studies, integration analyses with rapidly increasing data requiring automated bioinformatics. The presented web tool, Methods for Analyzing ViRal Integration

Collections (MAVRIC), takes a set of vector integration sequences as input and utilizes BLAST to align each virus-genome boundary to a location in the genome and is available at http://mavric.erasmusmc.nl/. MAVRIC returns the nearest gene, distance to transcription start site, common integration sites and patterns of nearby gene expression based on our data for human $CD34^+$ HSC analyzed on an Affymetrix U133 gene array. We used the integration data of the SCID gene therapy trials for evaluation of MAVRIC. The output demonstrates a specific gamma-retroviral integration pattern and illustrates how MAVRIC allows for direct multi-parameter comparison of integration patterns between different species, disease entities and viral vectors.

INTRODUCTION

The availability of PCR based techniques (inverse PCR, LAM-PCR, LM-PCR, splinkerette PCR) to identify viral integration sites has lead to efforts to classify the relationship between virus integration and leukemia when using replicating retroviruses (1-3) as well as in gene therapy to identify clonal dominance and insertional mutagenesis leading to oncogenic events (4,5). Further investigation into the integration profile of retroviruses and lentiviruses (6-8) has shown a relationship between the occurrence of insertional oncogenesis and the integration profile (8,9). Accurate analysis of the amplicon sequences retrieved after sequencing of the viral genome boundaries (10) is time consuming and sensitive to data processing errors. Furthermore, high throughput sequencing methodology quickly produces more sequences than can be reasonably analyzed using the available online tools. We therefore developed a bioinformatics pipeline aimed at high throughput data analysis for virus genome boundaries. Similar tools have been described (11,12) for alignment of genome virus boundary sequences, some providing exhaustive analysis of the genomic features surrounding the integrations (12). MAVRIC (Methods for Analyzing ViRal Integration Collections), automatically aligns input sequences to the human or murine genome using BLAST according to user-specified parameters and returns information on the genes surrounding the integration site. This paper outlines the design choices made in the process of developing the tools for virus integration analysis and investigates the sources of differences between analyses performed on the same datasets using different databases and analysis methods.

METHODS

Integration analysis

The annotation workflow can be summarized as series of steps which consist of data clean up, integration site detection, and identification of the nearest gene(s) to the viral integration (Figure 1). The virus genome boundary sequences, without vector, plasmid or LTR sequences, are loaded as a FastA formatted file. Subsequently these genome boundary sequences are pre-screened using a locally running version of RepeatMasker[21]. The Repeatmasker algorithm screens sequences for repeats using cross-match, an implementation of the Smith-Waterman-Gotoh algorithm developed by Phil Green[22]. Simple repeats, tandem repeats, segmental duplications and interspersed repeats such as pseudogenes, retrotranscripts, SINEs, DNA transposons, retrovirus retrotransposons

21 www.repeatmasker.org, Smit, AFA, Hubley, R & Green, P. RepeatMasker Open-3.0 1996-2004

22 http://www.phrap.org/, University of Washington

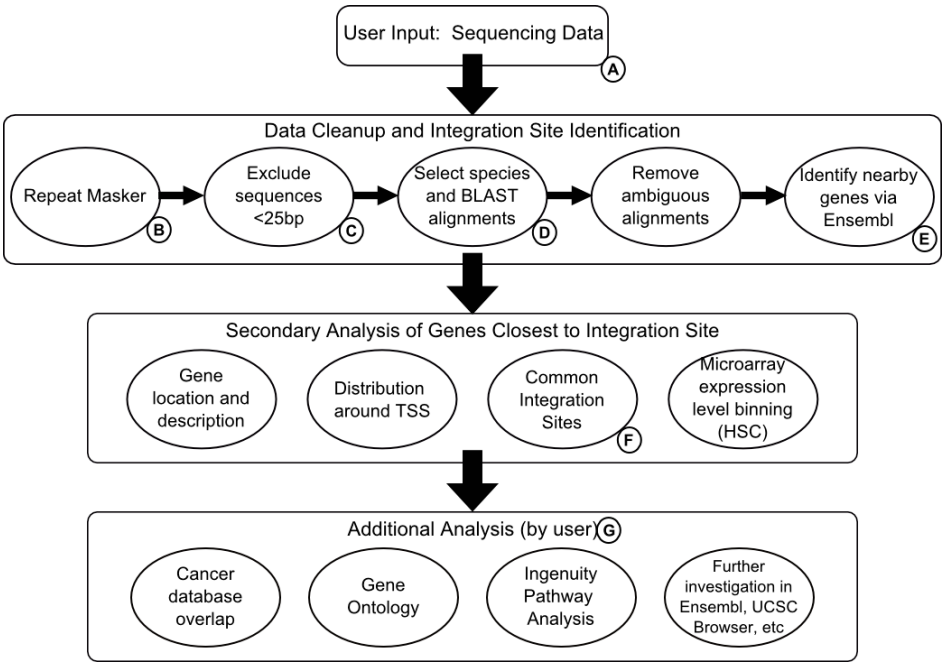


Figure 1. MAVRIC workflow.

and non-retroviral transposons (LINEs) are masked by this algorithm. The user inputs the RepeatMasker criteria (species) and the minimum acceptable sequence length. Short sequences may return a large number of redundant, uninformative hits in the genome. To speed up the annotation process for large runs, we recommend using a minimum sequence length of 25bp. Next, the user selects the BLAST species and version to be used and the maximum acceptable e-value. The e-value threshold is used to speed up the annotation process by minimizing the number of uninformative alignments returned by BLAST. Most sequences, in addition to the lowest e-value hit, return a large number of hits with higher e-values (partial alignments, etc), which can slow down the annotation. The maximum E-value threshold can be set to 10 if the user wishes to view all the BLAST hits. Finally the user chooses which Ensembl version, sets the size of the window flanking the integration site to search for genes, and chooses whether or not to return only RefSeq genes (13). Ensembl is used as the source for annotation data, since it provides a Perl interface which eases collection of data after the alignments have been made. MAVRIC aligns the screened sequences in the genome via BLAST (Basic Local Alignment Search Tool (14), a fast algorithm that allows alignment of query sequences (DNA or protein) to a predefined database. It finds the largest subsection of the query and tries to expand it in both directions, assigning penalties for the creation of gaps or mismatches. Given a query sequence, BLAST identifies the best fitting alignment in addition to similar alignments with higher penalty scores. In MAVRIC, the user can select the maximum e-value to be

returned, with the default maximum e-value set to 10^{-5}. The alignments of integration sites identified by BLAST are subsequently annotated using data available from Ensembl.

Annotation of retrieved integration locations

Our tool generates two outputs for Ensembl hits: one file that contains every gene located within a user-defined distance of the integration site, and another file which only contains the gene whose 5'end is closest to the integration site. The surrounding genes are taken into account because previous studies on deregulation caused by viral promoters showed that the genes surrounding virus insertions can be deregulated as well as the closest genes. This effect was shown to occur as far as 500 kbp from a virus insertion (4,15). MAVRIC generates a histogram showing the distribution of distances from the integration site to the nearest gene transcription start site. It also lists common integration sites (defined as 3 or more hits for a given gene in the dataset) and a summary of the analysis parameters and excluded sequences.

Nearest gene expression level binning

To determine the relationship between viral integration and the expression of nearby genes in hematopoietic stem cells, the target cells for gene therapy trials in inherited disorders of the blood cell / immune systems, we annotated the genes closest to each viral integration site (VIS) with their expression in hematopoietic stem cells. For human cells, we used gene expression data obtained from $CD34^+$ umbilical cord blood cells and for mouse cells we used Lin^- $Sca\text{-}1^+$ $c\text{-}kit^+$ cells. Genes from the file containing the unique closest gene to each VIS are sorted according to expression level in $CD34^+$ cells into 10 bins. MAVRIC creates a histogram displaying the number of VIS genes allotted to each bin, giving an overview of the relation between gene expression level and integration frequency. Expression levels in human $CD34^+$ cells and similar immature cells of mice were measured on an U133 plus 2.0 or Affymetrix Mouse 430 2.0 or gene arrays and compared to a random human integration data set generated by a perl script which chose a chromosome and bp integration site at random and located the nearest gene. The random dataset contained 1000 'integration sites', with 864 of them within 500 kbp of a human gene (Figure 4).

RESULTS

Effect of RepeatMasker

Genomes contain sequences of interspersed repeats and regions of low complexity: repetitive sequences make up almost 50% of the human genome (16). These repeat regions cause difficulties when intersecting with a virus genome boundary sequence.

When multiple alignments retrieved by BLAST share the lowest penalty score, the sequence will not point to a unique place in the genome. Repeat sequences also cause the throughput of the BLAST alignment to decrease. We approached this problem by pre-screening the virus genome boundary sequences using the online tool RepeatMasker. RepeatMasker allows batch uploads of sequences and returns sequences with the desired repeats masked. This data can then be aligned using BLAST without the throughput penalties that repeat sequences normally introduce. The numbers of different repeat consensus sequences varies between genomes and the user should indicate from which species their sequences are derived when selecting RepeatMasker. Figure 2 shows an example of the impact of repeat masking on sequence lengths. Repeatmasking clearly shortens the sequences by removal of the repeated regions.

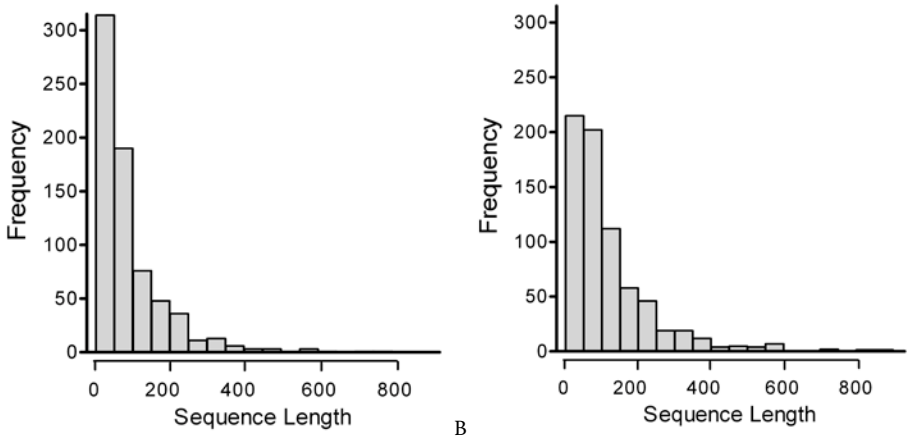


Figure 2. Changes in sequence length distribution caused by RepeatMasker. (A) The distribution of sequence lengths in the original data. (B) The distribution of sequence lengths after repeat masking.

Differences in annotation between genome databases

The annotation of virus integration sites is dependent on the analysis parameters provided. Although the different databases (UCSC, Ensembl, NCBI) use the same underlying genomic information[23], the annotations are slightly different. For instance, Ensembl[24] utilizes an automated annotation pipeline, using mRNA and protein data, combined with manually reviewed curated data from the Vega project[25] and reviewed protein coding transcripts from CCDS[26] (17). Any integration dataset therefore depends on the annotation database used. Although the majority of the genes identified will

23 http://genome.ucsc.edu/FAQ/FAQreleases, although UCSC only uses the C57Bl/6J assembly

24 http://www.ensembl.org/info/docs/genebuild/genome_annotation.html

25 http://vega.sanger.ac.uk/index.html

26 http://www.ensembl.org/info/docs/genebuild/ccds.html

be identical regardless of the database, there might be some differences between the annotations possibly leading to different genes being assigned as the closest gene to an integration. For example, the Mecom locus, which has several splice variants for Evi1 and Mds1, shows differences in annotation between NCBI, the UCSC genome browser and Ensembl (Supplementary figure 1, and Supplementary table 1).

Table 1 a Number of integrations mapped and genes identified by NCBI build, using published ADA-SCID as input. The consensus sequence is updated often, which can cause integration sites to be mapped to different places or to change the nearest gene. (b) Comparison of the genes linked to each viral integration site (VIS).The number of sequences with integrations is nearly identical, but there are some discrepancies in the Ensembl IDs of the genes linked to sequences. This means that the VIS occupies a slightly different place in the genome between the various builds. There are even greater discrepancies between gene names, indicating that the commonly used names for genes are subject to a great deal of turnover. Genome information and build dates were retrieved from ftp://ftp.ncbi.nih.gov/genomes/H_sapiens/.

	Human NCBI 34 (October 2003)	Human NCBI 35 (October 2004)	Human NCBI 36 (March 2008)
Number of integrations mapped	485	509	514
Number of genes annotated	442	465	470

Table 1b

	34 vs. 35	35 vs. 36	34 vs 36
Number of matching sequences with VIS	484	504	481
Number of matching EnsemblIDs	381	415	360
Number of matching gene names	268	360	265

The use of NCBI Reference sequences or broader annotation

The option to annotate only RefSeq genes has the advantage of returning only well described genes, but has the disadvantage of filtering out potentially interesting genomic features such as miRNAs, snRNAs and pseudogenes. In an analysis of a sample dataset (published ADA-SCID data described below), filtering for RefSeq genes resulted in only a 3.5% reduction in the number of sequences annotated but a 19.5% mismatch in the genomic features identified as nearest to the VIS (data not shown).

Performance Testing

The current application aligns 250 integrations per hour. This is suitable for moderately sized datasets with up to 500 sequences. The BLAST program can easily scale to be run in parallel (e.g. mpiBLAST), thus an increase in the number of processors (threads) available will directly result in a shorter analysis time. Therefore, we expect it will enhance the

performance of the application to increase the throughput when high throughput/deep sequencing results need to be analyzed. Alternatives to BLAST have been developed, such as SSAHA and BLAT. BLAT uses a client-server design together with a modified alignment algorithm, allowing efficient alignments of smaller sequences. It is optimized to recognize separated islands of perfect matches, which would be generated by the alignment of cDNA. SSAHA is used for fast searches in genome-size databases using a hashing approach, which enables it to perform sequence alignments 3 to 4 orders of magnitude faster than BLAST or FASTA and requires less memory (18).

Annotation of retrieved integration locations

For each integration site, there are surrounding genes upstream and downstream on both strands. In addition to the closest gene, an integration may have three or more neighboring genes that could be influenced by elements of the inserted provirus. Additionally, integrated promoter/enhancer regions can influence an area surrounding the integration over a distance possibly as large as 500 kbp (4). Studies of deregulation of the virus integration locus should therefore be focused on such large areas and might include up to 1Mbp regions. With this in mind, our tool generates two outputs for Ensembl hits: one file that contains every gene located within a user-defined window around the integration site (default is 100 kbp), and another file which only contains the gene whose 5' end is closest to each site. Assigning one unique 'closest' gene to each integration site allows for easier automated downstream analysis such as gene expression level binning, but user investigation of all the genes near each integration site is highly recommended.

Impact of analysis settings on annotation output

We have already discussed how repeat masking, RefSeq filtering and the size of the window around the VIS can affect the annotations returned. Additionally, changes between versions of genome builds can have an effect. Table 1a shows how the number of sequences hit and the number of genes identified increase with successive builds of the human genome. Table 1b illustrates how an identical dataset can produce different outputs depending on which version of NCBI was used by the BLAST algorithm. In particular the gene symbols are updated frequently. For instance, comparing the outputs from v34 and v35, we find that 95% of the aligned sequences are identical, but only 75% of the Ensembl IDs for those sequences match, and only 53% of the gene names are identical. We therefore recommend including the Ensembl ID, which is more stable, for any large scale analysis.

Comparison of MAVRIC and a published analysis for ADA-SCID gene therapy

We compared MAVRIC annotation and UCSC annotation on a dataset retrieved from a clinical gene therapy trial for ADA-SCID, patients transplanted with $CD34^+$ cells

transduced with the adenosine deaminase (*ADA*) gene using a retroviral vector (19). 708 total sequences were annotated. We ran 706 of those sequences through MAVRIC, which returned a varying number of annotated results depending on the pre-set RepeatMasker and e-value conditions. Increasing the e-value cutoff allows lower quality alignments to be annotated, while using repeat masking might increase the number of annotated alignments (Table 2). The mean number of alignments retrieved using any of these settings is, however, largely unaffected. For comparison to the published

Table 2. Comparison of the difference between the number integrations with at least one Refseq genes within 100kb identified when altering analysis parameters. Reducing the permissiveness of the BLAST e-value 100-fold has little effect on the number of sequences annotated, however using repeatmasking increases the number of sequences aligned to unique spots in the genome by approximately 25%.

Analysis parameters	Ensembl Hits	Mean number of alignments per hit (range)
No repeatmasking E-value = 0.01	395	3.11 (1-18)
No repeatmasking E-value = 1	419	3.06 (1-18)
With repeatmasking E-value = 0.01	511	3.32 (1-69)
With repeatmasking E-value = 1	514	3.35 (1-20)

data, we used the MAVRIC output using the following criteria: human RepeatMasker, minimum sequence length = 11bp (the lowest allowed by the BLAST algorithm), maximum e-value = 1. These criteria returned 608 annotated genes. The principal factor in the lower number of sequences returned by MAVRIC is the minimum sequence length requirement. On this basis, MAVRIC excluded 2 sequences before repeat masking, and an additional 63 sequences after repeat masking, to a total of 65 excluded sequences. This accounts for 66% of the sequences skipped by MAVRIC. Other sequences without annotated genes could be attributed to ambiguous BLAST results, ie multiple hits with identical lowest e-values, or no RefSeq genes being found within 500 kbp (see Supplemental Table 2). The distribution of VIS is similar between the two methods (Table 3a). For the combined sequences, MAVRIC analysis returned a 37%/63% split between intragenic and intergenic integrations, while the published analysis had a 44%/56% split. The percentage of integration sites less than 30 kbp upstream, 10 kbp upstream and within 5 kbp of the transcription start site (TSS) are likewise in agreement for the two methods. Similar common integration sites were found by the two methods (Table 3b). Figures illustrating the distribution of VIS around the nearest gene TSS are also comparable between the two analysis methods (Figure 3). MAVRIC found the same gene as the Aiuti analysis in 459 cases (76%). A list of the sequences where the two analysis methods returned different genes can be found in Supplemental Table 3.

Table 3a. Retroviral integration site distribution in hematopoetic stem cells from ADA-SCID patients, comparison between analysis methods.

	Number of hits	Intragenic	Intergenic	<30kb Upstream	<10kb Upstream	+/- 5kb from TSS
Aiuti In Vitro	212	50.90%	49.10%	19.40%	12.30%	23.60%
MAVRIC In Vitro	188	42.00%	58.00%	20.20%	12.80%	25.50%
Aiuti Ex Vivo	496	41.30%	58.70%	25.60%	19.60%	28.80%
MAVRIC Ex Vivo	420	33.80%	66.20%	31.40%	23.80%	32.10%
Aiuti Total	708	44.30%	55.70%	23.70%	17.40%	27.30%
MAVRIC Total	608	36.50%	63.50%	28.00%	20.40%	29.80%

Table 3b. Comparison of some common integration sites found by both analysis methods.

	Number of hits within window	
Gene	Aiuti *et al.*	MAVRIC
LMO2	6	6
BCL2	4	4
BLM	3	3
DYRK1A	3	3
CCND2	3	4
RNPC1	3	3
BTN3A2	2	2
MRPL39	2	3
RYBP	2	3
BHLHB2	1	3

ADA-SCID integrations compared to XSCID integration data

To assess to what extent the ADA-SCID data analyzed in this study overlaps with data from the XSCID clinical trial, previously published data from the London and Paris XSCID trials (20,21). The ADA SCID (19) and XSCID trials used similar gamma-retroviral vectors, which is reflected in a remarkable overlap (10.3-15.6%, Table 4). Ingenuity analysis of the ADA-SCID data shows that the gene functions being hit by the virus are very similar to those described for the XSCID and mouse integrations. The gene functions that were targeted were Gene Expression, Cell Death, Immune System Disorders, Cancer, and Hematological Disease, consistent with the target cells (supplementary Figure 2).

DISCUSSION

Automated analysis tools allow time efficient annotation of virus integration sites according to a predefined set of criteria, which makes the results less prone to user input

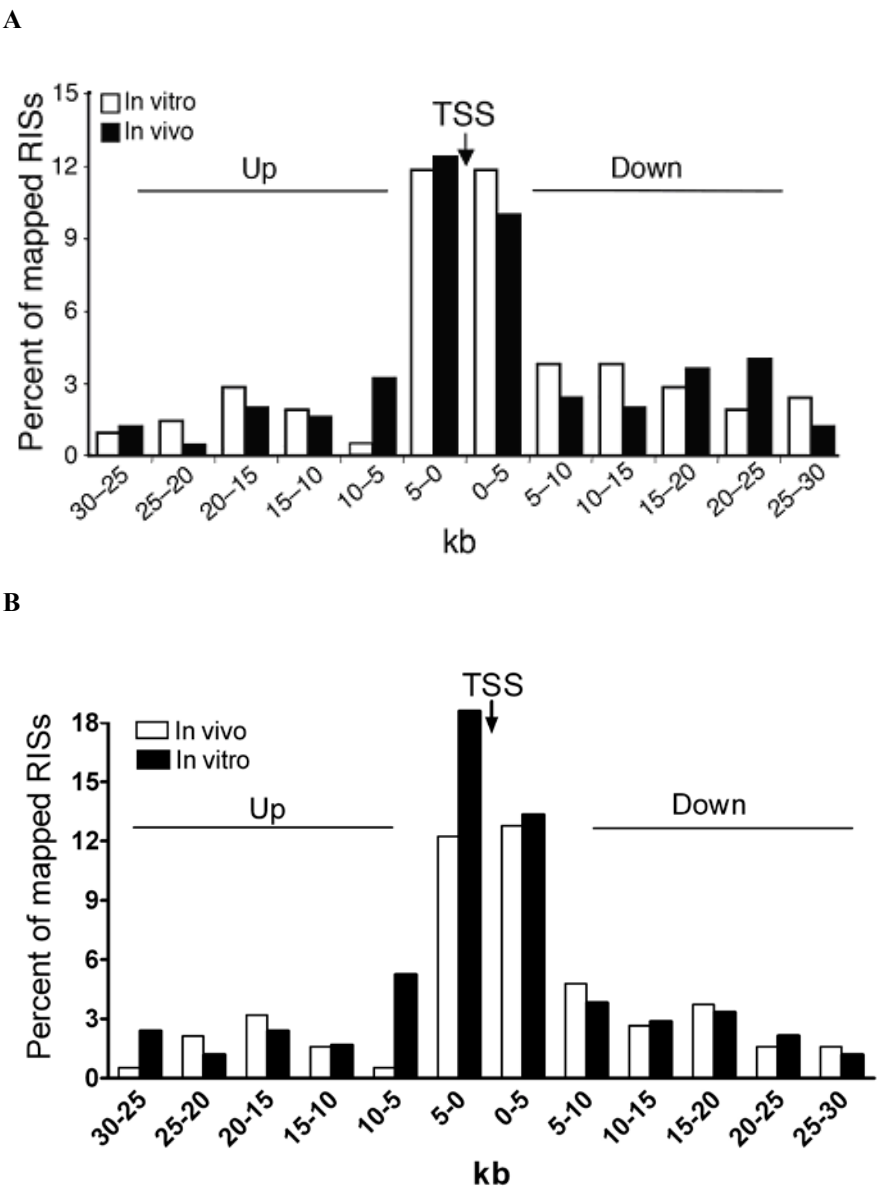


Figure 3. Distribution of retroviral integration site distances from nearest gene transcription start site (TSS) in ADA-SCID patients, based on published data (A) or MAVRIC annotation (B).

errors and allows for easy re-analysis in case annotations or database builds change. We created the MAVRIC web tool to automatically quality check integration site sequences, align those using BLAST and annotate using data retrieved from the Ensembl databases. MAVRIC removes repeat sequences, excludes sequences that fail the minimum length

Table 4. Genes co-occuring between X-SCID clinical data and ADA-SCID clinical data.

	ADA-SCID dataset (1536 genes)
X-SCID Paris (481 genes)	75 (15.6%)
X-SCID London (523 genes)	54 (10.3%)
X-SCID Combined (926 genes)	112 (12.1%)
X-SCID London 454 data (2089 genes)	263 (12.6%)

criteria, returns a list of genes near each integration site and annotates the genomic loci (Figure 1).

MAVRIC takes a collection of genomic sequences flanking viral integration sites and prepares those for analysis via several quality control checks (Figure 1: Flow chart of MAVRIC). First, MAVRIC uses RepeatMasker to screen and remove repeat sequences. It excludes sequences that are below a specified length to ensure a high number of unambiguous alignments. The remaining sequences are aligned to the genome of the appropriate species (rhesus, mouse or human) via BLAST. Sequences returning ambiguous BLAST hits are excluded from further analysis and should be investigated by hand. The details of the genomic loci surrounding the integration sites are annotated and packaged into a zip file along with the analysis parameters.

To address the need for sequence alignments to large genomes, BLAT (BLAST Like Alignment Tool) (22) was developed to handle alignments up to 500 times faster than BLAST. BLAT is currently used in the UCSC genome browser and Ensembl. It allows efficient alignments of smaller sequences and is optimized to recognize separate islands of perfect matches, which would be generated by the alignment of cDNA. LAM-PCR is widely used for the identification of viral insertion sites, (23). Due to the availability of restriction sites in the vicinity of the virus integration, each insertion will usually yield one specific amplicon. In samples obtained from mouse transplantation experiments

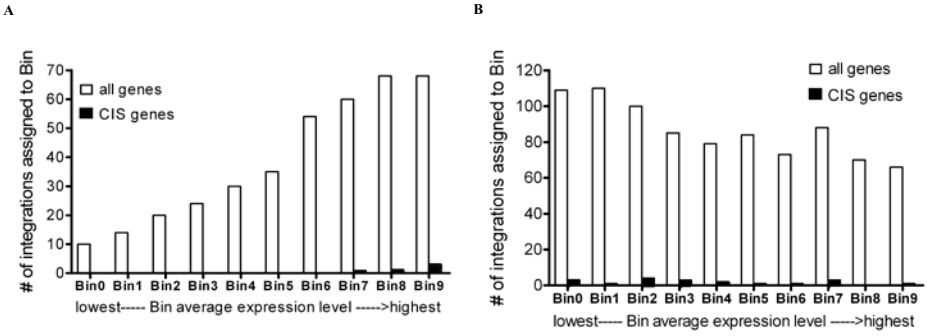


Figure 4: (A) Expression level binning of genes identified in ADA SCID trial via MAVRIC analysis/ Highly expressed genes are more likely to have VIS. All of the CIS genes found in this dataset were assigned to the three highest bins. Expression levels were generated from human HSC using an Affymetrix U133 plus 2.0 array. (B) Similar analysis of a dataset consisting of 1000 randomly generated in silico integrations.

we usually observe a limited number of amplicons. Pyrosequencing experiments, which can contain more than 500 thousand sequences, usually produce multiple sequences for each individual insertion. To remove unnecessary complexity in these datasets, highly similar sequences can be removed to decrease the amount of time necessary for BLAST alignments. Current hardware is then fast enough to perform BLAST alignments on a scale of thousands of integrations on commodity hardware. BLAST has been the *de facto* standard for sequence alignments, meaning that there are convenient ways of incorporating the Ensembl API into the current program; thus we chose to develop a system using BLAST alignments. To efficiently process larger datasets, we have employed a repeat masking strategy, which speeds up BLAST alignments and results in more alignments which can unambiguously be aligned to the genome. With this feature, more than 250 sequences may be annotated per hour. As larger sequencing batches become the norm, the number of processors running BLAST in parallel in can be increased in order to speed up sequence annotation.

Recently, insertion site recovery using LAM-PCR and similar restriction-site-dependent techniques was shown to rely on the availability of restriction sites in the vicinity of the insertion. Insertion sites lacking a nearby restriction site for the endonuclease of choice cannot be recovered. Improved methods circumvent this so called 'restriction bias' by using methods independent of the restriction sites surrounding virus insertion sites (24,25). Although these new approaches provide a increased coverage of the integrome, they also produce more complex datasets due to the presence of several different sequences for each single insertion, which will take considerably longer to align using current technology.

Using a default sequence length cut-off of 25bp allows us to restrict the number of integrations to those which can be aligned with high confidence and at the same time avoid excessively long alignment times. Smaller cut-offs can be used but the number of annotated sequences returned may be reduced. Since these processes are performed by a standardized, integrated analysis, a higher throughput is achieved than what can be offered using online BLAST alignment. The databases and annotation files are stored on a server, which allows analysis of a dataset with previous genome builds. This is of interest when new datasets need to be compared to earlier analyses. Alternatively, old data can easily be reanalyzed to conform to the newest genome builds and annotations. Such an updated analysis is necessary because the genome assemblies are updated with new sequence information and annotations roughly every half year[27] and although sequencing information has nearly been completed (except for centromeric and repeat regions), new and updated gene annotations are continuously introduced (e.g. build 36.3, March

27 http://www.ncbi.nlm.nih.gov/genome/guide/human/release_notes.html

2008, introduced the annotation of non coding RNAs). The most recent human genome build to introduce new sequence information was build 36.1 (March 2006).

Obviously, a method to annotate integrations following a defined set of criteria is less likely to make mistakes, such as typing errors. Yet there is a problem even when reporting the automatically annotated integrations. Usually, spreadsheets are a convenient way of displaying moderately sized datasets. The automatic correction methods built in to spreadsheets such as Microsoft Excel, however, can cause difficulties when the correction methods are used on gene symbols. This problem has been reported previously (26) and has also occurred in our analysis.

The difference in datasets obtained using RefSeq filtered versus unfiltered demonstrates how analysis results can be altered by merely changing annotation settings. It is therefore necessary to very explicitly state all parameters used in the analysis to be able to reproduce the exact alignments described. Even when all sequences are available, these parameters are necessary for a reproduction of alignments and annotations.

Comparing the output of the automated MAVRIC tool to the published analysis of an ADA-SCID gene therapy study demonstrates large similarities in the overall integration patterns while highlighting subtle differences in the annotation processes. The distributions of integration site distance from TSS were very similar, as were the common integration sites. Some of the sequences identified in the ADA- SCID study were excluded by MAVRIC due to failure to achieve a quality threshold (primarily minimum sequence length before or after repeat masking) and should be analyzed manually. MAVRIC successfully annotated 73% of the ADA SCID sequences in approximately 2 hours, representing a vast improvement in analysis throughput.

Integration site analysis of ADA-SCID and X-SCID clinical trials revealed a large overlap in genes that have nearby viral integrations. Ingenuity analysis showed the same functional pathways as frequent integration targets in both clinical trials and in mouse studies. This suggests that closely related vector types will have similar integration profiles across multiple treatments and species.

In basic structure, MAVRIC is similar to other automated sequence analysis tools such as QuickMap (12). QuickMap's advantages are the speed at which it analyzes sequence batches and the very elaborate output it provides on genomic features including repeats, transcription factor binding sites and fragile sites. MAVRIC's output produces less raw data, but it is more gene-based and presented in a way to highlight the potential safety implications of the viral integration patterns. MAVRIC also uses microarray data from HSCs to analyze the expression levels of VIS genes. Compared to QuickMap, MAVRIC has more front-end options for the user with regards to the analysis parameters (repeat masking, sequence length and e-value thresholds, RefSeq genes versus all genomic features, etc).

In summary, automated analysis tools such as MAVRIC can greatly increase the speed and efficiency of viral integration site annotation. MAVRIC returns detailed information on the genomic location of the integration sites and gives an overview of viral integration patterns. It is well-suited to generating annotations to compare historical datasets or to compare data from different labs. However, the user should carefully choose the quality thresholds and note which database versions are used, as small changes in the analysis parameters can have a significant effect on the resulting annotations. Re-analyzing datasets with identical analysis parameters helps to ensure that the resulting annotations are optimally suited for direct comparison to enable rapid cumulative data processing. As high throughput sequencing becomes routine and more gene therapy applications move towards clinical implementation, automated analysis tools will become essential for rapid patient monitoring.

Acknowledgments

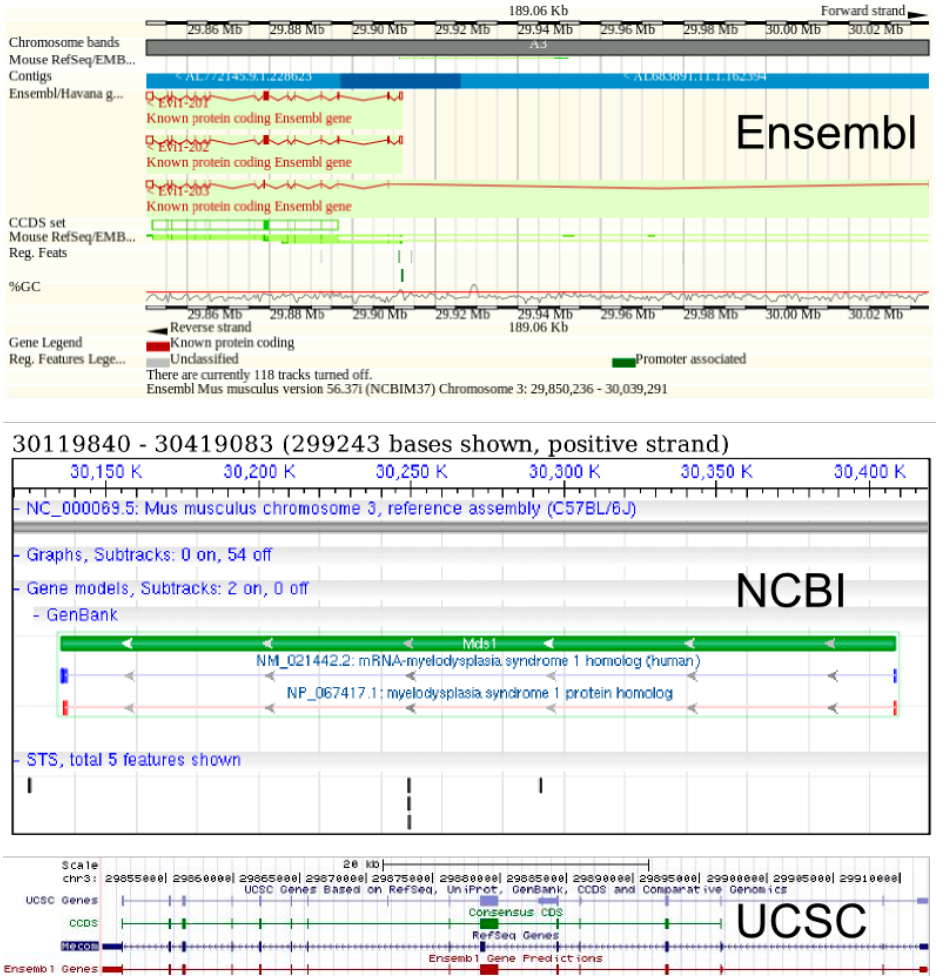
The authors acknowledge Drs. F.J.T. Staal and K. Pike-Overzet, Dept. Immunology, Erasmus University Medical Center, for assistance in the microarray analyses.

REFERENCES

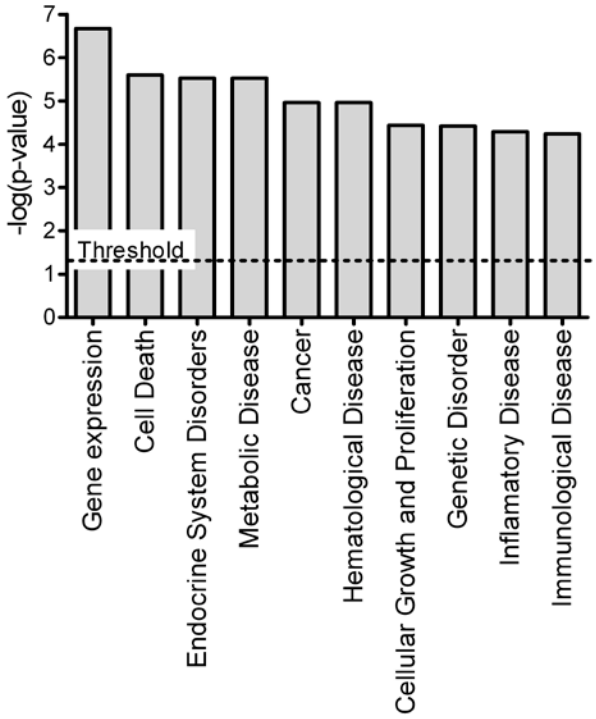
1. Mikkers H, Allen J, Knipscheer P, Romeijn L, Hart A, Vink E, et al. High-throughput retroviral tagging to identify components of specific signaling pathways in cancer. Nat Genet. 2002;32:153-9.
2. Joosten M, Vankan-Berkhoudt Y, Tas M, Lunghi M, Jenniskens Y, Parganas E, et al. Large-scale identification of novel potential disease loci in mouse leukemia applying an improved strategy for cloning common virus integration sites. Oncogene. 2002;21:7247-55.
3. Erkeland SJ, Verhaak RG, Valk PJ, Delwel R, Lowenberg B, Touw IP. Significance of murine retroviral mutagenesis for identification of disease genes in human acute myeloid leukemia. Cancer Res. 2006;66:622-6.
4. Kustikova O, Fehse B, Modlich U, Yang M, Dullmann J, Kamino K, et al. Clonal dominance of hematopoietic stem cells triggered by retroviral gene marking. Science. 2005;308:1171-4.
5. Modlich U, Kustikova OS, Schmidt M, Rudolph C, Meyer J, Li Z, et al. Leukemias following retroviral transfer of multidrug resistance 1 (MDR1) are driven by combinatorial insertional mutagenesis. Blood. 2005;105:4235-46.
6. 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.
7. Cattoglio C, Facchini G, Sartori D, Antonelli A, Miccio A, Cassani B, et al. Hot spots of retroviral integration in human CD34+ hematopoietic cells. Blood. 2007;110(6):1770-8.
8. Montini E, Cesana D, Schmidt M, Sanvito F, Ponzoni M, Bartholomae C, et al. Hematopoietic stem cell gene transfer in a tumor-prone mouse model uncovers low genotoxicity of lentiviral vector integration. Nat Biotechnol. 2006;24:687-96.
9. Zychlinski D, Schambach A, Modlich U, Maetzig T, Meyer J, Grassman E, et al. Physiological promoters reduce the genotoxic risk of integrating gene vectors. Mol Ther. 2008;16(4):718-25.
10. Kustikova OS, Modlich U, Fehse B. Retroviral insertion site analysis in dominant haematopoietic clones. Methods Mol. Biol. 2009;506:373-390.

11. Peters B, Dirscherl S, Dantzer J, Nowacki J, Cross S, Li X, et al. Automated analysis of viral integration sites in gene therapy research using the SeqMap web resource. Gene Ther. 2008 Sep;15(18):1294-1298.
12. Appelt J, Giordano FA, Ecker M, Roeder I, Grund N, Hotz-Wagenblatt A, et al. QuickMap: a public tool for large-scale gene therapy vector insertion site mapping and analysis. Gene Ther. 2009 Jul;16(7):885-893.
13. Pruitt KD, Tatusova T, Maglott DR. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. 2007 Jan;35(Database issue):D61-65.
14. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215(3):403-10.
15. Bartholomew C, Ihle JN. Retroviral insertions 90 kilobases proximal to the Evi-1 myeloid transforming gene activate transcription from the normal promoter. Mol Cell Biol. 1991;11:1820-8.
16. Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, et al. The diploid genome sequence of an individual human. PLoS Biol. 2007 Sep 4;5(10):e254.
17. Flicek P, Aken BL, Beal K, Ballester B, Caccamo M, Chen Y, et al. Ensembl 2008. Nucleic Acids Res. 2008 Jan;36(Database issue):D707-714.
18. Ning Z, Cox AJ, Mullikin JC. SSAHA: a fast search method for large DNA databases. Genome Res. 2001;11(10):1725-9.
19. Aiuti A, Cassani B, Andolfi G, Mirolo M, Biasco L, Recchia A, et al. Multilineage hematopoietic reconstitution without clonal selection in ADA-SCID patients treated with stem cell gene therapy. J Clin Invest. 2007;117(8):2233-40.
20. Schwarzwaelder K, Howe SJ, Schmidt M, Brugman MH, Deichmann A, Glimm H, et al. Gammaretrovirus-mediated correction of SCID-X1 is associated with skewed vector integration site distribution *in vivo*. J Clin Invest. 2007;117(8):2241-9.
21. Deichmann A, Hacein-Bey-Abina S, Schmidt M, Garrigue A, Brugman MH, Hu J, et al. Vector integration is nonrandom and clustered and influences the fate of lymphopoiesis in SCID-X1 gene therapy. J Clin Invest. 2007;117(8):2225-32.
22. Kent WJ. BLAT--the BLAST-like alignment tool. Genome Res. 2002;12:656-64.
23. Schmidt M, Hoffmann G, Wissler M, Lemke N, Mussig A, Glimm H, et al. Detection and direct genomic sequencing of multiple rare unknown flanking DNA in highly complex samples. Hum Gene Ther. 2001;12:743-9.
24. Gabriel R, Eckenberg R, Paruzynski A, Bartholomae CC, Nowrouzi A, Arens A, et al. Comprehensive genomic access to vector integration in clinical gene therapy. Nat. Med. 2009 Dec;15(12):1431-1436.
25. Pule MA, Rousseau A, Vera J, Heslop HE, Brenner MK, Vanin EF. Flanking-sequence exponential anchored-polymerase chain reaction amplification: a sensitive and highly specific method for detecting retroviral integrant-host-junction sequences. Cytotherapy. 2008;10(5):526-539.
26. Zeeberg BR, Riss J, Kane DW, Bussey KJ, Uchio E, Linehan WM, et al. Mistaken identifiers: gene name errors can be introduced inadvertently when using Excel in bioinformatics. BMC Bioinformatics. 2004;5:80.

SUPPLEMENTARY DATA



Supplementary figure 1. The MECOM locus as annotated by Ensembl, NCBI and UCSC.



Supplementary figure 2. Ingenuity functional pathways analysis of genes identified by MAVRIC as nearest to the ADA-SCID clinical trial VIS sequences.

Supplementary Table 1. Evi1 annotations in NCBI, Ensembl and UCSC.

Database	Annotated start site	Annotated end site
NCBI	29851638	29896586
Ensembl	29850236	30039291
UCSC	29851638	30039435

Supplementary Table 2. Aiuti *et al*. sequences not found by MAVRIC.

Sequence ID	EnsGene ID Aiuti*	Gene Name	Chromosome	MAVRIC result
S1_001	125968	ID1	20	Eval > 1
S1_006	180481	NP_689649.1	12	Eval > 1
S1_007	123179	EBPL	13	Eval > 1
S1_008	101000	PROCR	20	Eval > 1
S1_020	145526	CDH18	5	No gene hit
S1_027	7129	NP_291021.2	19	Eval > 1
S1_031	187839	XP_497922.1	3	Eval > 1
S1_037	120457	KCNJ5	11	Eval > 1
S1_041	29153	ARNTL2	12	Eval > 1
S1_048	130775	Q9NS90_HUMAN	1	Eval > 1
S1_055	64763	MLSTD1	12	Eval > 1
S1_065	92200	RPGRIP1	14	Eval > 1
S1_067	155034	FBXL18	7	Eval > 1
S1_082	140379	BCL2A1	15	Eval > 1
S1_084	146416	AIG1	6	Eval > 1
S2_004	5889	ZFX	X	Eval > 1
S2_016	187268	FAM9C	X	Eval > 1
S2_017	7047	MARK4	19	Eval > 1
S2_025	174839	NP_689891.1	3	Eval > 1
S2_026	25434	NR1H3	11	Eval > 1
S2_029	107951	PAPD1	10	Eval > 1
S2_030	177673	NP_689827.1	2	Eval > 1
S2_038	133872	NP_057211.4	8	Eval > 1
S2_P011	119866	BCL11A	2	Eval > 1
S2_P012	103479	RBL2	16	Eval > 1
S2_P014	183621	NP_877432.1	10	Eval > 1
S2_P035	120656	TAF12	1	Eval > 1
S2_P057	173301	NP_659487.1	8	No gene hit
S3_009	170873	MTSS1	8	Eval > 1
S3_011	180029	Q8NF14_HUMAN	1	Eval > 1
S3_020	111452	Q9NSM3_HUMAN	12	Eval > 1
S3_022	120563	LYZL1	10	Ambiguous
S3_023	137193	PIM1	6	Eval > 1
S3_028	175600	C7orf10	7	No gene hit
S3_030	79134	THOC1_HUMAN	18	Eval > 1
S3_034	134453	RBM17	10	Eval > 1
S3_035	164300	C5orf12	5	Eval > 1
S3_057	120645	Q9UPP2_HUMAN	12	Eval > 1
S3_064	166272	C10orf26	10	Eval > 1
S3_072	91972	CD200	3	Eval > 1
S3_073	165023	DIRAS2	9	Eval > 1
S3_077	172493	MLLT2	4	Eval > 1
S3_086	179573	Q9IUV1_HUMAN	19	Eval > 1
S3_088	172538	C10orf73	10	Eval > 1
S3_095	152520	NP_787050.3	13	Eval > 1

Sequence ID	EnsGene ID Aiuti*	Gene Name	Chromosome	MAVRIC result
S3_103	64999	ANKS1	6	Eval > 1
S3_113	196139	AKR1C3	10	Eval > 1
S3_114	185736	ADARB2	10	Eval > 1
S3_127	198589	LRBA	4	Eval > 1
S3_128	107165	TYRP1	9	No gene hit
S3_P028	176635	HORMAD2	22	Eval > 1
S3_P031	100219	XBP1	22	Eval > 1
S3_P032	186395	KTR10	17	Eval > 1
S3_P033	159674	SPON2	4	Eval > 1
S3_P044	132967	Q9NYD7_HUMAN	3	No gene hit
S3_P048	102606	ARHGEF7	13	Eval > 1
S3_P099	196891	Q8NED3_HUMAN	12	No gene hit
S4_006	115934	PPIL3	2	Eval > 1
S4_022	149179	NP_077018.1	11	Ambiguous
S4_030	167632	Q96Q05_HUMAN	8	Eval > 1
S4_031	183690	EFHC2	X	Ambiguous
S4_039	125510	OPRL1	20	Eval > 1
S4_041	170638	YV03_HUMAN	22	Eval > 1
S4_042	106992	AK1	9	Eval > 1
S4_P003	179177	Q87NC93_HUMAN	12	Eval > 1
S4_P008	170577	SIX2	2	Eval > 1
S4_P026	182805	XP_498131.1	6	No gene hit
S4_PO32	103449	SALL1	16	Eval > 1
S5_002	168546	GFRA2	8	No gene hit
S5_003	105402	NAPA	19	Eval > 1
S5_018	105492	SIGLEC6	19	Eval > 1
S5_031	168405	NR_002174.1	6	Eval > 1
S5_032	49618	Q8TEE4_HUMAN	6	No gene hit
S5_036	101255	TRIB3	20	Eval > 1
S5_058	153930	NP_694960.1	17	Eval > 1
S5_059	105643	ARRDC2	19	Eval > 1
S5_062	185189	NRBP2	8	Eval > 1
S5_069	160219	GAB3	X	Eval > 1
S5_078	138670	RASGEF1B	4	Eval > 1
S5_079	143013	LMO4_HUMAN	1	No gene hit
S5_083	198265	HELZ	17	Eval > 1
S5_086	160310	HRMT1L1	21	Eval > 1
S5_089	125952	MAX_HUMAN	14	Eval > 1
S5_099	101405	OXT	20	Eval > 1
S5_102	167258	CD2L7_HUMAN	17	Eval > 1
S5_107	170476	Q8WU39_HUMAN	5	Eval > 1
S5_114	171791	BLC2	18	Eval > 1
S5_131	90975	PITPNM2	12	Eval > 1
S5_132	178104	PDE4DIP	1	Eval > 1
S5_136	49618	Q8TEE4_HUMAN	6	No gene hit
S5_149	151287	NP_620134.2	13	Eval > 1

Sequence ID	EnsGene ID Aiuti*	Gene Name	Chromosome	MAVRIC result
S5_166	176657	EPS15L2	7	No gene hit
S5_P001	103489	XYLT1	16	Eval > 1
S5_P003	8083	JARID2	6	No gene hit
S5_P016	147697	MLZE	8	Eval > 1
S5_P028	66827	ZNF406	8	Eval > 1
S5_P033	174408	XM_063084.4	13	Ambiguous
S5_P035	170017	ALCAM	3	No gene hit
S5_P060	197218		2	Eval > 1

Supplementary Table 3. Annotation mismatches between MAVRIC and Aiuti *et al*.

Sequence ID	ENSGene ID Aiuti	Gene name Aiuti	ENSGene ID MAVRIC	Gene name MAVRIC	Chromosome
S1_004	148175	STOM	136848	DAB2IP	9
S1_009	112799	LY86	175753	NM_173675	6
S1_014	149617	Q6P094_HUMAN	171940	ZNF217	20
S1_019	169918	C15orf16	179964	XR_018466.1	15
S1_024	188092	GPR89	117262	GPR89A	1
S1_036	149617	Q6P094_HUMAN	171940	ZNF217	20
S1_044	102804	TB1I4_HUMAN	151778	C13orf21	13
S1_051	141748	XP_372668.2	161381	PLXDC1	17
S1_054	198769	XP_293026.4	115841	FAM82A	2
S1_060	145133	FAM44A	38219	FAM44A	4
S1_061	155093	PTPRN2	105993	DNAJB6	7
S1_062	150995	NP_002213.1	134107	BHLHB2	3
S1_069	141564	Q96C97_HUMAN	173814	LOC732236	17
S1_073	168675	C18orf1	177150	C18orf19	18
S1_083	188675	227291	157617	C21orf25	21
S1_085	70366	EST1A_HUMAN	177374	HIC1	17
S1_090	21776	AQR	159251	ACTC1	15
S1_093	196301	HLA-DRB9	204287	HLA-DR	6
S1_101	165966	PDRN4_HUMAN	15153	YAF2	12
S1_106	26652	AGPAT4	203702	C6orf59	6
S1_116	120910	PPP3CC	120896	SORBS3	8
S1_117	176966	952	120129	DUSP1	5
S2_002	167186	COQ7	205730	NP_001030013.1	16
S2_020	198294	NP_997209.1	143653	SCCPDH	1
S2_023	69020	MAST4	134061	CD180	5
S2_036	140577	Q8NF38_HUMAN	197299	BLM	15
S2_037	188739	K0117_HUMAN	207181	snoACA14	1
S2_039	163590	PPM1L	169255	B3GALNT1	3

Sequence ID	ENSGene ID Aiuti	Gene name Aiuti	ENSGene ID MAVRIC	Gene name MAVRIC	Chromosome
S2_043	96682	C6orf48	204388	HSPA1B	6
S2_047	186284	RUNX2	196284	SUPT3H	6
S2_P017	145491	ROPN1L	164236	LOC651746	5
S2_P024	26103	Q8IUB6_HUMAN	107796	ACTA2	10
S2_P025	38532	NP_056041.1	175643	C16orf75	16
S2_P029	111412	NP_079014.1	171471	KRTHB5	12
S2_P030	163084	TSN	211460	TSN	2
S2_P031	197959	DNM3	180999	C1orf105	1
S2_P036	159100	CU062_HUMAN	205929	C21orf62	21
S2_P044	157654	PALM2	188959	UCP1	4
S2_P045	185433	CU042_HUMAN	154719	MRPL39	21
S2_P054	99204	ABLIM1	169129	AFAP1L2	10
S3_001	185578	Q8NCT9_HUMAN	140563	MCTP2	15
S3_002	107104	Q86TE2_HUMAN	137090	DMRT1	9
S3_005	175925	XP_498446.1	95739	BAMBI	10
S3_013	198051	NP_998889.1	179399	GPC5	13
S3_018	154678	PDE1C	106341	C7orf16	7
S3_021	179050	MAFB	204103	MAFB	20
S3_032	165240	ATP7A	186076	PGAM4	X
S3_048	145133	FAM44A	38219	FAM44A	4
S3_050	196883	XP_496207.1	205045	Q6IEE8_HUMAN	17
S3_053	184588	PDE4B	118473	SGIP1	1
S3_056	198179	BCL2L7P1	149600	COMMD7	20
S3_060	141510	TP53	141499	WDR79	17
S3_061	89505	CKLF	140932	CMTM2	16
S3_062	100599	RIN3	100600	LGMN	14
S3_063	167632	Q97Q05_HUMAN	211505	C8orf17	8
S3_080	162896	PIGR	162894	FAIM3	1
S3_089	158062	NP_663321.1	142669	SH3BGRL3	1
S3_098	196951	Q8N984_HUMAN	196782	MAML3	4
S3_104	165915	SLC39A13	66336	SPI1	11
S3_106	107242	PIP5K1B	165059	PRKACG	9
S3_108	140577	Q8NF38_HUMAN	197299	BLM	15
S3_122	187768	Q5WOV6_HUMAN	89177	C20orf23	20
S3_125	182778	NP_787066.1	211978	IGHV5-78	14
S3_133	196876	SCN8A	135503	ACVR1B	12
S3_135	184588	PDE4B	118473	SGIP1	1
S3_137	179951	384844	184497	FAM70B	13
S3_P001	196186	130368	107175	CREB3	9

Sequence ID	ENSGene ID Aiuti	Gene name Aiuti	ENSGene ID MAVRIC	Gene name MAVRIC	Chromosome
S3_P002	165322	Q9NV28_HUMAN	204332	Q6ZUZ5_HUMAN	10
S3_P005	170310	STX8	65320	NTN1	17
S3_P007	179573	Q8IUV1_HUMAN	105738	SIPA1L3	19
S3_P015	153944	MSI2H_HUMAN	181610	MRPS23	17
S3_P016	128833	MYO5C	69966	GNB5	15
S3_P024	150995	NP_002213.1	134107	BHLHB2	3
S3_P039	151240	KIAA0934	15171	ZMYND11	10
S3_P056	183977	NP_653315.1	114166	PCAF	3
S3_P063	124535	WRNIP1	145949	SGK85_HUMAN	6
S3_P068	173848	NET1	178372	CALML5	10
S3_P070	100490	CDKL1	125375	ATP5S	14
S3_P072	196256	Q13715_HUMAN	196252	ADRB2	5
S3_P088	187613	Q7Z3I4_HUMAN	118515	SGK	6
S3_P095	103222	ABCC1	91262	ABCC6	16
S4_002	148513	ANKRD30A	198105	ZNF248	10
S4_003	132003	ZSWIM4	187556	NANOS3	19
S4_015	59122	Q9BQG6_HUMAN	131650	KREMEN2	16
S4_021	172081	MOBKL2A	99840	C19orf36	19
S4_025	63438	AHRR	180104	EXOC3	5
S4_026	136560	TANK	115233	PSMD14	2
S4_027	163832	Q9BW57_HUMAN	114650	SCAP	3
S4_029	165296	XP_291767.2	204656	0	10
S4_033	143569	UBAP2L	143612	C1orf43	1
S4_036	104964	AES	65717	TLE2	19
S4_040	196712	NF1	185862	EVI12B	17
S4_051	160584	NP_079440.2	118137	APOA1	11
S4_P004	183346	C10orf107	150347	ARID5B	10
S4_P006	134030	KIAA0427	101665	SMAD7	18
S4_P009	130076	IGHG3	198299	LOC649910	14
S4_P018	146684	KCTD7	154710	KCTD7	7
S4_P022	91986	NP_955805.1	206531	NP_001008784.4	3
S4_P024	166056	Q6PJ56_HUMAN	129562	DAD1	14
S4_P029	124203	C20orf174	124205	EDN3	20
S5_004	65809	C10orf45	151474	FRMD4A	10
S5_007	180891	CUEDC1	181610	MRPS23	17
S5_015	153956	CACNA2D1	19991	HGF	7
S5_019	188517	COL25A1	164089	AGXT2L1	4
S5_025	164744	SUNC1	136273	HUS1_HUMAN	7
S5_029	108753	TCF2	185128	LOC729873	17

CHAPTER 7

General Discussion

The safety aspect of viral gene therapy with the aim to restore gene function by the addition of a functional copy of the affected gene has become an important aspect of gene therapy after discoveries of leukemia development in a preclinical mouse model (1), XSCID patients in a gene therapy trial (2,3,4) and in a preclinical rhesus monkeys model (5).

In all studied species, these events occurred late (3 years after transplantation in the XSCID trial) after transduction and transplantation, leaving many open questions about the fate of the transduced clones and their progeny. From studies addressing the nature of the HSC, it is known that the $CD34^+$ cell population that is used in the XSCID (6), ADA SCID (7) and XCGD (8) studies only contains a small fraction of cells with repopulating capacity. Transduction of this rather heterogeneous fraction of cells will therefore hit a large number of cells that are not useful for repopulation as they lack repopulation capacity. It has however been shown using LSK subpopulations, which contain a higher fraction of hematopoietic stem cells in mice (9), and using mature T cells in mice (10) that the more differentiated cells in the bone marrow are not the main source of cells that give rise to leukemia. One of the initial concepts to reduce the risk (5 malignant proliferations in 9 treated patients in the XSCID trials) of leukemia occurrence in the patient was reduction of the number transplanted cells by using more pure populations of HSC. Since the source of the leukemia cells seems to be exactly this more purified cell fraction, this approach will probably not result in a reduction of leukemia occurrence. The number of cells of this specific cell fraction might still be reduced with the same effect.

For these reasons, screening the pre-transplant sample for possible dangerous integrations is not feasible if a $CD34^+$ bone marrow or PBMCs is transduced. First of all, the cell samples taken from the transduced pool do not end up in the patient and therefore account for cells that have not been transplanted. Furthermore, since 0.9-$20x10^6$ transduced cells/kg, mean $7x10^6$/kg (6-7,11) were transplanted on average and the cells that give rise to repopulating progeny only form a small fraction (0.05%-0.1%) (12) of this population, it will be exceedingly hard to draw conclusions about the cells that do actually repopulate even from high throughput sampling these cells.

Given this, safety assessment of viral gene therapy can only be based on preclinical animal and *in vitro* models. Given that in mouse (this thesis and 13) and rhesus monkey models (5) leukemia takes very long (50 weeks, in secondary transplantation, in mouse and 5 years in the rhesus monkey model) to occur, it is probably more feasible to test leukemogenic effects in *in vitro* model (14-16). While mouse and rhesus monkey transplantation models, which employ protocols close to those used in clinal gene therapy can provide insight in cell fate *in vivo*, they seem less suitable for vector development, where results about changes in architecture need to be visible faster to allow efficient development. In this situation, *in vitro* models will be applicable best. The *in vitro*

immortalization models are only able to produce surrogate incidences, mainly because each specific model has specific requirements (overcoming growth factor dependence by upregulation of Ghr in the model described by Bokhoven (15), introduction of autonomous growth by expression of Evi1 in the IVIM assay (16)) that need to be met before clones develop. These clones are then counted to determine immortalization frequency.

Similar problems exist when using tumor prone mice (17,18), because the deletions or expressed oncogenes that are present in these mouse models reduce the number of hits needed in the multi-hit process of leukemia formation and therefore do not necessarily reflect leukemogenesis incidences as those that are observed in unbiased mouse transplantation experiments. Specifics of culturing or gene deletion or -addition mouse models might even skew the incidences observed because the genes that need to be altered are different for each model. Although the different *in vitro* and accelerated *in vivo* models are useful for vector or vector architecture comparisons, they therefore are less indicative of the oncogenic risk in the patient.

The experiments described in chapter 3 were aimed at identifying the insertion sites of pLZRS gamma-retroviral vectors in a low copy transduction situation. When this study was initiated, the integration behavior of HIV derived lentiviruses was described (19) and integration profiles of ASLV and RV were described in cell lines (20). We set out to determine integration profiles of gamma-retroviral vectors in percoll-purified mouse bone marrow cells, from which we know that a only a small fraction of the cells is able to repopulate the mouse and give rise to subsequent life-long hematopoiesis. We considered that this mouse model would provide valuable information for clinical gene therapy of diseases affecting the blood or diseases where bone marrow transplantation might be a viable treatment option. By analyzing gene expression in very immature LSK cells, the cell subpopulation of bone marrow enriched for repopulating cells, and viral vector integrations obtained after retransplantation, we confirmed in this relevant cell type, the earlier finding (20) that gamma-retroviral insertions occur in genes which are expressed in the target cells. We then found a relation between the target cell and the functions of the genes that were targeted, linking the gene expression and insertion behavior more closely to the target cell.

For a large part, insertions seem to be guided by gene expression. In addition by analyzing the common insertion sites, we discovered that these do not exclusively end up in highly expressed genes, but are spread over the different expression values. This pointed to other parameters than only gene expression being involved. One of these parameters might very well be the effect of the neighboring gene on the growth characteristics of the target cells, as was demonstrated by the integrations in *Evi1* (16).

Since a viral vector insertion might influence expression of genes other than the closest gene (21,22) we also analyzed the effect of expression of the surrounding locus on integration and the functions of the genes in these loci. Both analyses confirmed

our previous findings with regards to expression and gene function. Gamma-retroviral insertion occurred statistically significantly more often in highly expressed genes and the insertion profile was related to the hematopoietic cells used as targets Functional analysis of vector insertions should therefore consider genes surrounding (+/-100 kbp) an insertion rather than just the closest ones.

In pre-clinical gene therapy models such as a rhesus monkey transplantation model (5) as well as clinical gene therapy using gamma-retroviral vectors, target cells need to be stimulated with cytokines before they can be efficiently transduced (23,24). This is due to the fact that gamma-retroviruses can only integrate in dividing cells. The cytokine stimulation drives cells into the cell cycle, thereby allowing the integration of the gamma-retrovirus. When the cytokine stimulation would alter vector insertion profiles, choosing a suitable cytokine cocktail might help steer insertions to selected genes. We investigated the effect of cytokines on insertion patterns and found no correlation in mice (chapter 3) or in human hematopoietic cells (Appendix, 25), therefore making it unlikely that the cytokine stimulation affects insertion behavior in LSK or human $CD34^+$ cells.

Since we considered gamma-retrovirus insertions in mice a suitable model for human insertions, we compared the insertion sites retrieved in mice with those retrieved in humans, using data from the London 25 or Paris XSCID 26 trials (see appendix). Using the appropriate human homologues to the retrieved mouse genes, we could identity that ~10% of the genes retrieved in mice were also found in clinical trials, were human $CD34^+$ cells were transduced with gamma-retrovirus. This leads us to the assumption that the described mouse model with long-term observation is suitable and predictive for insertion analysis of retroviral vectors.

In chapter 4, we focused on the analysis of virus integration sites in cells obtained from patients that underwent gene therapy for XSCID (25,26), ADA SCID (27), XCGD (8) and also from mouse and rhesus macaque pre-clinical experiments. The amount of sequences retrieved in each of these studies is limited due to the availability of DNA from these patients and the sequencing techniques used. The opportunity to combine the data and analyze them together should provide better insights in the behavior of gamma-retroviral insertions in hematopoietic stem cells, being of human, mouse or rhesus macaque origin. To have good comparability between the different datasets, the integrations in human cells were re-analyzed using the same genome build and alignment parameter (the effect of which is discussed in chapter 6). By doing so, we showed that 75% of insertions occurred within 10 kbp of RefSeq genes. The larger combined dataset also allow more thorough analysis of common insertion sites (CIS, defined by Suzuki's criteria 28). These CIS showed non-randomness as 40% of the insertion sites are in common genomic regions. Higher order CIS are near genes known to promote cellular transformation, as they are described in the RTCGD and CGC databases. When

comparing the pre-transplant and post-transplant samples, it is clear that selection takes place. This might in part be due to a limited number of cell capable of reconstitution being transplanted, but it might also be due to the effect of the insertions on the selected cells. The second option seems more likely, since the number of CIS increases over time.

The clinical trials analyzed here employed different transgenes to cure different monogeneic diseases (XCGD, XSCID, ADA SCID) and only shared the vector architecture (gamma-retroviral vectors) and the targeted cell subset ($CD34^+$ cells). Despite of the differences between the trials, high order CIS were not unique to each individual trial, but rather show the *in vivo* selection after transduction of $CD34^+$ cells with gamma-retroviral primers. The five most often retrieved *CIS*, *LMO2*, *MDS1/EVI1*, *PRDM16*, *SETBP1* and *CCND2*, were those that were involved in the onset of leukemia as a result of gamma-retroviral gene transfer. Analysis of the CIS in pre- and post-transplant data showed a strong selection effect *in vivo*.

When looking into the gene functions of the targeted genes in the trials, we observed, similar to the mouse experiments (described in chapter 3), that the insertions occurred in genes which have a physiological role in the target cells, already in the pre-transplant samples, arguing that pronounced selection of insertion sites is determined by the target cell type. Again, similar to the mouse study data, the human data showed that gamma-retroviral vectors tend to insert near expressed genes, but CIS do not uniquely occur near the highest expressed genes. This observation confirms that also in humans, CIS seem to undergo selection, perhaps due to the fact that insertions near CIS alter fitness of the target cell for long-term repopulation. Since the rhesus monkey genome is not annotated as well as the mouse or human genome, no clear direct results could be obtained from the rhesus macaque study. The analysis of the combined human clinical trials shows that gamma-retroviral insertions in hematopoietic cells show a similar behavior to those in mouse and rhesus macaque. The CIS retrieved in the trials and preclinical data confirm that in different species, CIS are more likely to be repopulate the recipient, and in addition, might lead to adverse events when these cells get deregulated.

In chapter 5, we looked into the survival of mice that received Percoll enriched bone marrow cells transduced with a gamma-retroviral vector carrying either a marker gene (eGFP) or a signaling molecule (wtStat5b) and a marker gene (eGFP) driven from an internal ribosomal entry site in a bi-cistronic setup. This mouse studies was originally designed to investigate the effect of the expression of wtStat5b in mouse bone marrow, but was extended to also evaluate the incidence insertional oncogenesis, shown by the occurrence of leukemia in the different experiment group. Expression of wtStat5b did not result in higher incidences of leukemia or the occurrence of different leukemia phenotypes. This allowed us to combine the eGFP and wtStat5b groups and analyze the

impact of gamma-retroviral insertion on the combined dataset. We performed transplantations with in total 16 x10^6 transgene positive cells and observed 3 leukemias in 187 transplanted mice. After secondary transplantation of pooled bone marrow of the primary recipients, another 21 leukemias in 126 mice were found. This is a 2.7 fold increase in leukemias over the 11 leukemias that were observed in 92 mice in the control groups. The need for long- term observations is clear, since the current study shows a median survival time of 378 days. This is longer than observed in the *Cdnk2a*$^{-/-}$ model (17), which had a 225 day median survival. The longer survival time arguably makes the current model more sensitive for less genotoxic vector backbones, since in the *Cdnk2a*$^{-/-}$ mouse, all mice start dying from background effects rather than the virus transduction (18).

Based on the number of transplanted transduced cells, we calculated an incidence of leukemia occurrence of 1.5x10^{-6} per transplanted cell, which was higher than the previous estimate of 2.2x10^{-7} based on immortalization experiments performed in TF-1 cells (29). It remains to be determined what the cause of these differences is, but transplantation of mice with a limited number of hematopoietic stem cells might cause replicative stress (30,31) during the *in vivo* expansion, which might influence the leukemia incidence. Looking at possible relations between known proto-oncogenes (RTCGD genes. 32) or dominant clones and survival time did not show that either of these had a negative influence on the survival of the mice.

During the experiments involving insertion site determination described in this thesis, we quickly realized that manual annotation of insertion sites is labor intensive and error prone. We therefore set out to automate parts of the annotation process. Using the Ensembl annotations, we were able to perform insertion annotation much faster than can be done by hand and removed user bias. Using the Ensembl API and annotations also allowed us to set strict parameters to make sure that every single insertion was analyzed as the other insertions in the dataset. The analysis parameters used in chapter 4 were closely inspected, and as a result, we decided to redo the analysis of all insertions described in chapter 4 with one set of analysis parameters rather than relying on previously published annotations (25-27). A closer look into the ADA SCID dataset (27) shows which impact analysis parameters might have on the annotated genes. While the amount of sequences in the clinical datasets were small enough to be aligned by hand, new insertion site studies based on pyrosequencing will provide at least one order of magnitude more sequence data. At this scale one can only rely on fast automated annotation, which was one of the reasons for developing MAVRIC. When high throughput (deep sequencing) approaches are to be established to monitor development of gene marked hematopoiesis in gene therapy patients, a rapid and reliable method for integration annotation is necessary.

The CIS in these studies seem to provide an advantage for integration, reconstitution or both, since 43% of the integrations occur in CIS among the different datasets and

the CIS are present in both early and late samples, and persist over a 4 year period, even when the clonal repertoire overall diminishes. When these integrations indeed do transfer this advantage, we can speculate that these integrations are important for successful gene therapy. This is even more clear from the high-throughput sequencing experiments that were performed in the London XSCID trial. Here, 6 CIS near oncogenes were identified at 3 different time points.

Analysis of pre-transplantation CIS compared to post-transplantation CIS showed that clonal selection is a process that apparently occurred shortly after the integrations occurred at these sites. However, there is a bias from the CGD studies, since the 7th order CIS were those in *Evi1* and *Prmd16* identified in that study.

Large scale cross study comparisons clearly show that integration analysis is rather sensitive to differences in genome builds and annotations used. It is clear that these should be harmonized in all studies to allow meaningful comparisons. This necessitates the development of automated tools, such as GTSG.org (33), SeqMap (34), MHH ISA[28] and others, which allow automated reanalysis to be performed within a reasonable time period.

The frequent occurrence of CIS in different human and preclinical samples near proto-oncogenes and with leukemia observed in 5 patients might lead to the suggestion that these virus integrations are frequently leukemogenic. However, leukemia development in rhesus monkeys and humans after transplantation with gamma-retrovirally transduced cells is still an uncommon event, even though we estimated frequencies for leukemia formation in the mouse to be between 1 in 10^6 to 10^7 transduced cells. The long latency time observed and the necessity for secondary transplantation in the mouse study point to the necessity of secondary mutations for leukemia development after gene therapy. The nature of these secondary events is still poorly understood and should be investigated further.

In conclusion, the pre-clinical data and the human studies show that the classical retroviral vector design probably does not have a favorable integration pattern from a integration safety standpoint and that potential improvements can be achieved in this area by using different vector backbones, self-inactivating designs or by using weaker internal promoter/enhancers. On the other hand, the CIS that were observed might actually help transduced cells to repopulate and survive better, which aids transplantation and cure of the patient.

A second, obvious point to increase retroviral gene therapy is the reduction of the number of transduced cells that is transplanted. Although the reduction of transplant size is obvious from a mathematical point of view, it might not be welcomed in the clinic, since it will inevitably lead to longer hospitalization and increased chances of infection

28 http://eh.mh-hannover.de/isa

during the time that the recipient is leukopenic. Here, a possible solution perhaps lies in the use of more purified target cells. Instead of $CD34^+$ cells, $CD34^+CD38^-$ cells might be used. These cells contain a higher fraction of the repopulating HSC and should therefore show similar transplantation kinetics. Two recent reports (9,10), however, described that insertional oncogenesis does not stem from the later progenitors. Using further purified HSC as targets for gamma-retroviral gene therapy is therefore unlikely to bring improvements in safety.

The adverse events in the clinical trials for XSCID and XCGD lead to increased interest in the analysis of clonality of the transplant. In mouse and rhesus monkey studies were retroviral vectors where used for transduction, oligoclonal repopulation patterns tended to arise, likely due to the fact that only a limited number of hematopoietic stem cells were transduced. In monkeys, the repopulation patterns of these clones was studied, by challenging the monkeys with chemotherapeutics to established how many dormant clones were present in the bone marrow. These results point to the fact that only a limited number of transduced clones give rise to long-term hematopoiesis. In the protocols used for clinical gene therapy, a host of clones carrying different integrations is usually transplanted, which makes it difficult to establish whether any malignant or pre-malignant clones are infused into patient. There is a dichotomy here: clinicians prefer to transplant larger numbers of cells, since this reduces the time to reconstitution. This has obvious clinical benefit, but when the number of transplanted clones is limited, to numbers below the estimated frequency of malignant clones that is expected with a certain virus backbone (for example, as determined in chapter 3) the chance of infusing a malignant clone is reduced at the expense of a longer time to reconstitution. Later reconstitution puts the patient at risk to get infected during the leukopenia. With self-inactivating vector backbones showing reduced frequencies of malignant transformation (18, 35) the reduction in cell numbers might not be very dramatic and allows protocols to reach the situation where the chance of infusing a malignant clone is limited. On the other hand, since the patients in the XSCID trials received no cytoreductive therapy at all (6,11), a reduction of the number of transplanted cells might be feasible down to a level were only very few cells finally repopulate the patient.

As shown in chapter 4, pre-transplant samples do provide insight into the integration pattern of a specific vector backbone in human cells, but fail to capture the post-transplant selection. It is therefore questionable whether firm assumptions on vector safety can be made on pre-transplant integration analysis. Furthermore, the occurrence of dominant clones and eventually leukemia is hard to predict from clinical samples or pre-clinical samples. In the case of LMO2 insertions, this is very clear, since these insertions lead to leukemias in five XSCID patients, but similar insertion in the ADA SCID trail until now did not result in such adverse events. In the London XSCID trial, an insertion in *LMO2* and a subsequent loss of the p-arm of chromosome 7 (4) was

believed to lead to the leukemia, showing that secondary events are necessary to lead to insertional oncogenesis.

In this respect, it is interesting to see that both FDA and EMEA currently do not prescribe which vectors can or cannot be used for gene therapy. It is perfectly feasible to start a clinical trial with a gamma-retroviral vector as were used in the XSCID gene therapy trials. The main concern, however, is the follow up of insertions in the patients. This follow-up is not a clear cut analysis, as we have shown in chapter 6. Small changes in the methods employed to study insertions can lead to different results. The question therefore rises, whether specialized laboratories should take over the task of analyzing patient data.

As described above, in clinical gene therapy protocols a large number of different clones (0.9-20x10^6 transduced cells/kg, mean 7x10^6/kg, (6-8, 11) is transplanted into patients. This is mainly because no HSC expansion protocols were available. Recently, cell culture protocols have been established for mouse (36) and human UCB $CD34^+$ cells (37), which allow expansion of stem cells up to 30x in 10 days. Such an expansion culture might allow selection of a limited number of possibly benign integrations, and expand them to a size suitable for reconstitution. A question untouched at this moment would be whether the expanding cells, in a clearly non-physiological setting might lead to accelerated accumulation of cell damage due to replicative stress or high oxygen levels in the cell culture, so that the effects of selection might be negated by the culture stress. Another avenue that is worth exploring is the gene correction of induced pluripotent stem cells (iPS) generated from patient tissue. Although the iPS field is currently quickly moving and proof-of-concept of gene addition therapy has been shown in mice (38), protocols to safely establish iPS clones, repair them or express additional copies of corrected genes and differentiate the clones into hematopoietic cells have not yet been developed. Again, since the iPS procedure allows growth of reprogrammed clones, a selection of benign clones should be feasible, but the same drawbacks of replicative stress need to be investigated.

With upcoming gene therapy trials for beta-thalassemia, Wiskott-Aldrich syndrome (39), Fanconi Anemia (40), metachromatic leukodystrophy (41) and a range of other storage diseases currently in translational research, gene therapy of monogeneic diseases promises to become a more routine treatment. Even though long-term follow data is still limited, one might say that for XSCID and ADA SCID, gene therapy provides a viable alternative to bone marrow transplantation. It is currently the treatment of choice, especially when no matched related donors are available. Since the conditioning in autologous gene therapy is usually milder compared to matched unrelated donor bone marrow transplant (no conditioning in XSCID and mild busulfan treatment in ADA SCID), gene therapy, if proven safe, will rapidly replace allogeneic stem cells trans-

plantation in these cases, with the expected added benefit of abolishing late effects of conditioning in these usually young patients.

REFERENCES

1. Li, Z. et al. Murine leukemia induced by retroviral gene marking. *Science* **296**, 497 (2002).
2. Hacein-Bey-Abina, S. et al. LMO2-associated clonal T cell proliferation in two patients after gene therapy for SCID-X1. *Science* **302**, 415-9 (2003).
3. Hacein-Bey-Abina, S. et al. Insertional oncogenesis in 4 patients after retrovirus-mediated gene therapy of SCID-X1. *J. Clin. Invest* **118**, 3132-3142 (2008).
4. Howe, S.J. et al. Insertional mutagenesis combined with acquired somatic mutations causes leukemogenesis following gene therapy of SCID-X1 patients. *J Clin Invest.* **118**, 3143–3150 (2008).
5. Seggewiss, R. et al. Acute myeloid leukemia is associated with retroviral gene transfer to hematopoietic progenitor cells in a rhesus macaque. *Blood* **107**, 3865-7 (2006).
6. Hacein-Bey-Abina, S. et al. Sustained correction of X-linked severe combined immunodeficiency by *ex vivo* gene therapy. *N Engl J Med* **346**, 1185-93 (2002).
7. Aiuti, A. et al. Correction of ADA-SCID by stem cell gene therapy combined with nonmyeloablative conditioning. *Science* **296**, 2410-3 (2002).
8. Ott, M.G. et al. Correction of X-linked chronic granulomatous disease by gene therapy, augmented by insertional activation of MDS1-EVI1, PRDM16 or SETBP1. *Nat Med* **12**, 401-9 (2006).
9. Kustikova, O.S. et al. Cell-intrinsic and Vector-related Properties Cooperate to Determine the Incidence and Consequences of Insertional Mutagenesis. *Mol. Ther* (2009).doi:10.1038/mt.2009.134
10. Newrzela, S. et al. Resistance of mature T cells to oncogene transformation. *Blood* **112**, 2278-2286 (2008).
11. Gaspar, H.B. et al. Gene therapy of X-linked severe combined immunodeficiency by use of a pseudotyped gammaretroviral vector. *Lancet* **364**, 2181-7 (2004).
12. Baum, C.M., Weissman, I.L., Tsukamoto, A.S., Buckle, A.M. & Peault, B. Isolation of a candidate human hematopoietic stem-cell population. *Proc. Natl. Acad. Sci. U.S.A* **89**, 2804-2808 (1992).
13. Modlich, U. et al. Leukemias following retroviral transfer of multidrug resistance 1 (MDR1) are driven by combinatorial insertional mutagenesis. *Blood* **105**, 4235-46 (2005).
14. Modlich, U. et al. Cell-culture assays reveal the importance of retroviral vector design for insertional genotoxicity. *Blood* **108**, 2545-53 (2006).
15. Bokhoven, M. et al. Insertional gene activation by lentiviral and gammaretroviral vectors. *J. Virol* **83**, 283-294 (2009).
16. Modlich, U. et al. Insertional Transformation of Hematopoietic Cells by Self-inactivating Lentiviral and Gammaretroviral Vectors. *Mol. Ther* (2009).doi:10.1038/mt.2009.179
17. Montini, E. et al. Hematopoietic stem cell gene transfer in a tumor-prone mouse model uncovers low genotoxicity of lentiviral vector integration. *Nat Biotechnol* **24**, 687-96 (2006).
18. Montini, E. et al. The genotoxic potential of retroviral vectors is strongly modulated by vector design and integration site selection in a mouse model of HSC gene therapy. *J. Clin. Invest* **119**, 964-975 (2009).
19. Schroder, A.R. et al. HIV-1 integration in the human genome favors active genes and local hotspots. *Cell* **110**, 521-9 (2002).
20. Mitchell, R.S. et al. Retroviral DNA integration: ASLV, HIV, and MLV show distinct target site preferences. *PLoS Biol* **2**, E234 (2004).
21. Bartholomew, C. & Ihle, J.N. Retroviral insertions 90 kilobases proximal to the Evi-1 myeloid transforming gene activate transcription from the normal promoter. *Mol Cell Biol* **11**, 1820-8 (1991).
22. Kustikova, O. et al. Clonal dominance of hematopoietic stem cells triggered by retroviral gene marking. *Science* **308**, 1171-4 (2005).

23. van Hennik, P.B. et al. Highly efficient transduction of the green fluorescent protein gene in human umbilical cord blood stem cells capable of cobblestone formation in long-term cultures and multilineage engraftment of immunodeficient mice. *Blood* **92**, 4013-22 (1998).
24. Wognum, A.W., Visser, T.P., Peters, K., Bierhuizen, M.F. & Wagemaker, G. Stimulation of mouse bone marrow cells with kit ligand, FLT3 ligand, and thrombopoietin leads to efficient retrovirus-mediated gene transfer to stem cells, whereas interleukin 3 and interleukin 11 reduce transduction of short- and long-term repopulating cells. *Hum Gene Ther* **11**, 2129-41 (2000).
25. Schwarzwaelder, K. et al. Gammaretrovirus-mediated correction of SCID-X1 is associated with skewed vector integration site distribution *in vivo*. *J Clin Invest* **117**, 2241-9 (2007).
26. Deichmann, A. et al. Vector integration is nonrandom and clustered and influences the fate of lymphopoiesis in SCID-X1 gene therapy. *J Clin Invest* **117**, 2225-32 (2007).
27. Aiuti, A. et al. Multilineage hematopoietic reconstitution without clonal selection in ADA-SCID patients treated with stem cell gene therapy. *J Clin Invest* **117**, 2233-40 (2007).
28. Suzuki, T. et al. New genes involved in cancer identified by retroviral tagging. *Nat Genet* **32**, 166-74 (2002).
29. Stocking, C. et al. Distinct classes of factor-independent mutants can be isolated after retroviral mutagenesis of a human myeloid stem cell line. *Growth Factors* **8**, 197-209 (1993).
30. Holyoake, T.L. et al. *In vivo* expansion of the endogenous B-cell compartment stimulated by radiation and serial bone marrow transplantation induces B-cell leukaemia in mice. *Br. J. Haematol* **114**, 49-56 (2001).
31. Baum, C., Kustikova, O., Modlich, U., Li, Z. & Fehse, B. Mutagenesis and oncogenesis by chromosomal insertion of gene transfer vectors. *Hum Gene Ther* **17**, 253-63 (2006).
32. Akagi, K., Suzuki, T., Stephens, R.M., Jenkins, N.A. & Copeland, N.G. RTCGD: retroviral tagged cancer gene database. *Nucleic Acids Res* **32**, D523-7 (2004).
33. Appelt, J. et al. QuickMap: a public tool for large-scale gene therapy vector insertion site mapping and analysis. *Gene Ther* **16**, 885-893 (2009).
34. Peters, B. et al. Automated analysis of viral integration sites in gene therapy research using the SeqMap web resource. *Gene Ther* **15**, 1294-1298 (2008).
35. Zychlinski, D. et al. Physiological promoters reduce the genotoxic risk of integrating gene vectors. *Mol Ther* **16**, 718-25 (2008).
36. Zhang, C.C. et al. Angiopoietin-like proteins stimulate *ex vivo* expansion of hematopoietic stem cells. *Nat Med* **12**, 240-5 (2006).
37. Zhang, C.C., Kaba, M., Iizuka, S., Huynh, H. & Lodish, H.F. Angiopoietin-like 5 and IGFBP2 stimulate *ex vivo* expansion of human cord blood hematopoietic stem cells as assayed by NOD/SCID transplantation. *Blood* **111**, 3415-3423 (2008).
38. Hanna, J. et al. Treatment of sickle cell anemia mouse model with iPS cells generated from autologous skin. *Science* **318**, 1920-1923 (2007).
39. Marangoni, F. et al. Evidence for long-term efficacy and safety of gene therapy for Wiskott-Aldrich syndrome in preclinical models. *Mol. Ther* **17**, 1073-1082 (2009).
40. Jacome, A. et al. Lentiviral-mediated genetic correction of hematopoietic and mesenchymal progenitor cells from Fanconi anemia patients. *Mol. Ther* **17**, 1083-1092 (2009).
41. Biffi, A. et al. Gene therapy of metachromatic leukodystrophy reverses neurological damage and deficits in mice. *J. Clin. Invest* **116**, 3070-3082 (2006).

CHAPTER 8

Summary

Retroviral gene therapy of monogeneic diseases has recently proven to be a successful method of treatment, often curing patients from severe and possibly lethal diseases such as X-linked severe combined immunodeficiency (XSCID), ADA SCID and chronic granulomatous disease (CGD). The cause of each of these diseases is a mutation in a single gene. The usual treatment for these diseases is bone marrow transplantation, however, not all patient have suitable HLA-matched related donors available. In such cases, bone marrow transplants using matched unrelated donors can be performed, but these often have severe side effects. Using retroviral vectors, derived from Moloney leukemia virus, an active, 'healthy' copy of such a mutated gene can be inserted at a random location in hematopoietic cells of the patient, usually $CD34^+$ cells. When these cells are then transplanted back into the patient and engraft, the patient can regenerate his hematopoietic system with the 'cured' cells. This approach has been tested for the diseases mentioned above in clinical setting and proved successful in 26 patients. After initial success, 5 patients in the XSCID trials developed leukemia and after very detailed examination of the locations were the retroviral vectors integrated, an effect of the virus was found to cause those leukemias. In four patients, the leukemia could be treated while their gene corrected hematopoiesis remained in place. In one patient, the leukemia could not be successfully treated, which lead to the death of the patient. Even when considering these severe adverse effects of gene therapy treatment, the overall results of these studies is positive and shows that monogeneic diseases of the hematopoietic system can be treated using gene therapy. It also clearly shows that the retroviral vectors used might not be optimal. This has caused the gene therapy field to investigate the viral vector backbone in detail, looking for readouts for genotoxicity and improved vector designs.

In chapter 3, we describe a study where we analyzed retroviral insertion sites retrieved from mice that received transduced bone marrow cells and were subsequently observed until retransplantation at different time points ranging from 117, 206, 245 and 342 days after transplantation. Secondary recipients were then observed for up to 645 days after transduction. When analyzing the insertions in the bone marrow and blood of these mice, we observed that there was a significant correlation between insertions and expression of the surrounding genes, measured by microarray. Interestingly, insertions sites that were more frequently observed did not show such strong correlations with gene expression, which indicates that the genes in these loci might have functions that make the cells that carry them repopulate more efficiently. We also found that the hematopoietic nature of the transduced cells was reflected in the repertoire of genes that were hit by the virus; the insertions occurs near genes with functions in the hematopoietic system, along with more general cell functions. Of note is also that ~6% of this limited number of genes had human homologues that were retrieved in the studies XSCID clinical trials described in chapter 4 (and appendix 1 and 2).

Chapter 4 extends our analysis in mice by investigating the insertions retrieved in the ADA-SCID, XSCID and CGD clinical trails, as well as a preclinical retroviral transduction study in rhesus monkeys. An occurrence of common insertion sites (CIS) both within and between studies was observed. These CIS were more frequently found after transplantation than in the pre-transplant samples, which leads us to believe that the CIS, were providing the cells carrying them with a survival/repopulation advantage. As in the mouse study, the insertions were significantly correlated with expression of the surrounding genes, but the CIS again did not show such close relationships.

To investigate how often insertions of replication deficient retroviral vectors would lead to leukemias, we analyzed the survival of 187 mice transplanted cells retrovirally transduced to carry either the marker gene *eGFP* or both *eGFP* and *wtStat5b*, a signaling molecule important in hematopoietic signal transduction. Unexpectedly, we found no difference in leukemia incidence or phenotype between these groups, showing that over-expression of a signaling molecule does not necessarily lead to malignant transformations. We observed 2.7x more leukemias in the mice transplanted with retrovirally transduced cells than in the control groups, with a median latency of 383 days. Since we controlled the number of cells initially transplanted, we could calculate an incidence of $1.9x10^{-7}$ leukemias/transgene positive cell (1 in 5.2 million cells), which increased to $1.5x10^{-6}$ (1 in 670,000 cells) in the secondary recipients. We analyzed the common insertion sites and retrieved the frequently found *Evi1* gene and insertion in genes associated with leukemia after retroviral infection (RTCGD). We were however unable to show that these genes or dominant clones in our mice lead to reduced latency.

In chapter 6, we describe the annotation tools we developed for analysis of retroviral insertion sites. During the experiments described in chapter 3 and 4, we noticed that the amount of available data for both the mouse and human genome increased almost every half year. We were therefore interested in re-analysis of the datasets, which is quite labor intensive. As a result, we developed a web tool that allows reasonably fast (re-)analysis of insertion sites. We reanalyzed a previously published dataset, containing insertion sites retrieved from ADA-SCID patients in a gene therapy trial, to assess the influence of each of the analysis parameters. We studied the effect of the use of removal of repeat sequences from the insertions and the differences between genome builds and annotations of these used. Differences in alignments between genome builds were usually small (~5%), but the annotations showed larger (~25%) differences, which also depended on the type of annotation that was used. For example, using RefSeq annotations proved to be less stable than EnsemblIDs. Since the influence of the analysis parameters proved to be considerable, we suggest that future studies of insertion sites should be very precise in the description of their methodology. Even small differences in analysis parameters might have implications for the conclusions obtained from such studies.

The research described in this thesis shows that although retroviral vectors can successfully be used for gene therapy, their integration profiles are not inherently safe. New viral vectors, with less efficient enhancers or carrying insulators might reduce the effect of the vector insertion on its surroundings. Similarly, a reduction of the number of transduced cells given in the transplant, might further reduce the frequency of adverse events. Mouse studies showed that long term observation resulted in increased leukemia incidence, which is important when mouse models are used for preclinical evaluation of new gene therapy vectors. Finally, the common insertions retrieved in mouse, rhesus macaque and human clinical trials point to effects of the genes in these common insertion loci that might increase cell survival or make cells more efficiently engraft the recipients and might therefore be interesting targets to investigate in hematopoietic reconstitution.

Nederlandstalige samenvatting

De retrovirale gentherapie van monogenetische ziekten is in de laatste jaren een effectieve behandelingsmethode gebleken. Vaak werden met behulp van deze therapie patiënten behandeld die anders geen, slechtere of duurdere behandelingen hadden moeten ondergaan. De ziekten die zijn behandeld zijn ernstige, soms levensbedreigende aandoeningen van het immuunsysteem, zoals XSCID (X-linked severe combined immunodeficiency), ADA SCID (adenosine deaminase deficiency) en CGD (chronic granulomatous disease). In al deze ziektes is een mutatie in één gen de oorzaak. Normaliter kunnen deze ziekten met behulp van beenmerg transplantatie worden behandeld, maar het is niet altijd mogelijk een geschikte, HLA compatibele familie donor te vinden. Bij afwezigheid wordt dan naar een HLA compatibele niet-familie donor gezocht. Het nadeel van een transplantatie van niet-familie donor beenmerg is een grotere morbiditeit.

Wanneer men retro-virale vectoren gebruikt, die bestaan uit een op het Moloney leukemie virus (MLV) gebaseerde virussen, dan is het mogelijk de $CD34^+$ beenmerg cellen van een patiënt met een monogenetische aandoening te voorzien van een additionele werkzame kopie van het defecte gen. Na transplantatie van deze cellen in de patiënt kunnen zij een nieuw hematopoietisch systeem vormen, waarbij het ingevoegde gen in alle dochter cellen van het originele transplantaat aanwezig is. Deze methode is in de afgelopen jaren in 26 patiënten getest en was tot nu toe voornamelijk succesvol. Echter, in 5 patiënten werd na initieel succes van de behandeling in de XSCID trials een leukemie gevonden. Na uitvoerig onderzoek van de leukemie cellen kon worden vastgesteld dat deze een insertie van de retrovirale vector zeer waarschijnlijk de expressie van de omliggende genen gedereguleerd had, wat aanleiding gaf tot ontwikkeling van de leukemie. In vier patiënten kon deze leukemie succesvol worden behandeld, waarna de patiënten met een normaal immuunsysteem inclusief additioneel gen genezen konden worden verklaard. In een geval kon de leukemie niet worden behandeld, wat tot het overlijden van de patiënt leidde. Zelf wanneer men de serieuze gevolgen van een virus integratie in een ongewenst locus en de daaruit volgende leukemie beschouwd, kan de behandeling van monogenetische ziekten van het hematopoietisch systeem met behulp van retrovirale vectoren over het algemeen als succesvol en effectief worden gezien. Uit deze klinische trials wordt ook duidelijk, dat de MLV gebaseerde retrovirale vector niet optimaal is. Het gentherapie vakgebied heeft zich als gevolg van de negatieve resultaten uit de trials geconcentreerd op de ontwikkeling van veiligere virale architectuur, tests voor genotoxiciteit en verbetering van de componenten van de vectoren.

In hoofdstuk 3 beschrijven we een studie waarin we retrovirale inserties die geïdentificeerd zijn in muizen die een transplantatie met getransduceerd beenmerg hebben ondergaan en vervolgens voor lange tijd gevolgd zijn, tot het moment dat beenmerg van deze muizen op 117, 206, 245 en 342 dagen na transplantatie in secundaire ontvanger muizen werd getransplanteerd. Deze secundaire ontvangers werden daarop tot maximaal 645 dagen geobserveerd. Bij de analyse van de virale inserties in het beenmerg

van deze muizen viel het op dat genen die in de doel cellen tot expressie kwamen (wat wij met microarray analyse aangetoond hebben) een grotere kans hadden door een retrovirus getroffen te worden. Dit fenomeen was al bekend voor de van HIV afgeleide lentivirussen en retrovirussen, maar de correlatie die wij konden aantonen was significant. Niet alleen de naastliggende genen leken van belang, maar ook het gehele locus. Opmerkelijk was dat een aantal veelvoorkomende doelwitten voor integratie niet bijzonder hoog tot expressie kwamen in de doel cellen. Hieruit maakten wij op dat deze genen wellicht een belangrijke taak in de hematopoiese hebben, waardoor de cellen met integraties in de buurt van deze genen beter in staat zijn de muis te repopuleren. Ook stelde we vast dat de hematopoietische aard van de doelcellen in het integratie patroon van de retrovirussen te herkennen was: een veelvoud van inserties bevonden zich in genen die een functie in de hematopoiese hadden, hoewel genen met meer algemene functies ook geraakt waren. In onze weliswaar beperkte integratie dataset vonden we een ~6% overlap met een integraties die in de patiënten in de XSCID studie gevonden waren (Hoofdstuk 4 en appendix 1 en 2).

In hoofdstuk 4 wordt de analyse van retrovirale inserties in hematopoietische cellen uitgebreid met inserties uit de XSCID, ADA SCID en CGD studies in mensen en een preklinische studie in resusapen. Hier werden veelvuldig voorkomende inserties (common insertion sites, CIS) niet alleen binnen de verschillende datasets gevonden, maar ook tussen de verschillende datasets. Ook werden deze CIS vaker gevonden na transplantatie in vergelijking tot voor transplantatie, waaruit op te maken valt dat cellen met deze CIS een voordeel in repopulatie of overleving hadden. Net als in de muizen studie beschreven in hoofdstuk 3, is een duidelijke correlatie tussen insertie en gen expressie zichtbaar, waarbij deze bij CIS minder duidelijk is.

De incidentie van leukemie als gevolg van retrovirale gen markering is geëvalueerd in 187 muizen, die zijn getransplanteerd met retroviral getransduceerde cellen, die de marker EGFP of het signaal transductie molecuul wtStat5 in combinatie met EGFP tot expressie brachten. De incidentie van leukemie was niet verschillend tussen deze groepen, waaruit we concluderen dat overexpressie van het signaal transductie molecuul wtStat5 niet noodzakelijkerwijs tot maligne transformatie leidt. In de muizen die getransduceerde cellen ontvangen werden 2.7x meer gevallen van leukemie gevonden dan in de controle groep, met een mediane latentietijd van 383 dagen. Omdat een bekend aantal getransduceerde cellen werd getransplanteerd, was het mogelijk te berekenen dat $1.9x10^{-7}$ getransduceerde cellen (1 per 5.2 miljoen cellen) in de primaire transplantatie tot een leukemie leidde. In de secundaire transplantatie was een toename zichtbaar in de leukemie-frequentie tot $1.5x10^{-6}$ per getransduceerde cel (1 per 670,000 cellen). In de analyse van de retrovirale integratie sites werd een frequente insertie in *Evi1* en in leukemie-geassocieerde genen (zoals samengevat in de RTCGD) gevonden. Inserties in

of nabij deze genen of andere dominante klonen in de polyclonale hematopoietische populatie leidde echter niet tot een versnelde progressie tot leukemie.

In hoofdstuk 6 introduceren we software voor annotatie van virus integratie sites. Bij de uitvoer van de experimenten beschreven in hoofdstuk 3 en 4 werd het duidelijk dat de beschikbaarheid van een programma voor automatische annotatie van insertie sites zowel de nauwkeurigheid van deze annotaties als de snelheid waarmee deze gegenereerd worden zou kunnen verbeteren. Om de functionaliteit van deze software te demonstreren, werd een eerder gepubliceerde dataset opnieuw geanalyseerd om de invloed de diverse parameters op de analyse te evalueren. Wij bekeken de effecten van het gebruik van verschillende genome builds, het verwijderen van repeats uit de sequenties die de insertie sites bepalen en verschillende manier annotatie van de geïdentificeerde integratie sites. Hieruit bleek dat er slecht kleine (5%) verschillen in alignments tussen de verschillende genome builds zijn, maar dat het gebruikt van verschillende annotaties een groter (~25%) verschil veroorzaakte. Omdat deze parameters een duidelijke invloed op de resultaten hadden, concluderen wij dat de analyse parameters duidelijk gespecificeerd dienen te worden in de beschrijving van zulke analyses. Het onderzoek beschreven in dit proefschrift maakt duidelijk hoewel het klinisch succes van het gebruik van retrovirale vectoren bewezen kan worden geacht, deze niet als geheel veilig kunnen worden beschouwd. Nieuwe ontwikkelingen in vector architectuur, waarbij van minder potente enhancer sequenties of van insulator sequenties gebruik wordt gemaakt, zouden het effect van de vector op het omliggende genoom kunnen reduceren en zo tot minder genotoxische vectoren kunnen leiden. Omdat het optreden van leukemie ook met het aantal getransplanteerde getransduceerde cellen samenhangt, zou een reductie van het aantal cellen in het transplantaat ook tot een geringere kans op leukemie kunnen leiden. In ons muizen-model was een duidelijke toename van het aantal leukemie gevallen zichtbaar, wat consequenties heeft voor de tijdsduur van zulke experimenten als ze als voor evaluatie van de veiligheid van een vector worden ingezet. De common insertion sites (CIS) die werden gevonden in het muizen model, de resus makaak en de patiënten-monsters duiden erop dat er een effect bestaat van genen die deze loci omringen op de overleving of de efficiëntie waarmee deze cellen het hematopoietisch systeem de ontvanger repopuleren na transplantatie. Het is daarom van belang de functie van deze genen in hematopoiese nader te onderzoeken.

Acknowledgments/Dankwoord

...we are like dwarfs on the shoulders of giants, so that we can see more than they, and things at a greater distance, not by virtue of any sharpness of sight on our part, or any physical distinction, but because we are carried high and raised up by their giant size.– Bernard of Chartres 12th century

Nieuw onderzoek ontstaat altijd als gevolg van vragen die door eerder onderzoek zijn opgeworpen. Elke onderzoeker is daarom schatplichtig aan zijn voorgangers, maar in het dagelijks lab-leven zijn collega's en vrienden van onschatbare waarde. De schouders van reuzen zijn noodzakelijk, maar schouder aan schouder staan met reuzen is wat je in je onderzoek vooruit helpt. Een aantal van deze giganten wil ik speciaal bedanken.

Om te beginnen een groot dankjewel aan alle medewerker van de groep Wagemaker. Als eerste aan Gerard Wagemaker, die me de mogelijkheid heeft gegeven zes jaar aan zeer interessant onderzoek te werken en diverse congressen te bezoeken. Van hem heb ik veel geleerd, variërend van hematologie, onderzoek, filosofie, volkenkunde, presentatie technieken en natuurlijk de experimenten uit de jaren 80, die soms, maar soms ook niet, beter zijn dan wat er tegenwoordig uitgevoerd wordt. Monique Verstegen, die me de eerste stapjes van mijn onderzoekscarrière heeft begeleid, heeft een bedankje ook meer dan verdient. Hopelijk is het prettig aan de andere kant van de tafel. Dank gaat ook uit naar Trui Visser die, altijd met gepeperde mening, haar inzicht en uitmuntende technische vaardigheden met me heeft gedeeld. Angeliki Zaniou, Kathelijn Peters, Shazia Arshad, Carla Oerlemans-Bergs, Erna Fränzel-Luiten, Geert Westerhuis, Merel Stok en de studenten bij de werkgroep Wagemaker, hebben allemaal een bijdrage aan dit proefschrift geleverd. Niek van Til heeft een deel van het onderzoek in dit proefschrift voortgezet. Dat leidde natuurlijk tot nieuwe vragen en vergelijkingen met de data in dit proefschrift, wat de nauwkeurigheid ervan heeft verbeterd. Niek, dank voor je kritische blik en goede ideeën. Steffen Dirkx heeft een groot deel aan de integratie analyses bijgedragen. Fatima Aerts, arts in een land van vol biologen, heeft me naast een groot aantal goede ideeën bijgebracht wat de samenwerking tussen artsen en biologen succesvol maakt. Leonie Kaptein verdient ook een woord van dank voor het net iets verder kijken in de wereld van gen therapie en waardevolle hulp en commentaren. With Marshall Huston I probably had the most in-depth discussions about data and programming, even though we got in over our heads on some occasions. This makes Marshall probably the persons with most insight into all the details of this thesis, and therefore a very suitable paranimf.

Maar mijn dank kan niet beperkt blijven tot de mensen in de werkgroep. Ik heb de hele afdeling Hematologie in Rotterdam als een zeer prettig werkomgeving ervaren. Iedereen daarom bedankt voor de bijdrage aan mijn werk of aan de algemene atmosfeer. Misschien worden secretaresses soms in dankwoorden vergeten. Dit is onterecht en

ik wil Monique Mes en Ans Mannens bedanken voor hun hulp bij het oplossen van onontkoombare administratie problemen.

My work on virus integration analysis started during an 2-week internship in the South of Germany in the lab of Manfred Schmidt and Christof von Kalle in Freiburg. I would like to thank the entire lab for the hospitality and for answering all my curious questions. I have enjoyed working with Manfred Schmidt and Annette Deichmann on my mouse and monkey samples first and later on the finer details of insertions in human, monkey and mouse datasets.

Met de werkgroep van Frank Staal bij de afdeling Immunologie hebben we in aantal jaar zeer interessante projecten opgezet. Floor Weerkamp en Karin Pike-Overzet, bedankt voor een hele prettige samenwerking en buitengewoon prettig gezelschap bij congres bezoeken. De inbreng van Miranda Baert, Peter Ng, Tom Schonewille, Edwin de Haas moet niet worden vergeten. Bij elkaar hebben zij een grote bijdrage aan de figuren in dit proefschrift geleverd. Zeer verhelderend en waardevol waren de invalshoeken van Dick de Ridder en prof.dr. Peter van der Spek op de integratie problematiek hadden. Roel Verhaak verdient een woord van dank omdat hij me de kracht van Perl voor het verwerken van grote datasets heeft laten zien, zonder hem zou ik ongetwijfeld nog steeds Excel gebruiken.

The collaborative projects in FP5 and 6 of the European Union gave me the opportunity to widen my view on gene therapy, mainly by visiting the Leukerbad meetings. Many thanks go to Dr. Steve Howe and colleagues with whom I spend a lot of bar-time discussing great projects, some of which made it into the lab and one of those actually made it into a paper I am very proud of.

Thanks are also due to all people I have and had the opportunity to work with at the department of Experimental Hematology at Hannover Medical School, a truly motivating environment to be in. Special thanks go to Axel Schambach, Ute Modlich, Tobias Maetzig, good friends with whom I already talked science before moving to Hannover, to my workgroup fellows Olga Kustikova, Maike Stahlhut, Stefan Bartels and technical man-at-arms Thomas Neumann and to prof. Christopher Baum for giving me the opportunity to work in his inspiring department.

Dank gaat ook uit naar prof.dr. Christopher Baum, prof.dr. Elaine Dzierzak en prof. dr. Peter van der Spek voor het lezen en becommentariëren van mijn proefschrift. Ik wil de overige leden van de promotiecommissie, prof.dr. Bob Löwenberg en prof.dr. Rob Hoeben danken voor de bereidheid met mij over het proefschrift van gedachte te willen wisselen.

Aan het einde van zo'n lange lijst namen staan dan zoals gebruikelijk familieleden en vrienden. Hoewel ik mijn onderzoek op een aantal conferenties heb toegelicht en bediscussieerd, is het het meest verhelderend in klare taal te zeggen wat je belangrijk vindt. Arnold, goede vriend en paranimf, hoewel onze vakgebieden mijlenver uit elkaar

liggen, is een babbel over de zaken die ons bezighouden altijd plezierig. Paul, Nicole, Muriel, Chaim, Opa en Oma van der Hidde, Theo en Lucie, Remco en Nora, Marleen en Otto, Marije, Pa en Ma, Shirley en Nina, ook jullie hebben allemaal, soms misschien zonder het te weten, een bijdrage aan dit werk geleverd.

Bedankt,

Hannover, mei 2010

Curriculum Vitae

CURRICULUM VITAE

Name:	Martijn H. Brugman, MSc.
Date of Birth	October 28, 1977
Place of Birth	Dordrecht, The Netherlands
Nationality	Dutch

Positions

2007-now	Ph.D. position at the Department of Experimental Hematology, Hannover Medical School, Hannover, Germany, led by Prof. Dr. C. Baum.

Education

2001- March 2007	Ph.D. Student at the Department of Hematology, ErasmusMC, Rotterdam, The Netherlands. Thesis:Insertional oncogenesis after retroviral gene transfer in hematopoietic stem cells. Supervisors: Dr. M.M.A. Verstegen PhD., Prof. Dr. G. Wagemaker PhD.
2001	Internship: Id proteins influence growth speed in angiogenesis. Department of Cellular Biochemistry, The Netherlands Cancer Institute (NKI), Amsterdam, The Netherlands. Supervisor: Dr. P. ten Dijke, PhD
2000	Graduation research project: The Effects of DCVC and Cisplatin on Kidney cells. Department of Toxicology, LACDR, Leiden University, Leiden, The Netherlands. Supervisor: Dr. B. van de Water, PhD.
1996-2001	MSc. in Biopharmaceutical Sciences, Leiden University, Leiden, The Netherlands.

Presentations

2004	7th Annual ASGT Meeting, Minneapolis, MN, USA.
2005	1st Annual CONSERT meeting, Leukerbad, Switzerland.
2005	13th Annual ESGT Congress, Prague, Tsjech Republic.
2006	2nd Annual CONSERT meeting, Leukerbad, Switzerland.
2006	9th Annual ASGT Meeting, Baltimore, USA.
2006	14th Annual ESGT Congress, Athens, Greece.
2007	3rd Annual CONSERT meeting, Leukerbad, Switzerland.

Awards

2004	Excellence in research award for students and fellows, ASGT.
2004	Travel Award, 7th Annual ASGT Meeting
2006	Travel Award, 9th Annual ASGT Meeting
2006	Travel Grant, 14th Annual ESGT Congress

Publications

PUBLICATIONS

1 Mechanisms controlling titer and expression of bidirectional lentiviral and gammaretroviral vectors.
Maetzig T, Galla M, Brugman MH, Loew R, Baum C, Schambach A.
Gene Ther. 2009 Oct 22.

2 Insertional transformation of hematopoietic cells by self-inactivating lentiviral and gammaretroviral vectors.
Modlich U, Navarro S, Zychlinski D, Maetzig T, Knoess S, Brugman MH, Schambach A, Charrier S, Galy A, Thrasher AJ, Bueren J, Baum C.
Mol Ther. 2009 Nov;17(11):1919-28. Epub 2009 Aug 11.

3 Cell-intrinsic and vector-related properties cooperate to determine the incidence and consequences of insertional mutagenesis.
Kustikova OS, Schiedlmeier B, Brugman MH, Stahlhut M, Bartels S, Li Z, Baum C.
Mol Ther. 2009 Sep;17(9):1537-47. Epub 2009 Jun 16.

4 Stem cell marking with promotor-deprived self-inactivating retroviral vectors does not lead to induced clonal imbalance.
Cornils K, Lange C, Schambach A, Brugman MH, Nowak R, Lioznov M, Baum C, Fehse B.
Mol Ther. 2009 Jan;17(1):131-43. Epub 2008 Nov 11.

5 Insertional mutagenesis combined with acquired somatic mutations causes leukemogenesis following gene therapy of SCID-X1 patients.
Howe SJ, Mansour MR, Schwarzwaelder K, Bartholomae C, Hubank M, Kempski H, Brugman MH, Pike-Overzet K, Chatters SJ, de Ridder D, Gilmour KC, Adams S, Thornhill SI, Parsley KL, Staal FJ, Gale RE, Linch DC, Bayford J, Brown L, Quaye M, Kinnon C, Ancliff P, Webb DK, Schmidt M, von Kalle C, Gaspar HB, Thrasher AJ.
J Clin Invest. 2008 Sep;118(9):3143-50.

6 Resistance of mature T cells to oncogene transformation.
Newrzela S, Cornils K, Li Z, Baum C, Brugman MH, Hartmann M, Meyer J, Hartmann S, Hansmann ML, Fehse B, von Laer D.
Blood. 2008 Sep 15;112(6):2278-86. Epub 2008 Jun 19.

7 Leukemia induction after a single retroviral vector insertion in Evi1 or Prdm16.
Modlich U, Schambach A, Brugman MH, Wicke DC, Knoess S, Li Z, Maetzig T, Rudolph C, Schlegelberger B, Baum C.
Leukemia. 2008 Aug;22(8):1519-28. Epub 2008 May 22.

8 Gammaretrovirus-mediated correction of SCID-X1 is associated with skewed vector integration site distribution *in vivo*.
Schwarzwaelder K, Howe SJ, Schmidt M, Brugman MH, Deichmann A, Glimm H, Schmidt S, Prinz C, Wissler M, King DJ, Zhang F, Parsley KL, Gilmour KC, Sinclair J, Bayford J, Peraj R, Pike-Overzet K, Staal FJ, de Ridder D, Kinnon C, Abel U, Wagemaker G, Gaspar HB, Thrasher AJ, von Kalle C.
J Clin Invest. 2007 Aug;117(8):2241-9.

9 Vector integration is nonrandom and clustered and influences the fate of lymphopoiesis in SCID-X1 gene therapy.
Deichmann A, Hacein-Bey-Abina S, Schmidt M, Garrigue A, Brugman MH, Hu J, Glimm H, Gyapay G, Prum B, Fraser CC, Fischer N, Schwarzwaelder K, Siegler ML, de Ridder D, Pike-Overzet K, Howe SJ, Thrasher AJ, Wagemaker G, Abel U, Staal FJ, Delabesse E, Villeval JL, Aronow B, Hue C, Prinz C, Wissler M, Klanke C, Weissenbach J, Alexander I, Fischer A, von Kalle C, Cavazzana-Calvo M.
J Clin Invest. 2007 Aug;117(8):2225-32.

10 Ectopic retroviral expression of LMO2, but not IL2Rgamma, blocks human T-cell development from CD34+ cells: implications for leukemogenesis in gene therapy.
Pike-Overzet K, de Ridder D, Weerkamp F, Baert MR, Verstegen MM, Brugman MH, Howe SJ, Reinders MJ, Thrasher AJ, Wagemaker G, van Dongen JJ, Staal FJ.
Leukemia. 2007 Apr;21(4):754-63. Epub 2007 Feb 1.

11 Retroviral vector insertion sites associated with dominant hematopoietic clones mark "stemness" pathways.
Kustikova OS, Geiger H, Li Z, Brugman MH, Chambers SM, Shaw CA, Pike-Overzet K, de Ridder D, Staal FJ, von Keudell G, Cornils K, Nattamai KJ, Modlich U, Wagemaker G, Goodell MA, Fehse B, Baum C.
Blood. 2007 Mar 1;109(5):1897-907. Epub 2006 Nov 21.

12 Gene therapy: is IL2RG oncogenic in T-cell development?
Pike-Overzet K, de Ridder D, Weerkamp F, Baert MR, Verstegen MM, Brugman MH, Howe SJ, Reinders MJ, Thrasher AJ, Wagemaker G, van Dongen JJ, Staal FJ.
Nature. 2006 Sep 21;443(7109):E5; discussion E6-7.

13 Human thymus contains multipotent progenitors with T/B lymphoid, myeloid, and erythroid lineage potential.
Weerkamp F, Baert MR, Brugman MH, Dik WA, de Haas EF, Visser TP, de Groot CJ, Wagemaker G, van Dongen JJ, Staal FJ.
Blood. 2006 Apr 15;107(8):3131-7. Epub 2005 Dec 29.

14 Stimulation of Id1 expression by bone morphogenetic protein is sufficient and necessary for bone morphogenetic protein-induced activation of endothelial cells.
Valdimarsdottir G, Goumans MJ, Rosendahl A, Brugman M, Itoh S, Lebrin F, Sideras P, ten Dijke P.
Circulation. 2002 Oct 22;106(17):2263-70.

Appended Publications

Gammaretrovirus-mediated correction of SCID-X1 is associated with skewed vector integration site distribution *in vivo*, Schwarzwaelder *et al*, JCI 2007

Vector integration is nonrandom and clustered and influences the fate of lymphopoiesis in SCID-X1 gene therapy, Deichmann *et al*, JCI 2007

Retroviral vector insertion sites associated with dominant hematopoietic clones mark "stemness" pathways, Kustikova *et al*, Blood 2007

Gammaretrovirus-mediated correction of SCID-X1 is associated with skewed vector integration site distribution in vivo

Kerstin Schwarzwaelder,[1,2,3] Steven J. Howe,[4] Manfred Schmidt,[1,2,5] Martijn H. Brugman,[6] Annette Deichmann,[1,2,5] Hanno Glimm,[1,2,5] Sonja Schmidt,[2] Claudia Prinz,[2] Manuela Wissler,[2,5] Douglas J.S. King,[4] Fang Zhang,[4] Kathryn L. Parsley,[4,7] Kimberly C. Gilmour,[7] Joanna Sinclair,[4] Jinhua Bayford,[7] Rachel Peraj,[7] Karin Pike-Overzet,[8] Frank J.T. Staal,[8] Dick de Ridder,[8,9] Christine Kinnon,[4] Ulrich Abel,[1,10] Gerard Wagemaker,[6] H. Bobby Gaspar,[4,7] Adrian J. Thrasher,[4,7] and Christof von Kalle[1,2,5,11]

[1]National Center for Tumor Diseases, Heidelberg, Germany. [2]Institute for Molecular Medicine and Cell Research and [3]Faculty of Biology, University of Freiburg, Freiburg, Germany. [4]Molecular Immunology Unit, Institute of Child Health, University College London, London, United Kingdom. [5]Department of Internal Medicine I, University of Freiburg, Freiburg, Germany. [6]Department of Hematology, Erasmus Medical Center, Rotterdam, The Netherlands. [7]Department of Clinical Immunology, Great Ormond Street Hospital for Children NHS Trust, London, United Kingdom. [8]Department of Immunology, Erasmus Medical Center, Rotterdam, The Netherlands. [9]Information and Communication Theory Group, Faculty of Electrical Engineering, Mathematics, and Computer Science, Delft University of Technology, Delft, The Netherlands. [10]Department of Medical Biostatistics, Tumor Center Heidelberg-Mannheim, Heidelberg, Germany. [11]Division of Experimental Hematology, Cincinnati Children's Research Foundation, Cincinnati, Ohio, USA.

We treated 10 children with X-linked SCID (SCID-X1) using gammaretrovirus-mediated gene transfer. Those with sufficient follow-up were found to have recovered substantial immunity in the absence of any serious adverse events up to 5 years after treatment. To determine the influence of vector integration on lymphoid reconstitution, we compared retroviral integration sites (RISs) from peripheral blood CD3$^+$ T lymphocytes of 5 patients taken between 9 and 30 months after transplantation with transduced CD34$^+$ progenitor cells derived from 1 further patient and 1 healthy donor. Integration occurred preferentially in gene regions on either side of transcription start sites, was clustered, and correlated with the expression level in CD34$^+$ progenitors during transduction. In contrast to those in CD34$^+$ cells, RISs recovered from engrafted CD3$^+$ T cells were significantly overrepresented within or near genes encoding proteins with kinase or transferase activity or involved in phosphorus metabolism. Although gross patterns of gene expression were unchanged in transduced cells, the divergence of RIS target frequency between transduced progenitor cells and post-thymic T lymphocytes indicates that vector integration influences cell survival, engraftment, or proliferation.

Introduction

Retroviral vectors have been widely used in human HSC gene therapy trials because they stably integrate into the genome and therefore provide an opportunity for sustained clinical effect. This principle has been applied successfully to treat inherited immunodeficiencies including X-linked SCID (SCID-X1) (1–3), adenosine deaminase–deficient SCID (4–6), and, more recently, X-linked chronic granulomatous disease (7). Despite highly encouraging results, evidence has accumulated in animal and human studies that mutagenic side effects occur as a direct result of vector integration (8–12). It has therefore become of particular importance to understand the risks of harmful mutagenesis and to define the patterns of retroviral insertion that may predispose to these events.

Recent studies have shown that the distribution of retroviral integration sites (RISs) within the genome is not arbitrary and is variable in pattern depending on the nature of the virus or vector. Murine leukemia virus– (MLV-), HIV-1–, and avian sarcoma leukemia virus–based (ASLV-based) vectors exhibit quite distinct target site preferences (13). Gammaretroviral vectors and HIV-1–based lentiviral vectors both preferentially integrate into gene coding regions (14), although gammaretroviruses particularly favor a 5-kilobase pair (5-kbp) window on either side of the transcription start site (TSS) (15). In contrast, ASLV exhibits only a weak preference for genes. The mechanisms that dictate the differential integration site patterns have not been clearly elucidated, but may depend to some extent on the accessibility of euchromatin to the preintegration complex, the transcriptional activity of the locus, and binding or tethering to specific DNA sequences via host proteins at the sites of insertion (16). It is therefore likely that integration patterns may also be skewed by the nature and activation status of the target cell.

Although integration patterns are easily defined in homogeneous cell populations in vitro, the influence of integration when measured in complex in vivo situations is more relevant for our understanding of the risks of harmful mutagenesis. In HSC gene therapy, starting cell populations that are transduced ex vivo are heterogeneous, and the minority of progenitor cells among them that do engraft are subject to postengraftment influences that dictate survival, homing to appropriate microenvironmental niches, and subsequent differentiation and proliferation in vivo.

Nonstandard abbreviations used: common γc, 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**:2241–2249 (2007). doi:10.1172/JCI31661.

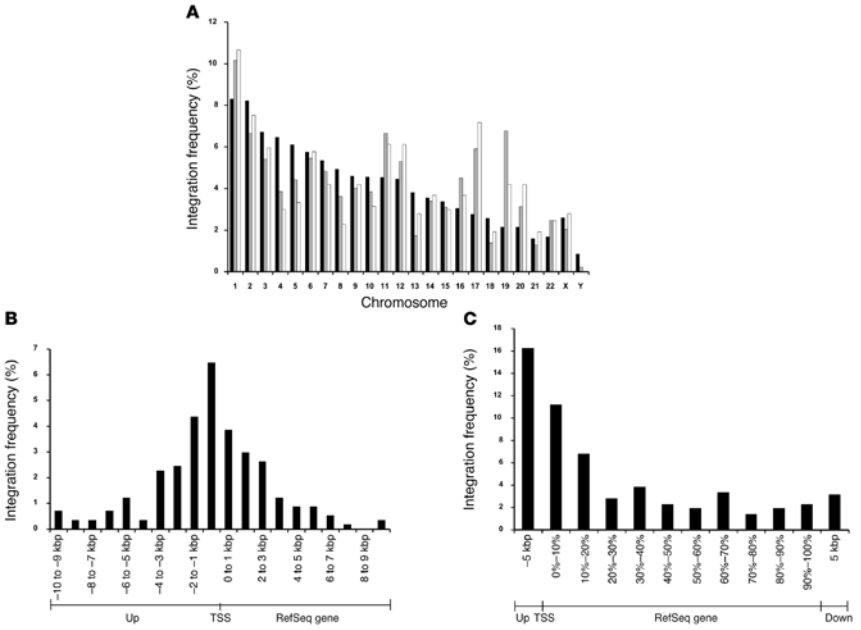


Figure 1
RIS distribution analysis of engrafted cells. (**A**) RIS distribution compared with chromosome size and gene content. The displayed chromosome distribution accounts for the double copy number of diploid autosomes. Black bars, size of chromosomes; gray bars, number of known genes; white bars, 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%; $P = 0.22$, χ^2 test), of targeting the TSS (within 5 kbp of TSS, 16% versus 26%; $P = 0.05$), of insertions in the proxim-

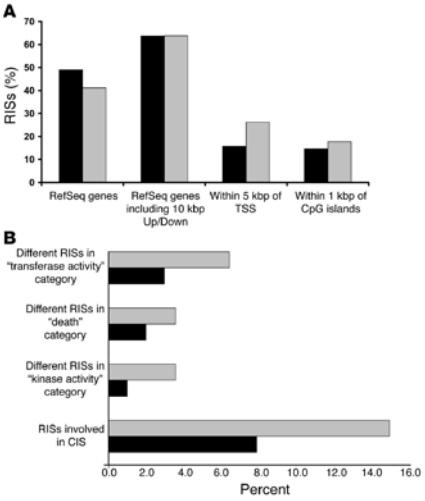


Figure 2
Comparison of pre- and posttransplant RIS distribution in Pt4. (**A**) Percentage of RISs detected in the indicated gene regions. (**B**) Distribution of vector-targeted genes (including the surrounding 10-kbp genomic region) with respect to GO and CIS formation. The GO categories were chosen 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.

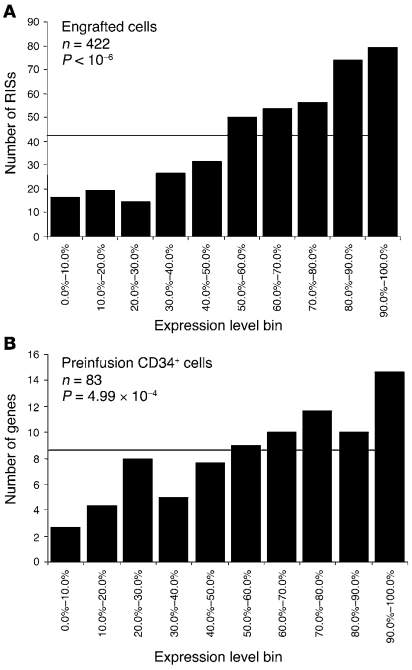


Figure 3
Association between vector integration and gene expression. (**A** and **B**) Number of RISs detected in engrafted cells (**A**) and in CD34+ cells prior to reinfusion (**B**) as a function of relative gene expression in stimulated peripheral blood CD34+ cells. For each gene, the probeset with 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 7×10^{-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 (3.12×10^9 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.

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.

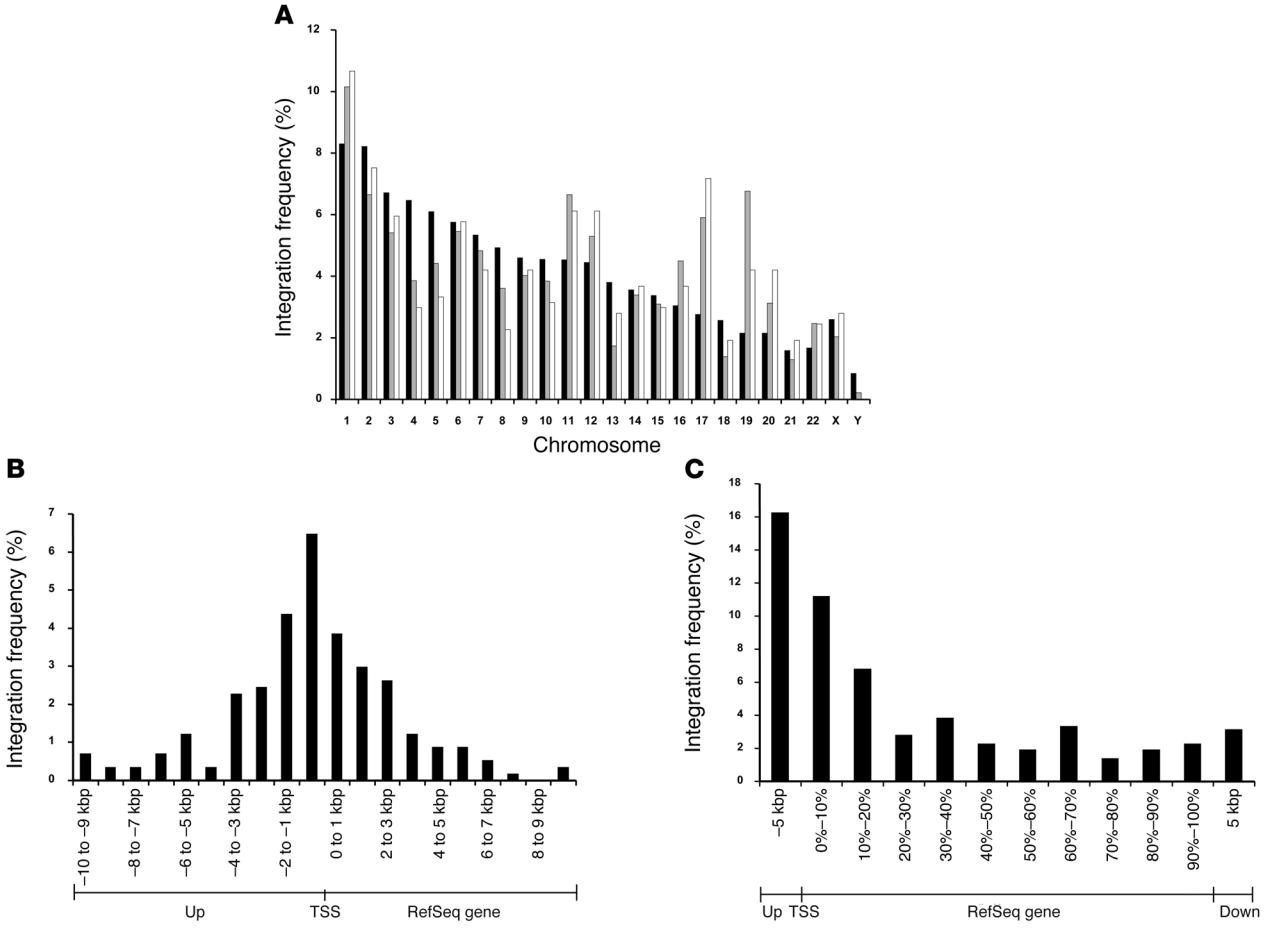


Figure 1
RIS distribution analysis of engrafted cells. (**A**) RIS distribution compared with chromosome size and gene content. The displayed chromosome distribution accounts for the double copy number of diploid autosomes. Black bars, size of chromosomes; gray bars, number of known genes; white bars, 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%; $P = 0.22$, χ^2 test), of targeting the TSS (within 5 kbp of TSS, 16% versus 26%; $P = 0.05$), of insertions in the proxim-

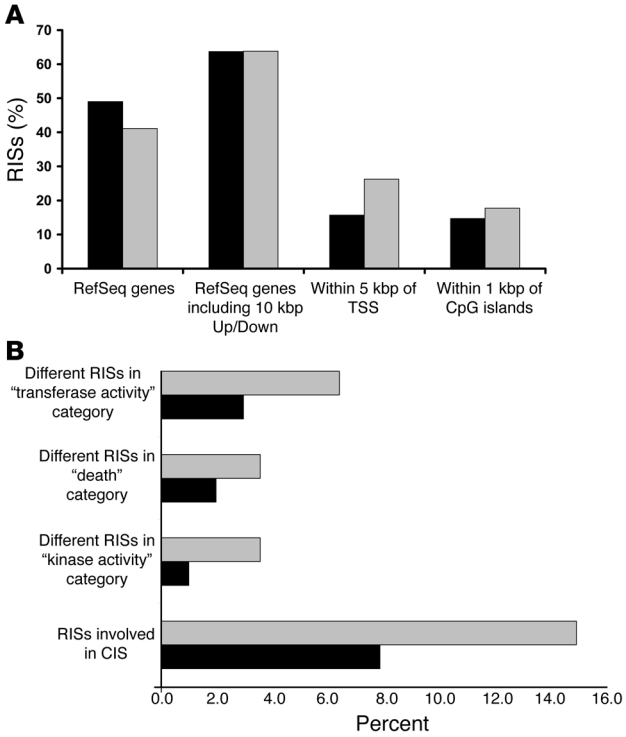


Figure 2
Comparison of pre- and posttransplant RIS distribution in Pt4. (**A**) Percentage of RISs detected in the indicated gene regions. (**B**) Distribution of vector-targeted genes (including the surrounding 10-kbp genomic region) with respect to GO and CIS formation. The GO categories were chosen 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.

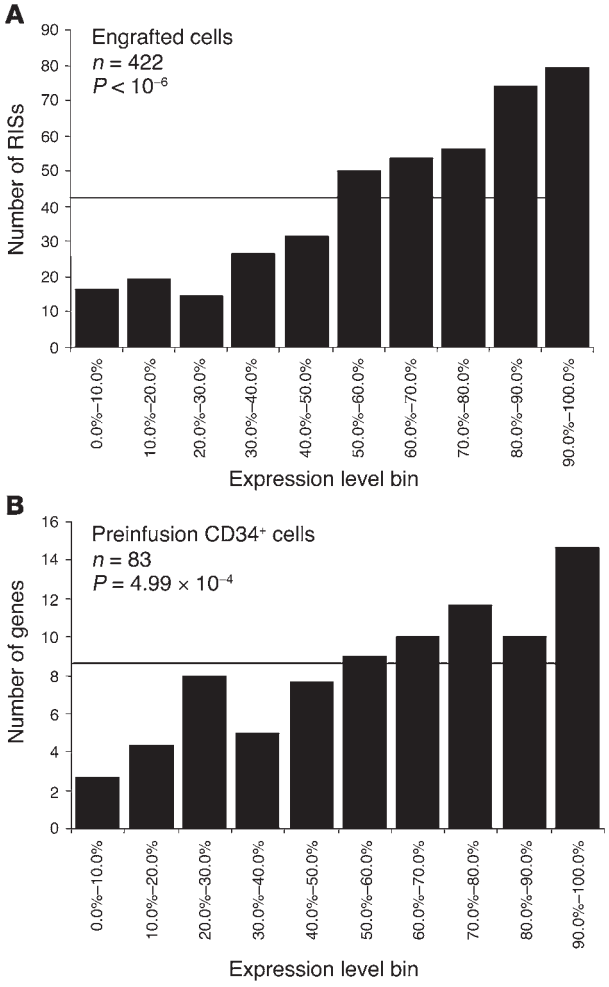


Figure 3
Association between vector integration and gene expression. (**A** and **B**) Number of RISs detected in engrafted cells (**A**) and in CD34+ cells prior to reinfusion (**B**) as a function of relative gene expression in stimulated peripheral blood CD34+ cells. For each gene, the probeset with 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 7×10^{-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 (3.12×10^9 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.

From www.bloodjournal.org at MHH-Bibliothek on September 30, 2008. For personal use only.

Retroviral vector insertion sites associated with dominant hematopoietic clones mark "stemness" pathways

Olga S. Kustikova,[1,2] Hartmut Geiger,[3] Zhixiong Li,[1] Martijn H. Brugman,[4] Stuart M. Chambers,[5] Chad A. Shaw,[6] Karin Pike-Overzet,[7] Dick de Ridder,[8] Frank J. T. Staal,[7] Gottfried von Keudell,[2] Kerstin Cornils,[2] Kalpana Jekumar Nattamai,[3] Ute Modlich,[1] Gerard Wagemaker,[4] Margaret A. Goodell,[5,6] Boris Fehse,[2] and Christopher Baum[1,3]

[1]Department of Experimental Hematology, Hannover Medical School, Germany; [2]Bone Marrow Transplantation, University Medical Center Hamburg-Eppendorf, Hamburg, Germany; [3]Division of Experimental Hematology, Cincinnati Children's Hospital Medical Center, OH; [4]Department of Hematology, Erasmus University Medical Center, Rotterdam, The Netherlands; [5]Center for Cell and Gene Therapy and Cell and Molecular Biology Program, Baylor College of Medicine, Houston, TX; [6]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX; [7]Department of Immunology, Erasmus Medical Center, Rotterdam, The Netherlands; [8]Information and Communication Theory Group, Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, The Netherlands

Evidence from model organisms and clinical trials reveals that the random insertion of retrovirus-based vectors in the genome of long-term repopulating hematopoietic cells may increase self-renewal or initiate malignant transformation. Clonal dominance of nonmalignant cells is a particularly interesting phenotype as it may be caused by the dysregulation of genes that affect self-renewal and competitive fitness. We have accumulated 280 retrovirus vector insertion sites (RVISs) from murine long-term studies resulting in benign or malignant clonal dominance. RVISs (22.5%) are located in or near (up to 100 kb [kilobase]) to known proto-oncogenes, 49.6% in signaling genes, and 27.9% in other or unknown genes. The resulting insertional dominance database (IDDb) shows substantial overlaps with the transcriptome of hematopoietic stem/progenitor cells and the retrovirus-tagged cancer gene database (RTCGD). RVISs preferentially marked genes with high expression in hematopoietic stem/progenitor cells, and Gene Ontology revealed an overrepresentation of genes associated with cell-cycle control, apoptosis signaling, and transcriptional regulation, including major "stemness" pathways. The IDDb forms a powerful resource for the identification of genes that stimulate or transform hematopoietic stem/progenitor cells and is an important reference for vector biosafety studies in human gene therapy. (Blood. 2007;109: 1897-1907)

© 2007 by The American Society of Hematology

Introduction

In analogy to their replication-competent ancestors,[1,2] the semirandom insertion of replication-deficient retrovirus-based vectors may alter cell fate by up-regulating cellular proto-oncogenes or disrupting tumor suppressor genes.[3-12] Such forms of insertional mutagenesis have always represented a safety concern in the development of human gene therapy, although initial studies did not reveal major consequences of random vector insertions.[13] The advent of sensitive technologies to detect vector insertion sites in mixed samples,[14-16] the completion of the murine and human genome projects,[17] the design of improved animal models with long-term follow-up,[3,18] and the increasing efficiency of retrovirus-mediated gene delivery in clinical trials[9,19-22] have all contributed to a revised interpretation of vector-mediated insertional mutagenesis. Clonal imbalance triggered by vector insertion is thus expected to represent the rule rather than the exception.[23-25]

Preclinical models and clinical trials revealed that the semirandom insertion of retrovirus-based vectors in the genome of long-term repopulating hematopoietic cells may increase self-renewal and/or initiate malignant transformation.[3-11] Increased self-renewal can be transitory, resulting in clonal succession such that a given dominant clone is replaced by others over time.[4,9,11] It is likely, although not always formally shown, that replication stress as caused by extended culture of cells prior to transplantation,[5] serial bone marrow transplantation (BMT) in myeloablated recipients,[3] cytotoxic chemotherapy,[10] or chronic infection[9] may trigger the clonal dominance. Long-term observation is required to detect such clones, as the growth kinetics of insertional mutants may be relatively slow and multiple competitor cells are often cotransplanted or present in the host.[3,4,10]

If more than one proto-oncogene is up-regulated by random vector insertion,[5] tumor-promoting sequences are encoded by the vector,[7] or cells with pre-existing tumor-promoting lesions are transduced,[26] clonal leukemias, lymphomas, or sarcomas may result in consequence of random vector insertion, as previously observed in studies with replication-competent retroviruses (RCRs) such as murine leukemia virus (MLV).[1,2,12] In contrast, clonal dominance was not detected following retroviral vector-mediated gene transfer in transplanted T cells, although a fifth of the retroviral vector insertion sites (RVISs) affected the expression of neighboring genes.[27] This supports the conclusion that clonal selection requires a triad consisting of dysregulated expression of genes that regulate cell fitness, a cell type with extensive self-renewal potential, and a milieu with a selection pressure for the fittest mutants.

Submitted August 28, 2006; accepted October 18, 2006. Prepublished online as *Blood* First Edition Paper, November 21, 2006; DOI 10.1182/blood-2006-08-044156.

The online version of this article contains a data supplement.

The publication costs of this article were defrayed in part by page charge payment. Therefore, and solely to indicate this fact, this article is hereby marked "advertisement" in accordance with 18 USC section 1734.

© 2007 by The American Society of Hematology

From www.bloodjournal.org at MHH-Bibliothek on September 30, 2008. For personal use only.

Our work has focused on a relatively simple serial BMT model in C57Bl6 mice. The "normal" genetic background of this strain, the relatively low incidence of host-derived tumors (< 3% under our experimental conditions), and the availability of an allelic variant in the CD45 panleukocyte antigen in a congenic strain (B6 CD45.1) to distinguish donor and host cells render this model particularly attractive for gene discovery by and preclinical safety studies of retroviral gene transfer into hematopoietic cells.

In the present report, we summarize data from several laboratories that used this model to develop a database of RVISs detected in dominant clones contributing to phenotypically intact, mildly dysplastic, and overtly malignant hematopoiesis. We describe the validation of our experimental conditions to detect genetic lesions underlying clonal dominance, and several important genetic and biological insights obtained from the newly established insertional dominance database (IDDb). These analyses underline the validity of our approach to discover genes that regulate fitness and potentially transform self-renewing cells in vivo, promoting a systematic extension for both gene discovery and vector biosafety studies in the context of different cell types and selection conditions.

Materials and methods

Transplantation conditions and analysis of healthy and leukemic hematopoiesis in mice

All BMT studies were performed in C57BL/6 mice. In brief, donor bone marrow cells were cultured ex vivo to stimulate gene transfer using vectors based on MLV, and cells were transplanted into lethally irradiated recipients aged 12 to 16 weeks. Mice were kept in the animal facilities of the participating institutions, according to local animal experimentation guidelines. Food and water were supplied ad libitum. Table 1 summarizes the transplantation conditions and vectors used (for further details, see Document S1, available on the *Blood* website; see the Supplemental Materials link at the top of the online article). Mice were humanely killed when symptomatic (leukemic) or after 2.5 to 7 months in the healthy cases and examined for pathologic abnormalities, including histologic, morphologic (blood smears and cytospins of bone marrow and spleen), and flow cytometry analyses.[5] Animal experiments were approved by the institutional animal research review boards of the principal investigators listed in Table 1.

Cell culture

K562 cells were cultivated and transduced as described.[28]

Ligation-mediated polymerase chain reaction

Ligation-mediated polymerase chain reaction (LMPCR) was performed as described.[4,5,15]

Insertion site analysis

Fragments containing retroviral genomic junctions were submitted to further analysis using the following websites: BLAST[29] searches were performed or, in some cases at Ensemble[30]; the mouse Retrovirus Tagged Cancer Gene Database (RTCGD)[31]; and/or the stem cell database (SCDb)[32] were used. Gene Ontology (GO) describes genes' biological roles and is arranged in a quasi-hierarchical structure from more general terms to more specific. To determine abundance for each GO category, the frequency of retroviral inserts was calculated and compared with the expected frequency observed by chance, as described.[33] GO analysis was confirmed by the Expression Analysis Systematic Explorer (EASE).[34]

Expression arrays

Mouse bone marrow cells were depleted from lineage-committed cells (CD5, CD45R [B220], CD11b, anti–Gr-1, 7-4m and Ter-119; Lineage depletion kit; Miltenyi Biotec, Bergisch-Gladbach, Germany) using AutoMACS (magnetic cell sorter) (Miltenyi Biotec) in 2 independent experiments. The lineage-depleted cells were selected for $CD117^+$ cells (c-kit selection kit; Miltenyi Biotec). $Lineage^-$/C-Kit^+/$Sca\text{-}1^+$ (LSK) cells were selected on a fluorescence-activated cell sorting (FACS) DiVa (BD Biosciences, San Jose, CA). Purity for both experiments was greater than 96%. RNA was isolated (Qiashredder and RNeasy; QIAGEN, Hilden, Germany) directly after sorting (day 0) or after maintaining the cells in serum-free medium supplemented with mSCF, mTPO, and Flt3L for 2 days. Quality was assessed using an Agilent 2100 BioAnalyzer (Agilent Technologies, Palo Alto, CA). Total RNA (100 ng) from LSK cells was used in the GeneChip Eukaryotic Small Sample Target Labeling Assay Version II (Affymetrix, Santa Clara, CA) to generate biotinylated cRNA. cRNA (11 μg) was fragmented for 35 minutes at 95°C. Fragmented cRNA (10 μg) was then hybridized to mouse 430 2.0 microarray (Affymetrix) for 16 hours at 45°C followed by washing, staining, and scanning at 570 nm, according to standard methods.[35] The expression data were normalized as described.[36,37] For each gene, the highest expression was determined. For some, only the most highly expressed probe set was used. To determine the association of vector insertion with gene expression, a Cochran-Armitage test for trend was performed.[38]

Pathway analysis

Gene symbols were entered into Netaffx (http://www.Affymetrix.com) and the corresponding Affymetrix IDs for the mouse 430 2.0 arrays were retrieved. The resulting Affymetrix IDs were entered in the Ingenuity

Table 1. Overview of murine bone marrow transplantation (BMT) experiments

Vector cDNA	Mice, no.	Principal investigator*	Donor cells (C57B16/J)	Hosts	Observation time, mo	
					First BMT	Second BMT
EGFP	8	B.F.	Lin^- BM cells, $CD45.1^+$	C57BL/6/J, $CD45.2^+$	6	7
EGFP	3	Z.L.	Lin^- BM cells $CD45.2^+$	C57BL/6/J, $CD45.2^+$	6	—
IRES.EGFP	7	H.G.	Low density BM cells, CD45.2, after 5-FU	C57BL/6/J, $CD45.1^+$	2.5	5
DsRed2	12	U.M.	Lin^- BM cells, $CD45.1^+$	C57BL/6/J, $CD45.2^+$	7	7
XRCC4	10	H.G.	Low-density BM cells, CD45.2, after 5-FU	C57BL/6/J, $CD45.1^+$	2.5	5
flCD34	11	Z.L.	Whole BM cells, $CD45.2^+$	C57BL/6/J, $CD45.2^+$	7	4.5
tCD34	17	Z.L.	Whole BM cells, $CD45.2^+$	C57BL/6/J, $CD45.2^+$	7	4.5
mflCD34	11	B.F.	Lin^- BM cells, $CD45.1^+$	C57BL/6/J, $CD45.2^+$	6	7
mtCD34	10	B.F.	Lin^- BM cells, $CD45.1^+$	C57BL/6/J, $CD45.2^+$	6	7
dLNGFR	16	Z.L.	Whole BM cells, $CD45.2^+$	C57BL/6/J, $CD45.2^+$	7	4.5
MDR1	8	U.M.	Lin^- BM cells, $CD45.1^+$ or whole BM cells $CD45.1^+$	C57BL/6/J, $CD45.2^+$	7	7
TAg	7	Z.L.	Lin^- BM cells, $CD45.2^+$	C57BL/6/J, $CD45.2^+$	6	—
TAg	2	Z.L.	32D cells	C3H/Hej	5	—

See legend of Figure 1 for abbreviations of cDNAs.
— indicates no second BMT.
*Initials of author.

From www.bloodjournal.org at MHH-Bibliothek on September 30, 2008. For personal use only.

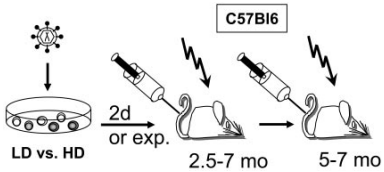


Figure 1. Experimental setup of murine BMT studies using donor cells modified with different retroviral vectors. The enhancer-promoter contained in the long terminal repeat (LTR), the cDNA encoded by the vector, and the 3′ untranslated region (3′ UTR) are indicated in Table 2. LD indicates low dose of retroviral vector; HD, high dose; and exp, expansion in vivo.

Pathway Analysis tool (http://www.ingenuity.com) to generate direct and indirect pathways. For each dataset, the 10 functions and diseases with the most genes assigned to it are displayed.

Results

Experimental setup

The RVISs described in this study are derived from murine experiments (mostly C57Bl6), using several replication-deficient MLV-based vectors for gene transfer into ex vivo–cultured hematopoietic cells. The vectors used include a group encoding fluorescent proteins (EGFP, DsRed), a group encoding transmembrane proteins that serve as selection markers (dLNGFR, human tCD34 and flCD34, murine tCD34 and flCD34, MDR1), and a vector expressing a gene associated with DNA repair (XRCC4). As a positive control for a transforming vector expressing a strong oncogene, the large T antigen (TAg) from simian virus 40 (SV40) was used (Figure 1, Table 2). TAg transforms cells by sequestering 2 tumor suppressor genes, Rb and p53.[39] The transforming potential of the TAg vector was initially evaluated in 32D cells, revealing insertion sites with potential contribution to transformation (Z.L., unpublished data, January 2006). Four RVISs from these studies were also included in the IDDb (1.4% of the database).

If the vectors do not encode oncogenic sequences, RVISs present in dominant clones may mark events that initiate increased self-renewal.[4] Importantly, we noted transcriptional dysregulation of the mutated alleles in all cases tested so far.[4] If the vectors encode oncogenic sequences such as TAg, the insertional events may either collaborate with the encoded oncogene to initiate tumor formation or promote the expansion of dominant malignant clones whose initial transformation is primarily dependent on the vector-encoded oncogene.[7] Mice were prospectively examined for several months; in a subset of the studies, serial BMT was performed to increase replicative stress and observation time (Figure 1; Table 1).

Validation of LMPCR

Different methods have been described to recover insertion sites from retrovirally transduced cells.[14-16,27,40,41] To identify insertion sites of dominant clones, it was crucial to neglect insertion sites present in minor clones. Ligation-mediated PCR (LMPCR) as opposed to the much more sensitive "linear amplification-mediated PCR" (LAMPCR) has previously been shown to lack the sensitivity to detect all insertion sites present in highly polyclonal samples.[16] However, we noted that the bands obtained by LMPCR correlated well with Southern blot results obtained in clonal samples, and recovery of RVISs was in the range of 80% when using a single restriction enzyme.[4,5,42] We thus decided to select dominant bands that are isolated from analytical gels for direct sequencing, ignoring weak bands that might reflect insertion sites present in minor clones.

We validated this approach by examining DNA from K562 clones that contained a known number of retroviral vector insertions and DNA from a K562 mass culture obtained after transduction with a high MOI of a marking vector.[28] Although LMPCR reproducibly showed "dominant bands" of molecular weights ranging from 100 to 800 base pair (bp) in clonal samples, polyclonal DNA yielded a smear of multiple minor bands (Figure 2A). To examine the minimal proportion of clonal DNA required for detection of dominant bands, we mixed DNA from a clone with 6 insertions (validated by Southern blot, not shown) with DNA from a polyclonal retrovirally transduced mass culture. If the clonal DNA constituted greater than 70% of the sample, LMPCR reproducibly revealed its insertion sites as dominant bands, whereas minor PCR products progressively disappeared. Major PCR products were recovered largely irrespective of their size (Figure 2B).

Direct sequencing of the PCR product confirmed the presence of RVISs (data not shown). We typically performed 2 LMPCRs to confirm reproducibility.

Composition and content of the IDDb

The sequence data obtained by LMPCR were blasted against the mouse genome to identify genes potentially affected by the insertion site. We also examined whether the hit loci were contained in the RTCGD,[2] and listed the experimental conditions as these may affect selection (vector, transplantation conditions, and potential development of malignancy; Table S1). In total, we identified 276 RVISs from a total of 120 C57Bl6 mice (receiving retrovirally engineered bone marrow cells), and 4 RVISs from 2 C3H/Hej mice (developing leukemia after receiving 32D cells transduced with a TAg vector). On average, we thus retrieved 2.3 insertions per animal, reflecting the low number of dominant clones. Only 16.4% of these mice presented with leukemia, manifesting with a latency of 5 to 10 months after gene transfer.

Table 2. Modules of retroviral vectors used in this study

Vector	LTR	cDNA	3′UTR
SF91DsRed2	SFFVp	DsRed2	wPRE
SF91EGFP	SFFVp	EGFP	wPRE
SF91IRESEGFP	SFFVp	IRES-EGFP	wPRE
SF91XRCC4	SFFVp	XRCC4-IRES-EGFP	wPRE
SF11flCD34	SFFVp	flCD34	—
SF11tCD34	SFFVp	tCD34	—
SF11mflCD34	SFFVp	mflCD34	—
SF11mtCD34	SFFVp	mtCD34	—
SF11dLNGFR	SFFVp	dLNGFR	—
HaMDR	HaMSV	MDR1	VL30
SF91TAg	SFFVp	EGFP2ASV40TAg	—

A high MOI was used in some experiments involving vectors HaMDR1 and SF91dsRed2.[5]

SFFV indicates spleen focus-forming virus; HaMLV, Harvey murine leukemia virus; EGFP, enhanced green fluorescent protein; DsRed2, red fluorescent protein; IRES-EGFP, internal ribosomal entry site followed by EGFP; XRCC4, x-ray repair complementing defective repair in Chinese hamster cells 4; flCD34, human full-length CD34; —, no additional element in 3′ UTR; tCD34, human truncated CD34; mflCD34, murine full-length CD34; mtCD34, murine truncated CD34; dLNGFR, deleted low-affinity nerve growth factor receptor; MDR1, multidrug resistance 1; EGFP2ASV40TAg, EGFP fusion protein linked to the large T antigen (TAg) of simian virus 40 using a self-cleaving 2A proteinase sequence; VL30, virus-like element 30; wPRE, woodchuck hepatitis virus posttranscriptional regulatory element.

From www.bloodjournal.org at MHH-Bibliothek on September 30, 2008. For personal use only.

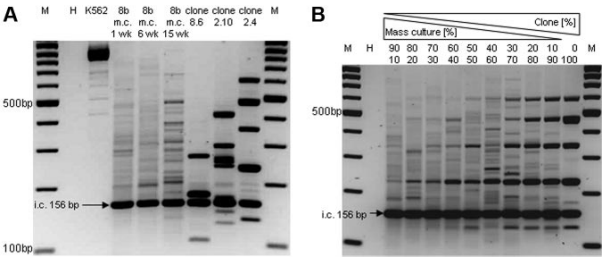


Figure 2. LMPCR validation. (A) DNA of K562 mass cultures and cell clones containing different numbers of retroviral insertions[28] was subjected to insertion site amplification by LMPCR using the conditions described in "Material and methods." In contrast to the clonal DNA, mass culture DNA does not reveal dominant bands except when cells were propagated for several weeks, revealing a clonal imbalance. (B) Mixing mass culture DNA with increasing amounts of DNA from clone 2.4 reveals that LMPCR recovers dominant bands if these contribute greater than 70% of the population.

Overall, 22.5% of the RVISs contained in the IDDb are located in or near to known proto-oncogenes as defined by the RTCGD[2] and additional literature (http://www.ncbi.nlm.nih.gov/entrez/), 49.6% in genes encoding proteins involved in various processes of cell signaling, 20% in other (often metabolic), and 7.9% in unknown genes. When bone marrow cells were transplanted to secondary recipients, the proportion of insertions in proto-oncogenes increased from 15% (primary recipients) to 24% (secondary recipients) in mice with normal hematopoiesis (Figure 3A). Thus, the IDDb perfectly reproduced the findings of our previous study performed in mice that showed no signs of hematopoietic malignancies.[4] Considering that proto-oncogenes represent 1.06% (n = 231) of the murine genome (Entrez Gene, May 11, 2006), this is a gross overrepresentation. For comparison, 37% of the RVISs recovered from leukemias were in or close to proto-oncogenes (Figure 3A), strongly suggesting that the RVISs were causally involved in promoting a competitive advantage and inducing transformation.[5]

We next asked whether the different transgenes encoded by the vectors affected clonal selection. We subdivided the hit genes into 4 groups: proto-oncogenes as defined by the RTCGD[2,31] and additional literature, signaling genes, other genes, and unknown genes. EGFP and DsRed encode fluorescent proteins which are not known to cause significant changes of signaling networks. Twenty-five percent (n = 70) of the hits were recovered using these vectors. In this subgroup, the distribution of hits in the 4 gene groups was almost identical to that obtained within the set of transgenes that encode surface marker proteins for which an effect on cellular signaling cannot be ruled out (MDR1, dLNGFR, human tCD34, human flCD34, murine tCD34, murine flCD34, XRCC4) (Figure 3B). In contrast, a control group in which the transgene encoded the potent oncoprotein TAg of SV40 showed a distinctively higher proportion of RVISs in proto-oncogenes (41% versus ~ 20% with other vectors). This supports the conclusion that RVISs recovered from tumors induced by replication-deficient vectors encoding oncogenes contribute to clonal selection.[7]

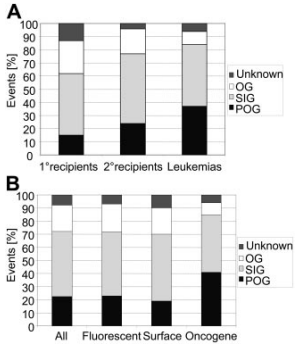


Figure 3. Retroviral vector insertion site (RVIS) distribution according to gene classes and type of transgene. (A) RVISs in known proto-oncogenes (POGs) increase in frequency over serial BMT and are most pronounced in leukemic clones. (B) No major impact of the transgene class was found except when the vector encoded a potent oncogene (TAg), which increased the probability to select for RVIS in POGs. SIGs indicates signaling genes; OGs, other genes.

Interestingly, RVISs of the oncogenic TAg vectors overlapped with those observed in healthy retrovirally marked hematopoiesis exhibiting clonal dominance. Four of the 13 proto-oncogene hits observed in tumors induced by TAg vector insertion were also observed in dominant clones transduced by other vectors *(Sema4b, AB041803, BC013781, Fli1)*, and 3 additional proto-oncogenes hits selected in TAg vector-transduced tumors occurred in gene families that were also marked using other vectors (eg, *BclX* was hit by the TAg vector and *Mcl1* by the dLNGFR vector; growth factor receptors *Axl* and *Csf1r* were hit by TAg vectors and *Csfr3* by the DsRed vector).

In further support of selection for clonal dominance largely irrespective of the type of transgene encoded, 4.6% (n = 13) of RVISs affect the *Mds/Evi1* locus which encodes a transcription factor expressed in primitive hematopoietic cells.[43] Rearrangement and ectopic expression of this allele contributes to human and murine leukemia.[43] *Evi1* represents the third most frequent insertion site listed in the RTCGD.[2,31] Sixteen percent of the other RVISs found in the IDDb are common RVISs (CRVISs); that is, independent insertion sites recovered from different cell clones but affecting the same gene. Because CRVISs are a strong indication of selection for an important biological function,[2] it is interesting that only 52% of the IDDb-CRVISs represent known proto-oncogenes (Table 3). Summarizing all RVISs in known proto-oncogenes, those forming novel CRVISs in our database and those occurring in genes with an established role in stem cell self-renewal and hematopoiesis, a group of at least 48 genes encoding growth factors, signal transducers, and transcription factors can be extracted which represent interesting candidates for future functional studies (Table 3). Interestingly, 81% of these RVISs were found in secondary and/or leukemic transplant recipients.

Insertion site distribution in relation to the transcription start site

To further address the potential selective pressure present on the mutated alleles, we analyzed the distribution of the RVISs with

From www.bloodjournal.org at MHH-Bibliothek on September 30, 2008. For personal use only.

Table 3. Insertions in known proto-oncogenes, genes with an established role in hematopoiesis, and common insertion sites in the insertional dominance database (IDDb)

Type of protein and locus	No. hits in IDDb	No. hits in RTCGD	Gene ID	Chromosome	Name/(proposed) function
Growth factor					
FasL	2	0	14103	1 H2.1	Fas ligand (TNF superfamily, member 6)
Vegfa	1	0	22339	17B3	Vascular endothelial growth factor A
Tnfsf10	2	2	22035	3A3	Tumor necrosis factor (ligand) superfamily, member 10, apoptosis induction
Receptor					
Sema4b	3	4	20352	7D1	Receptor activity
Axl	1	0	26362	7A3-B1	Receptor activity, protein kinase activity, human proto-oncogene
Csflr	1	9	12978	18D	Receptor of colony stimulating factor 1 (CSF-1)
Csfr3	1	0	12986	4D2+2	G protein coupled receptor for granulocyte colony stimulating factor (G-CSF)
Gpr43 = Ffar2	2	0	233079	7A3	G-protein coupled receptor, free fatty acid receptor
Igfbp4	2	0	16010	11 D-E1	Insulin-like growth factor binding protein 4
Ly78	2	0	17079	13D1	Lymphocyte antigen 78, receptor, signaling, CD180
Signal transducer					
Akt1	2	3	11651	12F1-F2	Intracellular signaling, kinase transforming protein
Bcl211	1	3	12048	2H1	Bcl2-like1, antiapoptosis
Ccnd3	1	13	12445	17B4	Cyclin D3, cell cycle
Mcl1	1	4	17210	3F2.1	Myeloid cell leukemia sequence 1, Bcl2-related antiapoptotic protein
Osbpl3	1	2	71720	6B3	Oxysterol binding protein-like 3, steroid metabolism
Pim2	2	5	18715	X A1.1	Serine/threonine-protein kinase Pim-2, proviral integration site 2, antiapoptosis
Plcg2	2	0	234779	8E1	Phospholipase C, gamma 2, survival signaling
Rab3gap2	2	0	381313	1H5	Similar to Rab3 GTPase activating protein, gene model 981
Rhof	1	4	23912	5F	Ras homolog gene family member
Sesn2 = Hi95	3	0	230784	4D2.3	Sestrin 2, induction in response to DNA damage
Transcription factor					
2610510B01Rik = Dopey2	1	4	70028	16 C4	Predicted leucine zipper transcription factor, dopey family
Bcll1a	2	6	14025	11A3+2	Zinc finger, essential for lymphopoiesis
Cbfa2t3h	1	7	12398	8E1	Core-binding factor, runt domain, alpha subunit 2, translocated 2, 3 homolog
Cutl1	1	3	13047	5 G2	Cut-like 1 (Drosophila) = CCAAT displacement protein = Cux/CDP homeoprotein
Elk4	1	3	13714	1E3-G	ELK4, ETS family member
Evi1	13	20	14013	3A3	Associated with murine and human leukemia, SMAD interacting
Fli1	2	7	14247	9A4	ETS family member
Fos (LOC627366)	2	5	14281	12 D2	FBJ osteosarcoma oncogene
Fosb (Ercc1)	1	0	14282	7A2	FBJ osteosarcoma oncogene B
Foxo3a	2	0	56484	10B2	Forkhead transcription factor, potentially pro-apoptotic
Gtf2i	2	0	14886	5G2	General transcription factor 2
Hhex	1	26	15242	19Cl	Hematopoietically expressed homeobox gene (T-cell oncogene)
Hic1	3	3	15248	11B5	Hypermethylated in cancer 1, transcription factor, Wnt antagonism
Hivep1	1	2	110521	13A4	HIV enhancer binding protein 1
HoxA7	2	19	15404	6B3	Homeobox gene A7 and surrounding cluster
Hoxb4/Hoxb5	1	0	15412	11D	Homeobox gene B4 and surrounding cluster
Lmo2	2	2	16909	2 E2	LIM domain only 2, T-cell oncogene
Mef2d	1	5	17261	3F1	Myocyte enhancer factor 2D
Mllt3	1	0	70122	4C4	Involved in leukemogenic translocations
Runx2	1	6	12393	17B3	Runt related, essential for hematopoiesis
Runx3	1	5	12399	4D2.3	Runt related, myeloid development
Sox4	1	64	20677	13A3-A5	Sry box, lymphocyte activation
Tal1	1	0	21349	4D1	Ebox family member, essential for hematopoiesis
Zfp3612	1	4	12193	17 E4	Zinc finger protein 36, C3H type-like
Metabolic					
Dph5	2	0	69740	3F3	DPH5 homolog (*Saccharomyces cerevisiae*) transferase
Unknown					
Lrrc6	2	0	54562	15D1	Leucin repeat containing 6 (testis)
Dym = 4933427L07Rik	1	3	69190	18 E2	Dymeclin, function unknown
AB041803	2	8	232685	6A3.3	Hypothetical protein, function unknown
BC031781	1	2	208768	1H4	Hypothetical protein, function unknown

The classification of proto-oncogenes follows the listing in the retrovirus-tagged cancer gene database (RTCGD; http://rtcgd.ncifcrf.gov/) and additional literature. A version of the table with the gene ID configured as a hyperlink to NCBI Entrez Gene (http://www.ncbi.nlm.nih.gov/entrez/) is available from the authors (http://www99.mh-hannover.de/kliniken/zellth/method.html). In the case of *Fos* and *Fosb*, Table S1 lists the loci shown in brackets.

Equals sign indicates alternative gene names.

respect to the transcriptional start site (TSS) of the next neighboring gene. In unselected freshly transduced cells, MLV vectors have a preference for insertions in the 10-kb (kilobase) window around the TSS,[40] with a peak in the ± 1-kb window,[41] whereas HIV and derived vectors tend to prefer actively transcribed sequences, in particular beyond +2 kb downstream of the TSS.[40,41,44] The reference data obtained in previous studies (kindly provided by D. Russell and G. Trobridge; Trobridge et al[41]) are shown in Figure 4A

From www.bloodjournal.org at MHH-Bibliothek on September 30, 2008. For personal use only.

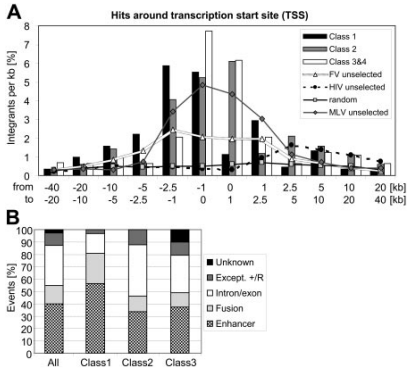


Figure 4. Type of mutations. Data are shown with respect to gene class 1 (common insertion sites, proto-oncogenes, and self-renewal genes), class 2 (signaling genes), and classes 3 and 4 (other and unknown genes). (A) Position of RVIS in the Insertional Dominance Database (IDDb) around the transcriptional start site. Reference data insertion sites of different vectors in freshly transduced cells, shown as lines, were kindly provided by G. Trobridge and D. Russell.[41] MLV indicates murine leukemia virus vector; FV, foamy virus vector; HIV, human immunodeficiency virus vector; random, computer-predicted random insertion pattern. (B) Overrepresentation of enhancer mutations in class 1 genes. RVISs were analyzed for the different types of retroviral insertional mutations proposed earlier.[12] Insertions located downstream but in an antisense orientation do not correspond to the definition of enhancer mutations suggested in Uren et al[12] and are therefore labeled "Except. +/R."

(lines), in comparison with our database (columns). For this comparison, we divided the IDDb hits into 4 classes: class 1 consists of known self-renewal genes, proto-oncogenes, and CRVISs present in the IDDb (Table 3); class 2 represents genes with a known or putative role in cellular signaling networks; class 3 collects other genes; and class 4 unknown genes.

Compared with the insertion pattern of MLV in unselected cells, the IDDb shows a clear overrepresentation of class 1 events in the window between -1 and -20 kb, and also between 5 and 10 kb downstream of the TSS (Figure 4A). No overrepresentation is found within 1 kb upstream of the TSS, and class 1 hits are even underrepresented in the first kilobase of transcribed sequence. A similar picture is observed in the window around +2.5 to 5 kb. Events in classes 3 and 4 serve as an internal control, showing no enrichment over the unselected MLV pattern in the windows around -5 to -1 kb and no counterselection in the +1-kb window. The region that is most likely to contribute to clonal dominance thus resides within -1 to -5 kb upstream of the TSS, whereas insertions closely downstream of the TSS tend to be counterselected.

Vectors based on foamy virus and even more so those based on lentiviruses have been shown to have a reduced bias for the region surrounding the TSS.[40,41] The IDDb with its focus on genes that support competitive fitness reveals that a simple switch to these vector types may not fully eliminate the risk of insertional mutagenesis. Looking at the window 5 kb upstream of the TSS, a switch to foamy virus–based vectors[41] might reduce the probability of "productive" class 1 insertions by a factor of less than 2 and a switch to HIV-based vectors by a factor of 10. However, hits in this window only account for less than 20% of class 1 events in our database. For the majority of events located further upstream or downstream, changing the retroviral backbone does not seem to change the risk.

The position and orientation of the vector with respect to the transcription unit allows a classification of insertional mutations as follows[12]: enhancer mutations are typically located upstream of the transcription unit in the antisense orientation or downstream in the sense orientation, fusion transcript mutations may originate from insertions upstream of transcription units in the sense orientation, and insertions within a transcription unit may lead to aberrant splicing or termination. In the IDDb, 40% (111 of 280) of the RVISs represented enhancer mutations, the majority (76%) occurring upstream of the TSS in the antisense orientation. Enhancer mutations were more relevant in class 1 genes (55%) than in the other classes (~ 34%). Fusion mutations represented 20% of the events in class 1, and approximately 14% in the other classes. Accordingly, insertions within transcription units were underrepresented in class 1 compared with the other classes (Figure 4B).

Together, the enrichment of insertions in class 1 genes over serial transplantation and with leukemic progression (Figure 3), the skewed distribution of insertions around the TSS (Figure 4A), and the counterselection of insertions within transcription units (Figure 4B) in favor of enhancer and fusion mutations all reveal that insertional mutations strongly contributed to the occurrence of clonal dominance in our experiments.

Overlap with stem cell databases and pathway analysis

MLV vectors preferentially target active genes, but extremely high gene expression levels might impede insertions.[45] To explore whether the RVISs selected in vivo represent genes expressed in primitive hematopoietic cells, as suggested from a previous study conducted with human cells,[46] we compared the genes listed in the IDDb with 3 different transcriptome databases. The first is the publicly accessible stem cell database (SCDb),[32] which represents a subtracted cDNA library derived from primitive hematopoietic cells present in murine fetal liver and marrow. The second database was generated from a genome-wide gene expression profiling experiment using Affymetrix array full genome mouse arrays on RNA extracted from highly purified hematopoietic stem/progenitor cells (Lin^- $Sca1^+$ $c\text{-}Kit^+$, LSK, > 96% pure after flow sorting) obtained from steady state murine bone marrow (M.H.B., K.P., D.R., F.J.T.S., M. M. A. Verstegen, G. W., unpublished observations, July 2006), and the third database was generated using the same array and RNA extracted from highly purified hematopoietic stem cells (side population [SP] combined with LSK)[47] (S.M.C. and M.A.G., unpublished data, July 2006).

We found 57% of the class 1 genes to be listed in the SCDb, as opposed to 32% for class 2 and 17% for class 3. With reference to the GO classification used in the SCDb, the IDDb shows an overrepresentation of genes encoding proteins involved in apoptosis, intracellular signaling, or transcriptional control, whereas the following gene classes are strongly underrepresented: cell adhesion, transport, chromatin regulators, protein processing, and protein synthesis (Table 4).

We further studied GO in its branching into a semihierarchical tree, describing genes in categories from very general (ie, regulation of biological process, levels 1-5) to very specific (level 10+). This analysis showed a highly significant ($P < .05$, hypergeometric) overrepresentation of the following processes: cell proliferation (level 4, $P = .016$), positive regulation of apoptosis (level 7, $P = .033$), and regulation of transcription, DNA-dependent (level 8, $P = .001$).

A network-based pathway analysis demonstrated that RVISs clustered near genes involved in cancer and were, in addition,

From www.bloodjournal.org at MHH-Bibliothek on September 30, 2008. For personal use only.

Table 4. Genes associated with clonal dominance preferentially belong to three GO categories: intracellular signaling, transcription factor, and apoptosis

Process	IDDb, %	SCDb, %	IDDb/ SCDb	Conclusion
Intracellular signaling	27	17	1.59	Overrepresented in IDDb
Transcription factor	23	14	1.64	Overrepresented in IDDb
Apoptosis	6	1	6.00	Overrepresented in IDDb
Cell adhesion	1	7	0.14	Underrepresented in IDDb
Transport	2	7	0.29	Underrepresented in IDDb
Chromatin regulators	0.5	3	0.17	Underrepresented in IDDb
Protein processing	0.5	7	0.07	Underrepresented in IDDb
Protein synthesis	0.5	4	0.13	Underrepresented in IDDb
Receptors	5	10	0.50	Underrepresented in IDDb
Metabolism	10	13	0.77	Similar representation
Unknown	17	NA	NA	Not comparable

Data of the stem cell database (SCDb) were derived from Ivanova et al.[62]
NA indicates not applicable.

strongly correlated with genes involved in hematologic and immune system development, functions, and disease (Tables S2 and S3). Canonical pathway analyses performed with IDDb genes revealed a significant overrepresentation of growth factor signaling pathways, death receptor signaling pathways, and associated intracellular networks (Table 5). Strikingly, most of the genes extracted in Table 3 are connected in 2 major networks (Figure 5). Figure 5A shows major pathways contributing to hematopoietic stemness (*Igf-1, Vegf*, *Pten*, apoptosis, death receptor), whereas Figure 5B reveals the association with additional nuclear players involved in hematopoietic self-renewal and lineage commitment.

This suggested that the RVISs selected in the IDDb occurred preferentially in a subset of genes expressed in primitive hematopoietic cells. We further approached this question by comparing the genes listed in the IDDb with gene expression microarray data obtained from purified fractions of hematopoietic stem/progenitor cells. With respect to the most primitive fraction analyzed, SP-LSK, RVISs present in the IDDb were clearly associated with expressed genes ($P < .01$, Wilcoxon test). Interestingly, the level of significance increased depending on serial transplantation and the degree of transformation: primary recipients ($P = .003$), secondary recipients ($P < .001$) and leukemias ($P < .001$). This reveals that the vast majority of genes whose deregulation causes clonal dominance is already expressed in primitive hematopoietic cells, rather than being activated from a silent state by insertional mutagenesis.

As the initial target population of retroviral gene transfer was not such a highly purified fraction, we also used gene expression array data from LSK cells to check whether the level of transcription correlates with RVISs. LSK cells contain both short-term and long-term repopulating cells,[48] and it is possible that some RVISs converted short-term to long-term clones, as can be observed in consequence of certain oncogenic translocations[49] (and references therein). On the basis of their relative expression level, genes were classified into 10 "bins" such that bin 1 represents the 10% of genes with the lowest expression levels, and bin 10 the 10% of genes with the highest. In agreement with findings made in unselected cells, RVISs present in the IDDb correlated with gene expression levels prior to transduction (Figure 6A-B). Comparing freshly isolated and cultured LSK cells, no major effect of culture conditions on the insertion profile was noted (Figure 6A-B). Interestingly, the association of RVISs with highly expressed genes tended to be more pronounced in class 1 than in classes 2 and 3 (Figure 6C).

Remarkably, the probability of forming a CRVIS does not seem to depend on the expression level. CRVISs are evenly distributed over all expression levels (Figure 6D) and even found in regions without transcriptional activity. A similar trend was observed for CRVISs that were hit 3 or more times; *Evi1*, the most frequent CRVIS in our dataset, does not show the highest expression level in the array (Figure 6D). Together, these data confirm the hypothesis that the risk of retroviral vector insertion in a given locus depends on its expression level in the target cell. However, the selection for

Table 5. Ingenuity pathway analysis of all genes listed in the IDDb (24 most significant results shown)

Pathway	Significance	Genes
p38 MAPK signaling	< .001	*Max, Faslg, Map3k5, Mef2d, Illr1, H3f3a, Ddit3*
Death receptor signaling	.001	*Faslg, Map3k5, Map3k14, Tnfsf10, Cflar*
PDGF signaling	.002	*Pdgfrb, Crk, Fos, Sos1, Plcg2*
Endoplasmic reticulum stress pathway	.002	*Ern1, Map3k5, Atf6*
PPAR signaling	.004	*Pdgfrb, Fos, Map3k14, Il1r1, Sos1*
IGF-1 signaling	.004	*Fos, Sos1, Foxo3a, Igfbp4, Akt1*
B-cell receptor signaling	.010	*Map3k5, Map3k14, Gab2, Sos1, Plcg2, Akt1*
PI3K/AKT signaling	.013	*Map3k5, RHEB, Gab2, Sos1, Akt1*
VEGF signaling	.020	*VEGF, Sos1, Plcg2, Akt1*
Neuregulin signaling	.021	*Crk, Sos1, Plcg2, Akt1*
PTEN signaling	.023	*Faslg, Sos1, Foxo3a, Akt1*
IL-6 signaling	.023	*Fos, Map3k14, Il1r1, Sos1*
Insulin receptor signaling	.023	*Crk, Sos1, Stxbp4, Foxo3a, Akt1*
Apoptosis signaling	.023	*Faslg, Map3k5, Map3k14, Plcg2*
IL-2 signaling	.026	*Fos, Sos1, Akt1*
Hypoxia signaling in the cardiovascular system	.032	*Vegf, Hic1, Akt1*
Neurotrophin/Trk signaling	.041	*Fos, Sos1, Akt1*
Natural killer cell signaling	.043	*Sos1, Klrb1c, Plcg2, Akt1*
ERK/MAPK signaling	.045	*Crk, Fos, Sos1, H3f3a, Plcg2*
FGF signaling	.048	*Crk, Map3k5, Sos1*
IL-10 signaling	.068	*Fos, Map3k14, Il1r1*
NF-κB signaling	.091	*Map3k14, Il1r1, Plcg2, Akt1*
SAPK/JNK signaling	.104	*Crk, Map3k5, Sos1*
Ephrin receptor signaling	.106	*Vegf, Crk, Sos1, Akt1*

For comparison, only 2 metabolic pathways were represented with more than one gene in the IDDb (purine metabolism, inositol phosphate metabolism).

From www.bloodjournal.org at MHH-Bibliothek on September 30, 2008. For personal use only.

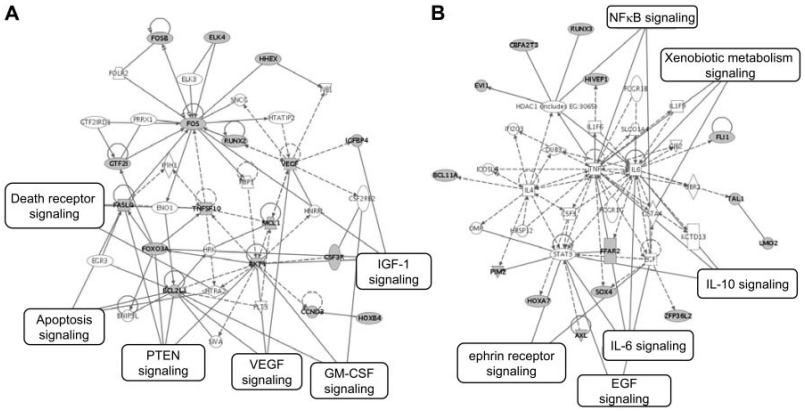


Figure 5. Ingenuity analyses of the genes listed in Table 3 reveal 2 major pathways. Note that further members of these pathways (A-B) may be highlighted when extending the analysis to the full IDDb. That is, *Siva* shown on the bottom of Figure 5A is a chromosomal neighbor of *Akt1*; this locus represents a CRVIS in the IDDb (Table 3; Table S1).

CRVIS is not a sole function of the initial expression level. We conclude that CRVISs are selected based on the biological consequences of target gene dysregulation and do not necessarily reflect a higher probability of retroviral integration.

Association of proto-oncogenes with leukemogenesis

To address whether the IDDb contains novel information regarding proto-oncogenes associated with leukemogenesis, we compared our data with tumor phenotypes listed in the RTCGD (obtained in animals infected with RCRs). The 2 databases are not redundant: Many CRVISs observed in the IDDb are not listed as CRVISs in the much larger RTCGD *(Igfbp4*, *Dph5*, *FasL*, *Gpr43*, *Gtf2i*, *Ly78*, *Lrcc6*, *Plcg2*, *Sesn2*, *Tnfsf10*, *Rab3gap2)*. These genes may be more likely to confer clonal dominance in healthy hematopoiesis than to contribute to malignant transformation. Furthermore, the IDDb shows some genes to have almost identical RVISs as listed in the RTCGD, however, frequently in association with distinct tumor phenotypes. The expansion and comparative analysis of these 2 databases may thus provide deeper insights into the association of genes with the induction of clonal dominance and malignant tumors.

Interestingly, the few RVISs identified to date that were associated with clonal dominance or malignant transformation in primate models and clinical trials are all found in the IDDb: *BCL-2A1* was identified as the RVIS in the single case of a malignant transformation observed to date in a nonhuman primate model following the use of replication-deficient retroviral vectors.[10] It is highly related to *BclX* and *Mcl1* listed in the IDDb. *LMO2* was found as a CRVIS in cases of lymphatic leukemia

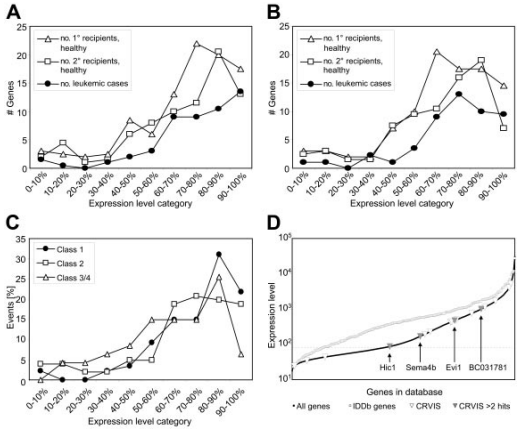


Figure 6. The probability of retroviral vector insertion but not the probability of forming a common insertion site depends on the expression level of the affected gene. (A) Array data from enriched hematopoietic progenitors containing both long-term and short-term repopulating cells (LSK cells, freshly isolated) were divided into 10 equal bins according to relative gene expression levels. The curves show the number of genes marked by RVISs in the different bins. Irrespective of the selection conditions (primary recipient, secondary recipient, or leukemia), the probability of RVIS is highest in the 40% most highly expressed genes. (B) Similar results were obtained when examining array data from LSK cells that were cultured for 2 days. (C) The selection for insertions in highly expressed genes is most pronounced for class 1 genes. (D) Expression levels of all genes detected by the arrays of LSK cells versus all RVIS genes of the IDDb, showing that the latter clearly have a much higher expression. The CRVIS genes of the IDDb are superimposed, showing that these do not cluster in the highest expression levels. Labeled genes represent CRVISs that were hit 3 times or more. Genes below the dotted line are not expressed in LSK cells.

From www.bloodjournal.org at MHH-Bibliothek on September 30, 2008. For personal use only.

occurring in gene therapy for X-linked severe combined immunodeficiency (SCID-X1) disease,[8] and the murine homolog is contained in the IDDb. Finally, the most frequent CRVIS in the IDDb found in association with both clonal dominance and myeloid transformation is *Evi1*; CRVISs in the human homolog were observed in association with clonal dominance in a recent report of patients undergoing retroviral vector-mediated gene therapy for chronic granulomatous disease (CGD).[9]

Discussion

The present study introduces a novel database (IDDb) listing RVISs associated with clonal dominance in cases of normal, potentially preleukemic hematopoiesis or malignant transformation of hematopoietic cells. We showed that our experimental conditions select RVISs of dominant clones that contribute the majority (> 50%) of a polyclonal population. Under our experimental conditions that involve a rather profound replication stress, the dysregulated cellular genes most likely have the potential to promote proliferation and/or survival of long-term repopulating hematopoietic cells. Consistent with present concepts of oncogenesis and leukemogenesis,[50,51] GO analysis revealed that 3 major gene functions contribute to clonal dominance: regulation of proliferation, apoptosis, and transcription. More importantly, pathway studies revealed that these genes are functionally connected in 2 major signaling networks (Figure 5). Of note, many of the genes listed in these classes have previously not been implicated in hematopoietic stem cell (HSC) biology.

Another important conclusion was that those genes which contribute to clonal dominance following insertional mutagenesis are more likely to be hit if already being transcriptionally active at a relatively high level in primitive hematopoietic cells. This was also observed with reference to the transcriptome of freshly isolated cells, independent of prior cytokine stimulation. The same conclusion was derived from an independent study performed with an even longer observation period (M.H.B., K.P., D.R., F.J.T.S., M. M. A., Verstegen, G.W., unpublished observations, July 2006). Interestingly, similar findings were made with retrovirally marked human cells observed in the nondiabetic obese (NOD)/SCID xenotransplant setting,[46] whereas insertion sites observed in murine tumors induced by RCR rather overlap with the transcriptome of human leukemias.[52] We would therefore assume that all vectors that show an insertion bias for expressed genes and contain strong enhancer sequences raise the probability of inducing clonal dominance by insertional mutagenesis. This also implies that the risk of clonal dominance or even malignant transformation should be much lower if gene transfer occurs in cells that have partially or completely silenced the self-renewal program.

The leukemias occurring in our model typically require combinatorial genetic lesions, either by the presence of RVISs in more than one leukemia-promoting gene,[5] or by a single proto-oncogenic RVIS in combination with signal alterations evoked by the vector-encoded transgene.[3] Although animals were examined for hematopoietic abnormalities in compliance with recommendations,[53,54] preleukemic clonal expansion might have been overlooked. Leukemogenic signal alterations are expected to be dose related, as previously observed for *Evi1* and *Hoxb4*.[55-57] The potential utility of these genes for stem cell expansion will thus depend on the ability to identify the required level of transcriptional dysregulation. Accordingly, we would assume that RVISs in such genes are only selected in vivo if the resulting extent of transcriptional dysregulation fits the selective pressure encountered in the given conditions. Insertional mutagenesis by RVISs may thus represent a powerful approach to identify genes that promote clonal survival under different selection conditions, such as exposure to cytotoxic drugs, inhibitory cytokines, irradiation, or disease-specific conditions.

Notably, not all IDDb entries can be considered as potential inducers of clonal dominance. Some genes may be accidentally marked in clones that contain more than one insertion, and intrinsic, potentially stochastic differences in cell fitness may also contribute to clonal dominance (reviewed in Spangrude et al[58]). A stronger focus on serially transplanted HSCs and experimental conditions favoring a single integration per cell may further increase the stringency of the screen. However, final proof requires functional studies. For a number of genes contained in the IDDb an essential role in the regulation of cellular survival is experimentally validated. This applies to the majority of the proto-oncogenes listed in Table 3 and the genes involved in the networks presented in Figure 5. However, only a smaller subset of these genes has previously been implicated to regulate "stemness." Examples are *Akt1*, which is known to be essential for self-renewal of murine embryonic stem cells,[59] *Hoxb4*, which stimulates HSC self-renewal without necessarily inducing leukemia,[60] and *Evi1*, which triggers self-renewal of myeloid progenitor cells in vitro and might give rise to a myelodysplastic syndrome and myeloid leukemia.[6,43] Interestingly, *Akt1* together with *Foxo3a* and *Cyclin D* regulates the hibernation of HSCs,[61] and all 3 genes are found in the center of the network shown in Figure 5A. Other genes that are functionally related to the last 2 examples are also found: The IDDb (Table S1) lists additional homeobox genes *(Hoxa7, Hhex, Cutl1, Dlx2, Dlx3)* and *Ski*, which is related to *Evi1* in its function to interact with SMAD signaling. An extended analysis of the IDDb also reveals further members of other pathways that are not (yet) recognized by the Ingenuity software tool.

Expanding the IDDb is also of major importance for the safety analysis of RVISs in preclinical and clinical studies. Although mice and humans differ in their susceptibility to transformation and some underlying mechanisms,[50] the IDDb nevertheless contains the 3 leading gene families associated with leukemia induction or preleukemic alterations observed to date in nonhuman primates and clinical trials: the *Bcl2*-related genes,[10] *Lmo2*,[8] and *Evi1*.[9] Expanding our approach to studies with other animal models might eventually even reveal basic biological principles regulating stem cell fitness that have been genetically and functionally conserved between different species. A general database for vector insertion sites that also includes data from clinical trials would be of great value.

Acknowledgments

We thank Manfred Schmidt and Christof von Kalle for their support in establishing LMPCR and for contributing insertion sites as published in references 3 and 5. We thank Anita Badbaran for technical assistance and Kristoffer Weber for help with the figures.

This work was supported by the Deutsche Forschungsgemeinschaft (grant DFG SPP1230) (B.F., Z.L. and C.B.) and (grant DFG-FE568/5-1,2) (B.F.), the European Union (grants INHERINET-QLK3-CT-2001-00427 and CONSERT-LSHB-CT-2004-005242) (G.W., F.J.T.S., and C.B.), and the National Cancer Institute (grant R01-CA107492-01A2) (C.B.).

From www.bloodjournal.org at MHH-Bibliothek on September 30, 2008. For personal use only.

Authorship

Contribution: O.S.K., G.v.K., and K.C. performed LMPCR, sequence analyses; O.S.K. organized the database and performed associated biostatistics; H.G., Z.L., K.J.N., and U.M. designed and performed animal experiments (including associated molecular biology); M.H.B., S.M.C., C.A.S., K.P.-O., D.d.R., F.J.T.S., G.W., and M.A.G. performed transcriptome studies and associated bioinformatics; and B.F. and C.B. initiated and coordinated the work, and wrote the paper together with the above colleagues.

Conflict of interest disclosure: The authors declare no competing financial interests.

Correspondence: Boris Fehse, Bone Marrow Transplantation, University Hospital Eppendorf, Martinistr. 52, 20251 Hamburg, Germany; e-mail: fehse@uke.uni-hamburg.de; and Christopher Baum, Experimental Hematology, OE6960, Hannover Medical School, Carl-Neuberg-Straße 1, 30625 Hannover, Germany; e-mail: baum.christopher@mh-hannover.de.

References

1. Mikkers H, Berns A. Retroviral insertional mutagenesis: tagging cancer pathways. Adv Cancer Res. 2003;88:53-99.
2. Akagi K, Suzuki T, Stephens RM, Jenkins NA, Copeland NG. RTCGD: retroviral tagged cancer gene database. Nucleic Acids Res. 2004;32: D523-D527.
3. Li Z, Dullmann J, Schiedlmeier B, et al. Murine leukemia induced by retroviral gene marking. Science. 2002;296:497.
4. Kustikova OS, Fehse B, Modlich U, et al. Clonal dominance of hematopoietic stem cells triggered by retroviral gene marking. Science. 2005;308: 1171-1174.
5. Modlich U, Kustikova O, Schmidt M, et al. Leukemias following retroviral transfer of multidrug resistance 1 are driven by combinatorial insertional mutagenesis. Blood. 2005;105:4235-4246.
6. Du Y, Jenkins NA, Copeland NG. Insertional mutagenesis identifies genes that promote the immortalization of primary bone marrow progenitor cells. Blood. 2005;106:3932-3939.
7. Du Y, Spence SE, Jenkins NA, Copeland NG. Cooperating cancer-gene identification through oncogenic-retrovirus-induced insertional mutagenesis. Blood. 2005;106:2498-2505.
8. Hacein-Bey-Abina S, Von Kalle C, Schmidt M, et al. LMO2-associated clonal T cell proliferation in two patients after gene therapy for SCID-X1. Science. 2003;302:415-419.
9. Ott MG, Schmidt M, Schwarzwaelder K, et al. Correction of X-linked chronic granulomatous disease by gene therapy, augmented by insertional activation of MDS1-EVI1, PRDM16 or SETBP1. Nat Med. 2006;12:401-409.
10. Seggewiss R, Pittaluga S, Adler RL, et al. Acute myeloid leukemia associated with retroviral gene transfer to hematopoietic progenitor cells of a rhesus macaque. Blood. 2006;107:3865-3867.
11. Calmels B, Ferguson C, Laukkanen MO, et al. Recurrent retroviral vector integration at the Mds1/Evi1 locus in nonhuman primate hematopoietic cells. Blood. 2005;106:2530-2533.
12. Uren AG, Kool J, Berns A, van Lohuizen M. Retroviral insertional mutagenesis: past, present and future. Oncogene. 2005;24:7656-7672.
13. Cornetta K, Morgan RA, Anderson WF. Safety issues related to retroviral-mediated gene transfer in humans. Hum Gene Ther. 1991;2:5-14.
14. Nolta JA, Dao MA, Wells S, Smogorzewska EM, Kohn DB. Transduction of pluripotent human hematopoietic stem cells demonstrated by clonal analysis after engraftment in immune-deficient mice. Proc Natl Acad Sci U S A. 1996;93:2414-2419.
15. Schmidt M, Hoffmann G, Wissler M, et al. Detection and direct genomic sequencing of multiple rare unknown flanking DNA in highly complex samples. Hum Gene Ther. 2001;12:743-749.
16. Schmidt M, Zickler P, Hoffmann G, et al. Polyclonal long-term repopulating stem cell clones in a primate model. Blood. 2002;100:2737-2743.
17. Waterston RH, Lindblad-Toh K, Birney E, et al. Initial sequencing and comparative analysis of the mouse genome. Nature. 2002;420:520-562.
18. Li Z, Fehse B, Schiedlmeier B, et al. Persisting multilineage transgene expression in the clonal progeny of a hematopoietic stem cell. Leukemia. 2002;16:1655-1663.
19. Abonour R, Williams DA, Einhorn L, et al. Efficient retrovirus-mediated transfer of the multidrug resistance 1 gene into autologous human long-term repopulating hematopoietic stem cells. Nat Med. 2000;6:652-658.
20. Hacein-Bey-Abina S, Le Deist F, Carlier F, et al. Sustained correction of X-linked severe combined immunodeficiency by ex vivo gene therapy. N Engl J Med. 2002;346:1185-1193.
21. Gaspar HB, Parsley KL, Howe S, et al. Gene therapy of X-linked severe combined immunodeficiency by use of a pseudotyped gammaretroviral vector. Lancet. 2004;364:2181-2187.
22. Aiuti A, Slavin S, Aker M, et al. Correction of ADA-SCID by stem cell gene therapy combined with nonmyeloablative conditioning. Science. 2002; 296:2410-2413.
23. Baum C, Dullmann J, Li Z, et al. Side effects of retroviral gene transfer into hematopoietic stem cells. Blood. 2003;101:2099-2114.
24. Baum C, Kustikova O, Modlich U, Li Z, Fehse B. Mutagenesis and oncogenesis by chromosomal insertion of gene transfer vectors. Hum Gene Ther. 2006;17:253-263.
25. Nienhuis AW, Dunbar CE, Sorrentino BP. Genotoxicity of retroviral integration in hematopoietic cells. Mol Ther. 2006;13:1031-1049.
26. Montini E, Cesana D, Schmidt M, et al. Hematopoietic stem cell gene transfer in a tumor-prone mouse model uncovers low genotoxicity of lentiviral vector integration. Nat Biotechnol. 2006;24: 687-696.
27. Recchia A, Bonini C, Magnani Z, 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 U S A. 2006;103:1457-1462.
28. Kustikova OS, Wahlers A, Kuhlcke K, et al. Dose finding with retroviral vectors: correlation of retroviral vector copy numbers in single cells with gene transfer efficiency in a cell population. Blood. 2003;102:3934-3937.
29. National Center for Biotechnology Information. BLAST: Basic Local Alignment Search Tool. http://www.ncbi.nlm.nih.gov/BLAST. Accessed October 2006.
30. European Bioinformatics Institute and Wellcome Trust Sanger Institute. Ensembl. http://www.ensembl.org. Accessed January 2006.
31. National Cancer Institute-Frederick. RTCGD: retrovirus tagged cancer gene database http://rtcgd.ncifcrf.gov. Accessed January 2006.
32. Lemischka IR, Moore KA, Stoeckert C. SCDb: stem cell database. http://stemcell.princeton.edu. Accessed January 2006.
33. Venezia TA, Merchant AA, Ramos CA, et al. Molecular signatures of proliferation and quiescence in hematopoietic stem cells. PLoS Biol. 2004;2: e301.
34. National Institute of Allergy and Infectious Diseases. EASE: Expression Analysis Systematic Explorer. http://david.niaid.nih.gov/david/ease.htm. Accessed July 2006.
35. Staal FJ, Weerkamp F, Baert MR, et al. Wnt target genes identified by DNA microarrays in immature CD34+ thymocytes regulate proliferation and cell adhesion. J Immunol. 2004;172:1099-1108.
36. de Ridder D, Staal FJ, van Dongen JJ, Reinders MJ. Maximum significance clustering of oligonucleotide microarrays. Bioinformatics. 2006;22: 326-331.
37. Dik WA, Pike-Overzet K, Weerkamp F, et al. New insights on human T cell development by quantitative T cell receptor gene rearrangement studies and gene expression profiling. J Exp Med. 2005; 201:1715-1723.
38. Armitage P, Berry G, Matthews JNS. Statistical methods in medical research. 4th ed. Oxford, United Kingdom: Blackwell Science Ltd; 2002.
39. Sullivan CS, Pipas JM. T antigens of simian virus 40: molecular chaperones for viral replication and tumorigenesis. Microbiol Mol Biol Rev. 2002;66: 179-202.
40. Wu X, Li Y, Crise B, Burgess SM. Transcription start regions in the human genome are favored targets for MLV integration. Science. 2003;300: 1749-1751.
41. Trobridge GD, Miller DG, Jacobs MA, et al. Foamy virus vector integration sites in normal human cells. Proc Natl Acad Sci U S A. 2006;103: 1498-1503.
42. Modlich U, Bohne J, Schmidt M, et al. Cell culture assays reveal the importance of retroviral vector design for insertional genotoxicity. Blood. 2006; 108:2545-2553.
43. Nucifora G, Laricchia-Robbio L, Senyuk V. EVI1 and hematopoietic disorders: history and perspectives. Gene. 2006;368:1-11.
44. 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.
45. Maxfield LF, Fraize CD, Coffin JM. Relationship between retroviral DNA-integration-site selection and host cell transcription. Proc Natl Acad Sci U S A. 2005;102:1436-1441.
46. Wagner W, Laufs S, Blake J, et al. Retroviral integration sites correlate with expressed genes in hematopoietic stem cells. Stem Cells. 2005;23: 1050-1058.
47. Camargo FD, Chambers SM, Drew E, McNagny KM, Goodell MA. Hematopoietic stem cells do not engraft with absolute efficiencies. Blood. 2006; 107:501-507.
48. Adolfsson J, Mansson R, Buza-Vidas N, et al. Identification of Flt3+ lympho-myeloid stem cells lacking erythro-megakaryocytic potential a revised road map for adult blood lineage commitment. Cell. 2005;121:295-306.

From www.bloodjournal.org at MHH-Bibliothek on September 30, 2008. For personal use only.

49. Krivtsov A, Twomey D, Feng Z, et al. Transformation from committed progenitor to leukaemia stem cell initiated by MLL-AF9. Nature. Nature. 2006; 442:818-822.
50. Hahn WC, Weinberg RA. Rules for making human tumor cells. N Engl J Med. 2002;347:1593-1603.
51. Gilliland D, Tallman M. Focus on acute leukemias. Cancer Cell. 2002;1:417-420.
52. Erkeland SJ, Verhaak RG, Valk PJ, Delwel R, Lowenberg B, Touw IP. Significance of murine retroviral mutagenesis for identification of disease genes in human acute myeloid leukemia. Cancer Res. 2006;66:622-626.
53. Kogan SC, Ward JM, Anver MR, et al. Bethesda proposals for classification of nonlymphoid hematopoietic neoplasms in mice. Blood. 2002;100: 238-245.
54. Morse HC, Anver MR, Fredrickson TN, et al. Bethesda proposals for classification of lymphoid neoplasms in mice. Blood. 2002;100:246-258.
55. Boyd KE, Xiao YY, Fan K, et al. Sox4 cooperates with Evi1 in AKXD-23 myeloid tumors via transactivation of proviral LTR. Blood. 2006;107:733-741.
56. Schiedlmeier B, Klump H, Will E, et al. High-level ectopic HOXB4 expression confers a profound in vivo competitive growth advantage on human cord blood CD34+ cells, but impairs lymphomyeloid differentiation. Blood. 2003;101:1759-1768.
57. Brun AC, Bjornsson JM, Magnusson M, et al. Hoxb4-deficient mice undergo normal hematopoietic development but exhibit a mild proliferation defect in hematopoietic stem cells. Blood. 2004; 103:4126-4133.
58. Spangrude GJ, Smith L, Uchida N, et al. Mouse hematopoietic stem cells. Blood. 1991;78:1395-1402.
59. Watanabe S, Umehara H, Murayama K, Okabe M, Kimura T, Nakano T. Activation of Akt signaling is sufficient to maintain pluripotency in mouse and primate embryonic stem cells. Oncogene. 2006;25:2697-2707.
60. Buske C, Humphries RK. Homeobox genes in leukemogenesis. Int J Hematol. 2000;71:301-308.
61. Yamazaki S, Iwama A, Takayanagi S, et al. Cytokine signals modulated via lipid rafts mimic niche signals and induce hibernation in hematopoietic stem cells. EMBO J. 2006;25:3515-3523.
62. Ivanova NB, Dimos JT, Schaniel C, Hackney JA, Moore KA, Lemischka IR. A stem cell molecular signature. Science. 2002;298:601-604.