

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}

¹Department of Experimental Hematology, Hannover Medical School, Germany; ²Bone Marrow Transplantation, University Medical Center Hamburg-Eppendorf, Hamburg, Germany; ³Division of Experimental Hematology, Cincinnati Children's Hospital Medical Center, OH; ⁴Department of Hematology, Erasmus University Medical Center, Rotterdam, The Netherlands; ⁵Center for Cell and Gene Therapy and Cell and Molecular Biology Program, Baylor College of Medicine, Houston, TX; ⁶Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX; ⁷Department of Immunology, Erasmus Medical Center, Rotterdam, The Netherlands; ⁸Information 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 On-

tology 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 infection⁹ 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

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 cytopins 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: BLAST²⁹ searches were performed or, in some cases at Ensemble³⁰; 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-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.

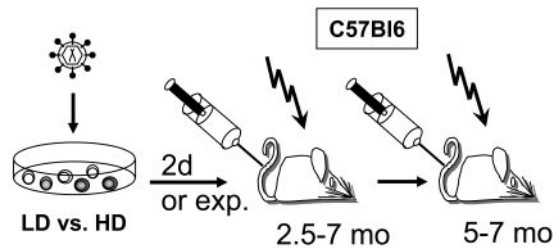


Figure 1. Experimental setup of murine BMT studies using donor cells modified with different retroviral vectors. The enhancer-promoter contained in the long terminal repeat (LTR), the cDNA encoded by the vector, and the 3' untranslated region (3' UTR) are indicated in Table 2. LD indicates low dose of 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
SF91IRESEGF	SFFVp	IRESE-EGFP	wPRE
SF91XRCC4	SFFVp	XRCC4-IRES-EGFP	wPRE
SF11flCD34	SFFVp	flCD34	—
SF11tCD34	SFFVp	tCD34	—
SF11mflCD34	SFFVp	mflCD34	—
SF11mtCD34	SFFVp	mtCD34	—
SF11dLNGFR	SFFVp	dLNGFR	—
HaMDR1	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.

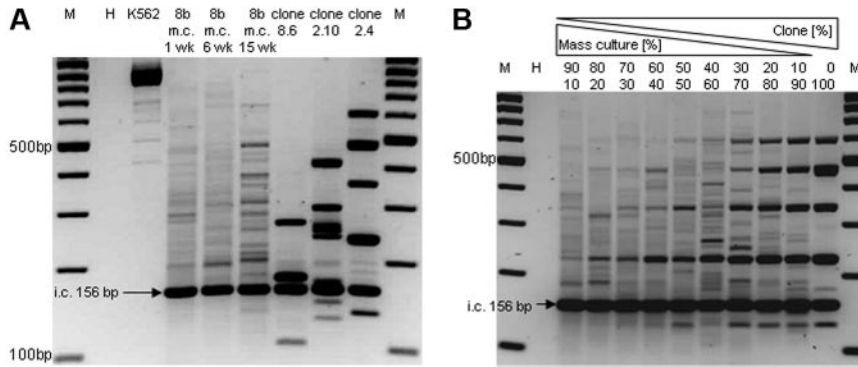


Figure 2. LMPCR validation. (A) DNA of K562 mass cultures and cell clones containing different numbers of retroviral insertions²⁸ was subjected to insertion site amplification by LMPCR using the conditions described in "Material and methods." In contrast to the clonal DNA, mass culture DNA does not reveal 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 RTCGD² 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 fCD34, murine tCD34, murine fCD34, 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

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 *Csf3r* 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.

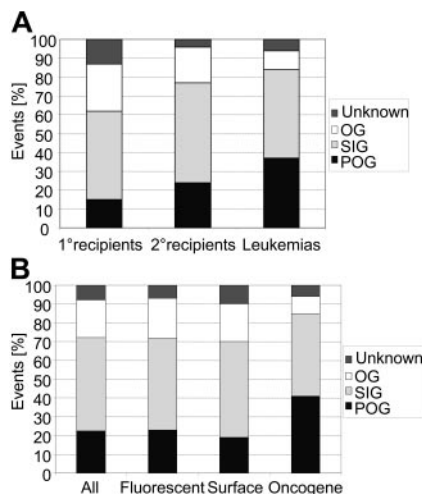


Figure 3. Retroviral vector insertion site (RVIS) distribution according to gene classes and type of transgene. (A) RVISs in known proto-oncogenes (POGs) increase in frequency over serial BMT and are most pronounced in leukemic clones. (B) No major impact of the transgene class was found except when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

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

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					
<i>FasL</i>	2	0	14103	1 H2.1	Fas ligand (TNF superfamily, member 6)
<i>Vegfa</i>	1	0	22339	17B3	Vascular endothelial growth factor A
<i>Tnfsf10</i>	2	2	22035	3A3	Tumor necrosis factor (ligand) superfamily, member 10, apoptosis induction
Receptor					
<i>Sema4b</i>	3	4	20352	7D1	Receptor activity
<i>Axl</i>	1	0	26362	7A3-B1	Receptor activity, protein kinase activity, human proto-oncogene
<i>Csflr</i>	1	9	12978	18D	Receptor of colony stimulating factor 1 (CSF-1)
<i>Csfr3</i>	1	0	12986	4D2+2	G protein coupled receptor for granulocyte colony stimulating factor (G-CSF)
<i>Gpr43 = Ffar2</i>	2	0	233079	7A3	G-protein coupled receptor, free fatty acid receptor
<i>Igfbp4</i>	2	0	16010	11 D-E1	Insulin-like growth factor binding protein 4
<i>Ly78</i>	2	0	17079	13D1	Lymphocyte antigen 78, receptor, signaling, CD180
Signal transducer					
<i>Akt1</i>	2	3	11651	12F1-F2	Intracellular signaling, kinase transforming protein
<i>Bcl211</i>	1	3	12048	2H1	Bcl2-like 1, antiapoptosis
<i>Ccnd3</i>	1	13	12445	17B4	Cyclin D3, cell cycle
<i>Mcl1</i>	1	4	17210	3F2.1	Myeloid cell leukemia sequence 1, Bcl2-related antiapoptotic protein
<i>Osbpl3</i>	1	2	71720	6B3	Oxysterol binding protein-like 3, steroid metabolism
<i>Pim2</i>	2	5	18715	X A1.1	Serine/threonine-protein kinase Pim-2, proviral integration site 2, antiapoptosis
<i>Plcg2</i>	2	0	234779	8E1	Phospholipase C, gamma 2, survival signaling
<i>Rab3gap2</i>	2	0	381313	1H5	Similar to Rab3 GTPase activating protein, gene model 981
<i>Rhof</i>	1	4	23912	5F	Ras homolog gene family member
<i>Sesn2 = Hi95</i>	3	0	230784	4D2.3	Sestrin 2, induction in response to DNA damage
Transcription factor					
<i>2610510B01Rik = Dopey2</i>	1	4	70028	16 C4	Predicted leucine zipper transcription factor, dopey family
<i>Bcl11a</i>	2	6	14025	11A3+2	Zinc finger, essential for lymphopoiesis
<i>Cbfa2t3h</i>	1	7	12398	8E1	Core-binding factor, runt domain, alpha subunit 2, translocated 2, 3 homolog
<i>Cutl1</i>	1	3	13047	5 G2	Cut-like 1 (Drosophila) = CCAAT displacement protein = Cux/CDP homeoprotein
<i>Elk4</i>	1	3	13714	1E3-G	ELK4, ETS family member
<i>Evi1</i>	13	20	14013	3A3	Associated with murine and human leukemia, SMAD interacting
<i>Fli1</i>	2	7	14247	9A4	ETS family member
<i>Fos (LOC627366)</i>	2	5	14281	12 D2	FBJ osteosarcoma oncogene
<i>Fosb (Ercc1)</i>	1	0	14282	7A2	FBJ osteosarcoma oncogene B
<i>Foxo3a</i>	2	0	56484	10B2	Forkhead transcription factor, potentially pro-apoptotic
<i>Gtf2i</i>	2	0	14886	5G2	General transcription factor 2
<i>Hhex</i>	1	26	15242	19C1	Hematopoietically expressed homeobox gene (T-cell oncogene)
<i>Hic1</i>	3	3	15248	11B5	Hypermethylated in cancer 1, transcription factor, Wnt antagonism
<i>Hivep1</i>	1	2	110521	13A4	HIV enhancer binding protein 1
<i>HoxA7</i>	2	19	15404	6B3	Homeobox gene A7 and surrounding cluster
<i>Hoxb4/Hoxb5</i>	1	0	15412	11D	Homeobox gene B4 and surrounding cluster
<i>Lmo2</i>	2	2	16909	2 E2	LIM domain only 2, T-cell oncogene
<i>Mef2d</i>	1	5	17261	3F1	Myocyte enhancer factor 2D
<i>Mlit3</i>	1	0	70122	4C4	Involved in leukemogenic translocations
<i>Runx2</i>	1	6	12393	17B3	Runt related, essential for hematopoiesis
<i>Runx3</i>	1	5	12399	4D2.3	Runt related, myeloid development
<i>Sox4</i>	1	64	20677	13A3-A5	Sry box, lymphocyte activation
<i>Tal1</i>	1	0	21349	4D1	Ebox family member, essential for hematopoiesis
<i>Zfp3612</i>	1	4	12193	17 E4	Zinc finger protein 36, C3H type-like
Metabolic					
<i>Dph5</i>	2	0	69740	3F3	DPH5 homolog (<i>Saccharomyces cerevisiae</i>) transferase
Unknown					
<i>Lrrc6</i>	2	0	54562	15D1	Leucin repeat containing 6 (testis)
<i>Dym = 4933427L07Rik</i>	1	3	69190	18 E2	Dymeclin, function unknown
<i>AB041803</i>	2	8	232685	6A3.3	Hypothetical protein, function unknown
<i>BC031781</i>	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.ncicrf.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 \pm 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 al⁴¹) are shown in Figure 4A

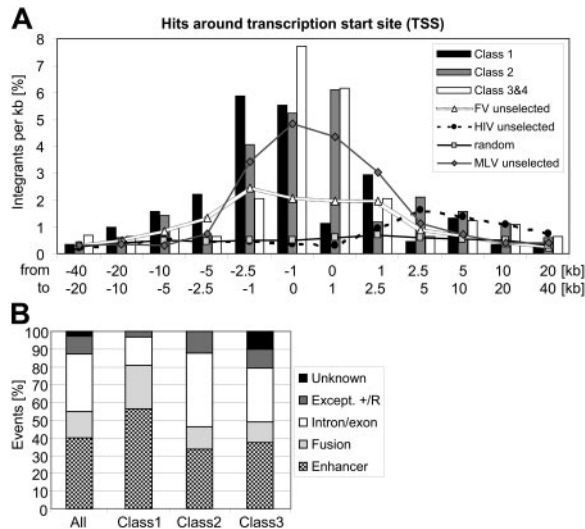


Figure 4. Type of mutations. Data are shown with respect to gene class 1 (common insertion sites, proto-oncogenes, and self-renewal genes), class 2 (signaling genes), and classes 3 and 4 (other and unknown genes). (A) Position of RVIS in the Insertional Dominance Database (IDDb) around the transcriptional 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 insertions proposed earlier.¹² Insertions located downstream but in an antisense orientation do not correspond to the definition of enhancer mutations suggested in Uren et al¹² 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 vectors⁴¹ 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 follows¹²: 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 ($\text{Lin}^- \text{Sca1}^+ \text{c-Kit}^+$, 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,

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*, *Pen*, 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 translocations⁴⁹ (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	<i>Max, Faslg, Map3k5, Mef2d, Il1r1, H3f3a, Ddit3</i>
Death receptor signaling	.001	<i>Faslg, Map3k5, Map3k14, Tnfrsf10, Cflar</i>
PDGF signaling	.002	<i>Pdgfrb, Crk, Fos, Sos1, Plcg2</i>
Endoplasmic reticulum stress pathway	.002	<i>Ern1, Map3k5, Atf6</i>
PPAR signaling	.004	<i>Pdgfrb, Fos, Map3k14, Il1r1, Sos1</i>
IGF-1 signaling	.004	<i>Fos, Sos1, Foxo3a, Igbp4, Akt1</i>
B-cell receptor signaling	.010	<i>Map3k5, Map3k14, Gab2, Sos1, Plcg2, Akt1</i>
PI3K/AKT signaling	.013	<i>Map3k5, RHEB, Gab2, Sos1, Akt1</i>
VEGF signaling	.020	<i>VEGF, Sos1, Plcg2, Akt1</i>
Neuregulin signaling	.021	<i>Crk, Sos1, Plcg2, Akt1</i>
PTEN signaling	.023	<i>Faslg, Sos1, Foxo3a, Akt1</i>
IL-6 signaling	.023	<i>Fos, Map3k14, Il1r1, Sos1</i>
Insulin receptor signaling	.023	<i>Crk, Sos1, Stxbp4, Foxo3a, Akt1</i>
Apoptosis signaling	.023	<i>Faslg, Map3k5, Map3k14, Plcg2</i>
IL-2 signaling	.026	<i>Fos, Sos1, Akt1</i>
Hypoxia signaling in the cardiovascular system	.032	<i>Vegf, Hic1, Akt1</i>
Neurotrophin/Trk signaling	.041	<i>Fos, Sos1, Akt1</i>
Natural killer cell signaling	.043	<i>Sos1, Klrb1c, Plcg2, Akt1</i>
ERK/MAPK signaling	.045	<i>Crk, Fos, Sos1, H3f3a, Plcg2</i>
FGF signaling	.048	<i>Crk, Map3k5, Sos1</i>
IL-10 signaling	.068	<i>Fos, Map3k14, Il1r1</i>
NF- κ B signaling	.091	<i>Map3k14, Il1r1, Plcg2, Akt1</i>
SAPK/JNK signaling	.104	<i>Crk, Map3k5, Sos1</i>
Ephrin receptor signaling	.106	<i>Vegf, Crk, Sos1, Akt1</i>

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

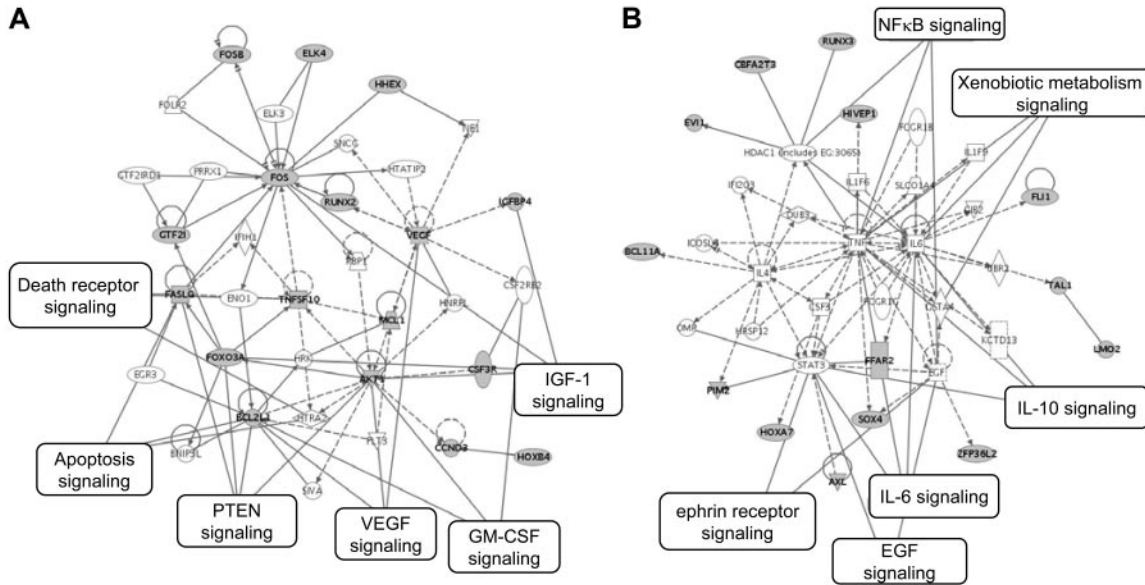


Figure 5. Ingenuity analyses of the genes listed in Table 3 reveal 2 major pathways. Note that further members of these pathways (A-B) may be highlighted when extending the analysis to the full IDDb. That is, *Siva* shown on the bottom of Figure 5A is a chromosomal neighbor of *Akt1*; this locus represents 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 (*Igf1p4*, *Dph5*, *FasL*, *Gpr43*, *Gtf2i*, *Ly78*, *Lrcc6*, *Plcg2*, *Sesn2*, *Tnfrsf10*, *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

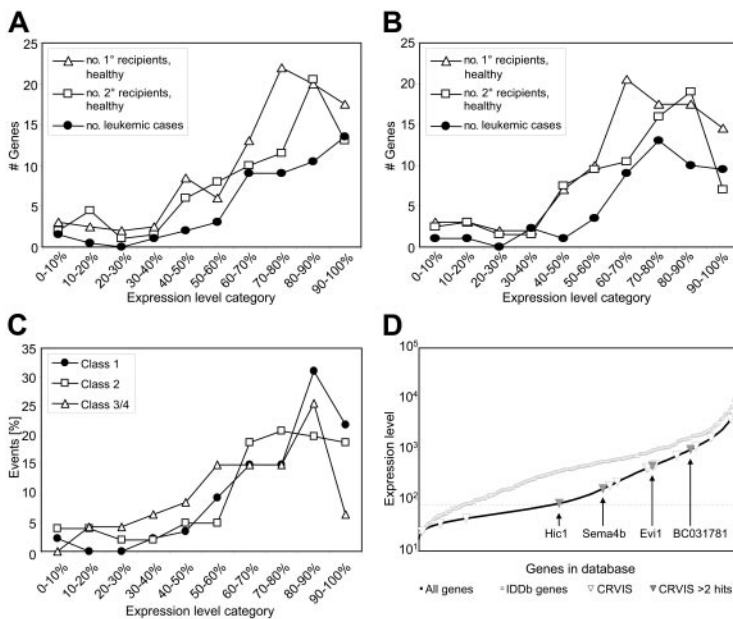


Figure 6. The probability of retroviral vector insertion but not the probability of forming a common insertion site depends on the expression level of the affected gene. (A) Array data from enriched hematopoietic progenitors containing both long-term and short-term repopulating cells (LSK cells, freshly 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.

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 transcrip-

tional 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 al⁵⁸). 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 INHERIT-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.).

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

- Mikkers H, Berns A. Retroviral insertional mutagenesis: tagging cancer pathways. *Adv Cancer Res.* 2003;88:53-99.
- Akagi K, Suzuki T, Stephens RM, Jenkins NA, Copeland NG. RTCGD: retroviral tagged cancer gene database. *Nucleic Acids Res.* 2004;32:D523-D527.
- Li Z, Dullmann J, Schiedlmeier B, et al. Murine leukemia induced by retroviral gene marking. *Science.* 2002;296:497.
- Kustikova OS, Fehse B, Modlich U, et al. Clonal dominance of hematopoietic stem cells triggered by retroviral gene marking. *Science.* 2005;308:1171-1174.
- 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.
- 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.
- Du Y, Spence SE, Jenkins NA, Copeland NG. Cooperating cancer-gene identification through oncogenic-retrovirus-induced insertional mutagenesis. *Blood.* 2005;106:2498-2505.
- 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.
- 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.
- 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.
- 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.
- Uren AG, Kool J, Berns A, van Lohuizen M. Retroviral insertional mutagenesis: past, present and future. *Oncogene.* 2005;24:7656-7672.
- Cornetta K, Morgan RA, Anderson WF. Safety issues related to retroviral-mediated gene transfer in humans. *Hum Gene Ther.* 1991;2:5-14.
- 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.
- 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.
- Schmidt M, Zickler P, Hoffmann G, et al. Polyclonal long-term repopulating stem cell clones in a primate model. *Blood.* 2002;100:2737-2743.
- Waterston RH, Lindblad-Toh K, Birney E, et al. Initial sequencing and comparative analysis of the mouse genome. *Nature.* 2002;420:520-562.
- 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.
- 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.
- 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.
- 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.
- 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.
- Baum C, Dullmann J, Li Z, et al. Side effects of retroviral gene transfer into hematopoietic stem cells. *Blood.* 2003;101:2099-2114.
- 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.
- Nienhuis AW, Dunbar CE, Sorrentino BP. Genotoxicity of retroviral integration in hematopoietic cells. *Mol Ther.* 2006;13:1031-1049.
- 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.
- 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.
- 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.
- National Center for Biotechnology Information. BLAST: Basic Local Alignment Search Tool. <http://www.ncbi.nlm.nih.gov/BLAST>. Accessed October 2006.
- European Bioinformatics Institute and Wellcome Trust Sanger Institute. Ensembl. <http://www.ensembl.org>. Accessed January 2006.
- National Cancer Institute-Frederick. RTCGD: retrovirus tagged cancer gene database. <http://rtcgd.ncifcrf.gov>. Accessed January 2006.
- Lemischka IR, Moore KA, Stoeckert C. SCDB: stem cell database. <http://stemcell.princeton.edu>. Accessed January 2006.
- Venezia TA, Merchant AA, Ramos CA, et al. Molecular signatures of proliferation and quiescence in hematopoietic stem cells. *PLoS Biol.* 2004;2:e301.
- National Institute of Allergy and Infectious Diseases. EASE: Expression Analysis Systematic Explorer. <http://david.niaid.nih.gov/david/ease.htm>. Accessed July 2006.
- 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.
- de Ridder D, Staal FJ, van Dongen JJ, Reinders MJ. Maximum significance clustering of oligonucleotide microarrays. *Bioinformatics.* 2006;22:326-331.
- 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.
- Armitage P, Berry G, Matthews JNS. *Statistical methods in medical research.* 4th ed. Oxford, United Kingdom: Blackwell Science Ltd; 2002.
- 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.
- 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.
- 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.
- 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.
- Nucifora G, Laricchia-Robbio L, Senyuk V. EVI1 and hematopoietic disorders: history and perspectives. *Gene.* 2006;368:1-11.
- 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.
- 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.
- 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.
- Camargo FD, Chambers SM, Drew E, McNagny KM, Goodell MA. Hematopoietic stem cells do not engraft with absolute efficiencies. *Blood.* 2006;107:501-507.
- 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.

49. Krivtsov A, Twomey D, Feng Z, et al. Transformation from committed progenitor to leukaemia stem cell initiated by MLL-AF9. *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. Schiedmeier 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.