Target Site Selection of Retroviral Vectors in the Human Genome: Viral and Genomic Determinants of Non-Random Integration Patterns in Hematopoietic Cells

Thesis

How to cite:

Cattoglio, Claudia (2008). Target Site Selection of Retroviral Vectors in the Human Genome: Viral and Genomic Determinants of Non-Random Integration Patterns in Hematopoietic Cells. PhD thesis The Open University.

UNRESTRICTED

CLAUDIA CATTOGLIO

TARGET SITE SELECTION OF RETROVIRAL VECTORS IN THE HUMAN GENOME: VIRAL AND GENOMIC DETERMINANTS OF NON-RANDOM INTEGRATION PATTERNS IN HEMATOPOIETIC CELLS

PhD thesis in fulfillment of the requirements of the Open University for the degree of Doctor of Philosophy in Molecular and Cellular Biology

2008

Director of studies	External Supervisor
Prof. Fulvio Mavilio	Prof. Andrew M. L. Lever

Vita-Salute San Raffaele University
Department of Biological and Technological Research
San Raffaele Scientific Institute
Milan, Italy

Submission date: 19 August 2008
Date of award: 19 Dec. 2008

ABSTRACT

Integration of gamma-retroviruses (RV) and lentiviruses (LV) follows different, non-random patterns in mammalian genomes. To obtain information about the viral and genomic determinants of integration preferences, I mapped > 2,500 integration sites of RV and LV vectors carrying wild type or modified LTRs in human CD34$^+$ hematopoietic cells. Recurrent insertion sites (hot spots) account for > 20% of the RV integration events, while they are significantly less frequent for LV vectors. Genes controlling growth, differentiation and development of the hematopoietic and immune system are targeted at high frequency by RV vectors and further enriched in RV hot spots, suggesting that the cell gene expression program is instrumental in directing RV integration. To investigate the role of transcriptional regulatory networks in directing RV and LV integration, I evaluated the local abundance and arrangement of putative transcription factor binding sites (TFBSs) in the genomic regions flanking integrated proviruses. RV, but not LV integrations are flanked by specific subsets of TFBSs, independently of their location with respect to genes (within genes, outside, or around their transcription start sites). Hierarchical clustering and a Principal Components Analysis of TFBSs flanking integration sites of RV vectors carrying different LTRs, and LV vectors packaged with wild type or RV-LV hybrid integrase, showed that both the protein and the DNA component of the pre-integration complex (PIC) have a causal role in directing proviral integration in TFBS-rich regions of the genome. Chromatin immunoprecipitation analysis indicated that TFs are bound to unintegrated LTR enhancers into the nucleus, and might synergize with the viral integrase in tethering retroviral PICs to specific domains of transcriptionally active chromatin. The results of this project suggest

substantial differences in the molecular mechanisms tethering RV and LV PICs to human chromatin, and predict a different insertional oncogenesis risk of RV *vs.* LV vectors for human gene therapy.

TABLE OF CONTENTS

1. Introduction

The transfer of a therapeutic gene into somatic cells (gene therapy) is a promising medical approach for the management of many inherited and acquired diseases. Among others, blood disorders are of special interest for gene therapy interventions, thanks to the easy accessibility and hierarchical structure of the blood system, with a relatively limited number of hematopoietic stem cells (HSCs) giving origin to all lineages of differentiated blood cells. Moreover, modification of a small number of long-term repopulating stem cells is often sufficient to achieve therapeutic efficacy in the entire system. Among several strategies developed for gene delivery, replication-defective viral vectors derived from retroviruses, especially from gamma-retroviruses (RV) and lentiviruses (LV), are the most widely used. In fact, after entering the target cell, retroviral vectors deliver their genomic material directly to the cell nucleus, where it is stably inserted into the host cell genome by the virally encoded integrase protein. Gene-transfer vectors derived from murine RV, such as the Moloney murine leukemia virus (Mo-MLV), have been extensively used to transduce human HSCs in several gene therapy clinical trials, in some cases allowing correction of life-threatening blood disorders[1-4]. These vectors were considered relatively safe, until lymphoproliferative disorders were reported in five patients treated with MLV-transduced HSCs for X-linked severe combined immunodeficiency (X-SCID)[5-8]. All adverse events were correlated with the insertional activation of T-cell proto-oncogenes operated by MLV long terminal repeats (LTRs). The oncogenic potential of murine RV has been known for decades, but the risk of insertional mutagenesis by retroviral vectors was estimated to be low, on the assumption of random proviral integration into the genome. The X-SCID adverse outcomes indicated the importance of understanding the molecular basis of retroviral integration and boosted a series of large-scale insertion studies aimed at evaluating genotoxic risks and general integration preferences of both RV and LV vectors. From these studies it became clear that RV and LV vectors integrate non-randomly in mammalian genomes, with a strong preference for active and gene-dense chromatin regions. In particular, RV, but not LV vectors preferentially target gene transcription start sites (TSSs) and CpG islands[9-12], where the insertion of viral enhancers contained in the LTRs has a high probability to interfere with gene regulation[13]. Moreover, analysis of RV integration patterns in natural or experimentally induced hematopoietic tumors showed the existence of insertion sites recurrently associated with a malignant phenotype. These common insertion sites (CISs) include proto-oncogenes and other genes controlling cell growth and proliferation, deregulation of which has a causal relationship with neoplastic transformation[14]. Some of these CISs have been also retrieved at high frequency in the nonmalignant progeny of transduced HSCs in mice[15], non-human primates[16], and humans[4], suggesting that insertion into certain genes may cause clonal amplification of transduced progenitors *in vivo*. However, pretransplant, unselected HSCs were never rigorously analyzed short-term after RV transduction, leaving the possibility that the clonal dominance observed *in vivo* is favored by the existence of highly preferred regions of retroviral integration that make clonal amplification more likely to occur.

This thesis analyzes thoroughly a large collection of RV and LV integration sites retrieved from human CD34$^+$ HSCs at an early time point after infection, when clonal selection in culture is very unlikely to have occurred. The general goal of the project was to describe the integration preferences of RV and LV vectors in the genome of clinically relevant cells, and possibly provide new insights into the molecular mechanisms responsible for their differential integration targeting. I found that a large proportion (21%) of RV insertion sites are clustered in hot spots, targeting genes involved in the control of growth, differentiation and development of hematopoietic cells, including several CISs. On the contrary, only 8% of LV integration sites formed hot spots, with no apparent bias for hematopoietic-specific genes. This suggested that the gene expression program of HSCs is somewhow involved in directing RV preintegration complexes (PICs) to preferred sites in the genome. To further investigate the link between transcription and RV integration, I evaluated the local abundance and arrangement of putative transcription factor binding sites (TFBSs) in the genomic regions flanking integrated RV and LV proviruses. Again RV, but not LV vectors favor genomic regions flanked by specific subsets of TFBSs, independently of their location with respect to genes or TSSs. Analysis of RV and LV mutants showed that the MLV LTR enhancer, together with the MLV integrase, has a causal role in directing proviral integration towards TFBS-rich regions of the genome. Chromatin immunoprecipitation assays indicated that cellular TFs binding unintegrated LTR enhancers in the nucleus might synergize with the integrase in tethering RV PICs to specific domains of transcriptionally active chromatin.

Providing evidence that RV vectors tend to target hot spots of integration in the proximity of regulatory regions and of genes controlling cell growth and proliferation, this thesis also predicts a higher genotoxic risk in using gamma-retroviral *vs.* lentiviral vectors for human gene therapy applications.

2. A brief review of retroviruses and retroviral vectors

2.1 Retroviruses

2.1.1 Structure and classification

Retroviruses comprise a large and diverse family of enveloped RNA viruses, replicating through a DNA intermediate[17]. The virions (80-100 nm in diameter) consist of an outer lipid envelope and of an internal protein core, accommodating two identical copies of single-stranded viral RNA genome (**Figure 1**). The lipid bilayer derives from the plasma membrane of infected cells into which virally encoded envelope proteins are inserted (transmembrane and surface components, linked by disulphide bonds). The internal core is composed of non-glycosylated structural proteins (matrix, capsid and nucleocapsid) and contains the full-length viral genome and the virally encoded enzymes (reverse transcriptase, integrase and protease).

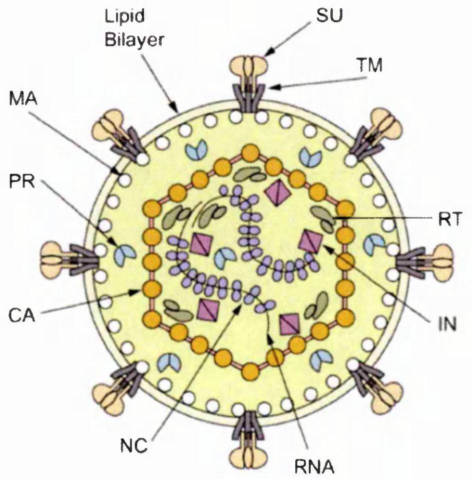


Figure 1. General structure of a retrovirus. Envelope proteins, consisting of a transmembrane (TM) and a surface (SU) component, are inserted in a lipid bilayer, where they bind host cell receptors to promote viral entry. Each virion contains two copies of full-length RNA viral genome, embedded in the protein core encoded by *gag* domain (NC, nucleocapsid; CA, capsid; MA, matrix). Functional proteins of viral origin indispensable for replication (reverse transcriptase, integrase and protease) are also encapsidated. Shapes are merely representative and do not necessarily reflect the real geometry.

Depending on their genomic structure, retroviruses are broadly divided into two categories: simple and complex (**Figure 2**); both simple and complex genomes contain four elementary genes (*gag*, *pro*, *pol*, *env*) coding for the essential viral proteins, as follows:

gag: matrix, capsid, nucleocapsid

pro: protease

pol: reverse transcriptase, integrase

env: surface and transmembrane components of the envelope protein

In addition to the basic coding domains, complex retroviruses also encode several accessory genes, derived from multiple splicing (**Figure 2B**). Both ends of all viral genomes contain terminal noncoding sequences, composed of 5' and 3' unique sequences (U5 and U3 regions), and of two direct repeats (R) where the transcription start site and the polyadenylation signal are located.

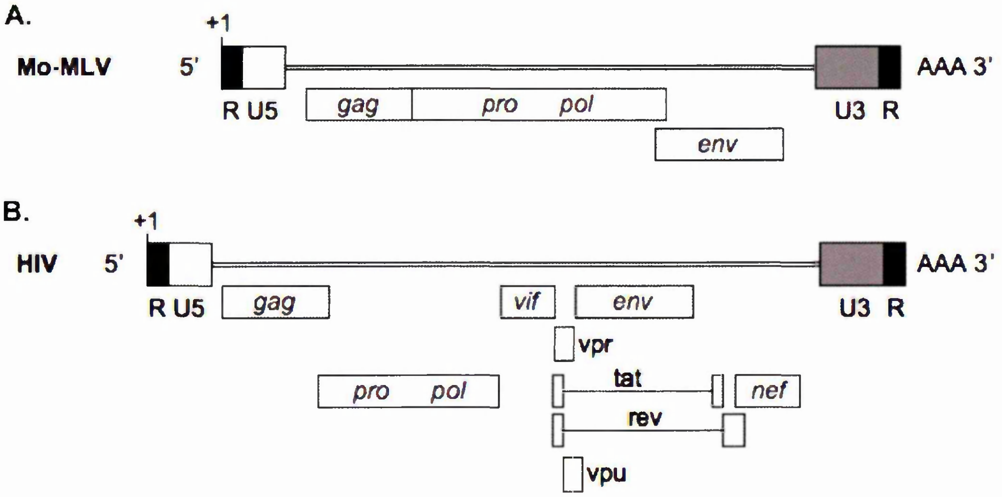


Figure 2. Simple and complex retroviral genomes. (A) Moloney murine leukemia virus (Mo-MLV) genomic RNA is only made up of four elementary coding regions, *gag*, *pro*, *pol*, and *env*. Terminal, noncoding R, U5 and U3 regions are depicted; transcription start site (+1) and polyadenylation signal (AAA) are specified. **(B)** Human immunodeficiency virus (HIV) is a complex retrovirus, with six accessory, partially overlapping genes (*vif*, *vpr*, *tat*, *rev*, *vpu*, and *nef*) in addition to the four basic coding domains.

Based on evolutionary relatedness, retroviruses are further classified into seven *genera* (**Figure 3**). Moloney murine leukemia virus (Mo-MLV) and human immunodeficiency virus-type 1 (HIV-1), object of this thesis, belong to the *genus* of gamma-retroviruses (also known as oncoretroviruses) and of lentiviruses, respectively.

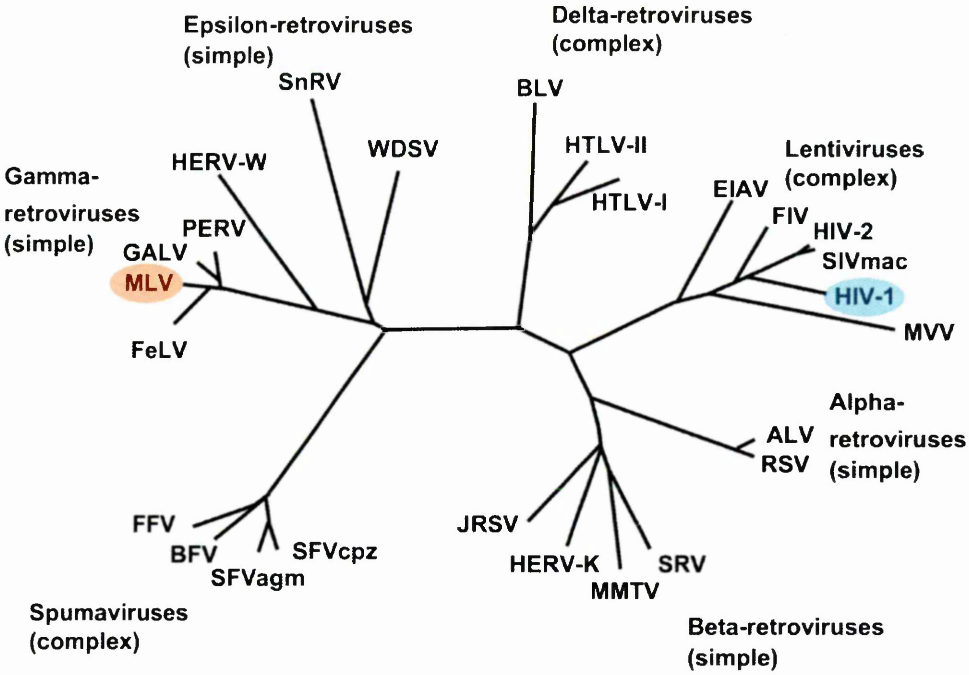


Figure 3. Phylogeny of retroviruses. Classification of complex and simple retroviruses into seven *genera*. Moloney murine leukemia virus (MLV) and human immunodeficiency virus (HIV) position along the phylogenetic tree is highlighted.

2.1.2 Replication cycle

Fusion of virus and host cell plasma membrane occurs upon interaction between envelope glycoproteins and specific cellular receptors. RNA genome is released in the cytoplasm, where it remains associated with the core structural proteins and the viral reverse transcriptase to form the so-called reverse transcription complex (RTC). Each viral RNA molecule is then retrotranscribed into a blunt-ended DNA copy by two jumps of the reverse transcriptase from the 3' to the 5' terminus of the template strand; this results in the duplication of U3 and U5 sequences located at the 3' and 5' edges of the RNA molecule, and in the formation of two identical long terminal repeats (LTRs) at both DNA ends (**Figure 4**). After completion of reverse transcription, RTCs are reorganized into preintegration complexes (PICs), containing viral DNA, the integrase enzyme and other viral and cellular proteins (see section 2.1.4). PICs are then translocated to the nucleus, where the viral DNA is permanently integrated into the host genome by viral integrase. Nuclear entry represents a critical step for gamma-retrovirus infectivity, since their PICs are not able to traverse nuclear membrane. Therefore MLV-related viruses can integrate exclusively in cells undergoing mitosis, when the nuclear envelope is disassembled. On the contrary, lentiviral PICs are translocated to the interphase nucleus through nuclear pores by an energy-dependent, nuclear localization signal-mediated import. The mechanism is at the basis of lentivirus capability of infecting quiescent, nondividing cells.

Once stably integrated, the viral DNA (now called provirus) is transcribed by the RNA polymerase II-dependent transcription machinery to generate both full length (unspliced) and messenger (spliced) RNAs. Viral RNAs are next translocated from the nucleus to the cytoplasm, where the host translational machinery synthesizes and modifies viral proteins. The structural components of the viral inner core and replication enzymes (products of the *gag* and *pol* genes, respectively) are in fact translated, transported and assembled as polyprotein precursors (Gag and GagPol). Full-length viral RNAs and newly synthesized envelope and core proteins assemble together at the cell periphery and at the plasma membrane, where virions finally bud out of the cell. Maturation to infectious particles is completed soon after budding, when the viral protease cleaves Gag and GagPol polyproteins into individual domains.

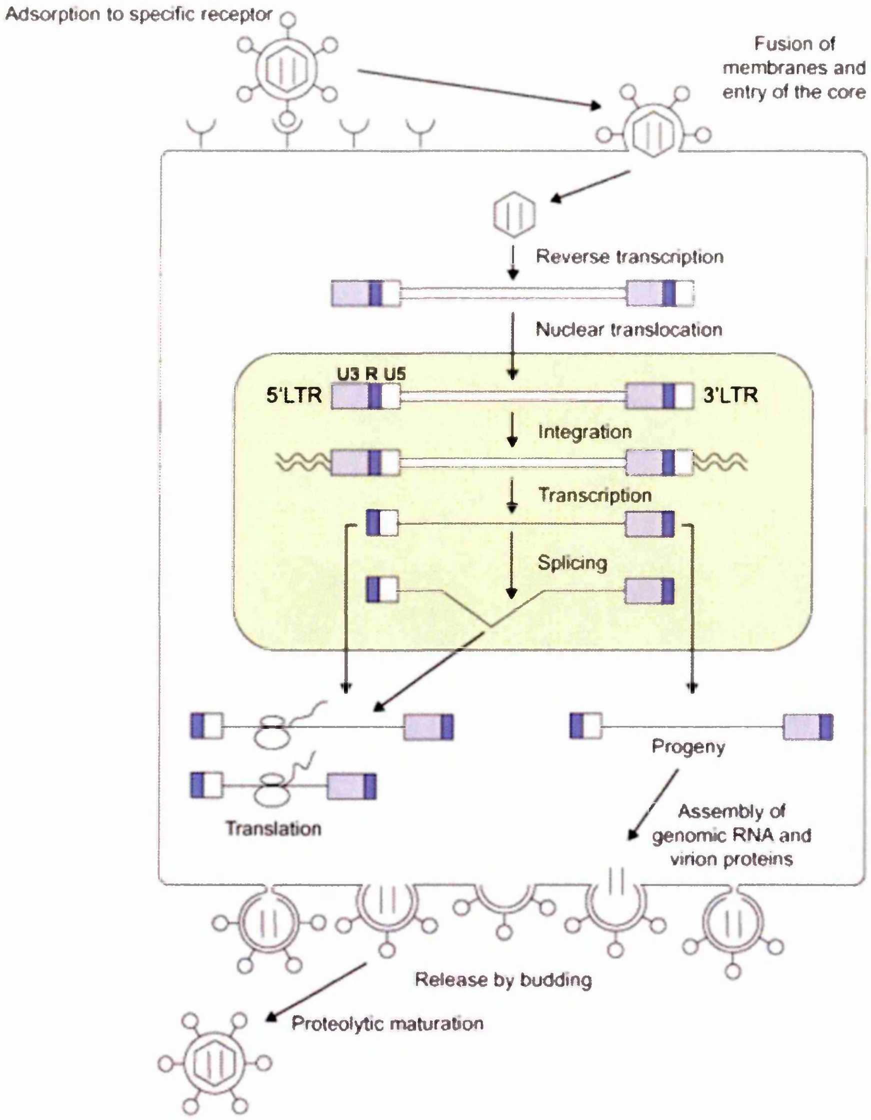


Figure 4. Retroviral life cycle. Envelope proteins interact with specific receptors on the surface of host cells, membrane fusion occurs and viral core is released in the cytoplasm. After reverse transcription, two identical long terminal repeats (LTRs), each one composed of U3, R and U5 regions, form at both ends of viral DNA. Proviral DNA is then translocated to the nuclear compartment and stably integrated into host DNA. Cellular transcription, splicing and translation machineries orchestrate expression and maturation of viral proteins. Virions are assembled at the cell periphery and released from host cell membrane; maturation to infectious particles occurs soon after budding.

2.1.3 Integration reaction: products and kinetics

Integration is an essential step in the life cycle of most retroviruses. In fact, the permanent insertion of viral cDNA in the host cell genome ensures both stable expression of viral genes in the infected cells and perpetuation of the provirus to the host cell progeny. The integration process is a two-step reaction catalyzed by the viral integrase protein; substrate for integration is the double-stranded, blunt-ended linear DNA molecule originating from retrotranscription. The reaction takes place in the context of the preintegration complex (PIC), a nucleoprotein agglomerate consisting of viral DNA, integrase dimers or multimers, a subset of virion core proteins, and specific cellular proteins. Composition of PICs is variable among different retroviruses (section 2.1.4).

The integration reaction is a multistep process (**Figure 5**). Soon after completion of viral DNA synthesis, the integrase removes two nucleotides from the 3' end of both strands of viral DNA, adjacent to a conserved *CA* dinucleotide, generating recessed 3'-hydroxyl groups; in the subsequent cleavage-ligation reaction, the processed 3'-hydroxyl ends are joined to protruding 5' ends of the target DNA. Complete integration is achieved when cellular enzymes repair gaps at each host-virus DNA junction, resulting in a 4- to 6-base pair repeat in the host DNA flanking each proviral end (bases +1 to +4/5/6 downstream of the insertion nucleotide).

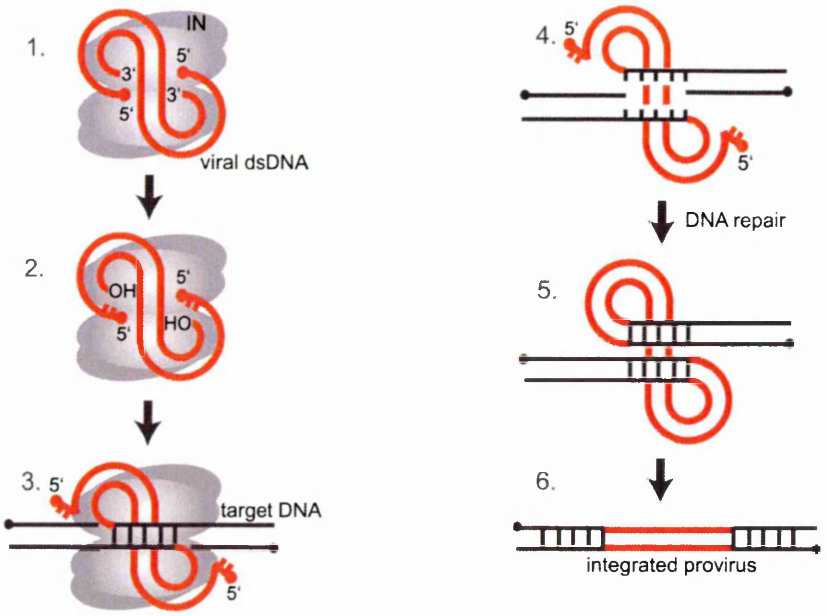


Figure 5. **Two-step integration reaction**. Gray ovals represent integrase monomers (IN), thick red lines represent viral DNA, black lines represent target DNA, and dots represent 5′ ends. Linear blunt-ended viral cDNA is bound by integrase in the context of the preintegration complex **(1)**. In the first "processing" step **(2)**, integrase removes two nucleotides from the 3′ ends of the viral DNA, exposing recessed 3′ hydroxyl groups (-OH). In the second "joining" step **(3)**, IN binds the recessed 3′ ends of viral DNA to the target DNA, in a concerted cleavage-ligation reaction. Unpairing of the target DNA between the joined ends of the viral DNA yields gaps in the target DNA **(4)**. Integration is completed when host DNA repair enzymes fill in the gaps in host DNA flanking the provirus, remove the overhangs of two nucleotides at the 5' ends of the viral DNA, and perform covalent ligation between host and viral DNA **(5-6)**.

Of the total linear cDNA coming from retrotranscription, only a certain fraction is actually integrated, resulting in a functional provirus. A significant proportion is instead degraded, while a certain amount is converted into by-products, detectable at considerable levels in the nucleus at late time-points after infection. These are dead-end circular forms that stay as extrachromosomal viral DNA molecules until degraded. There are three classes of circular unintegrated DNA molecules (**Figure 6**):

a) 1-LTR circles, originating from homologous recombination between the LTRs of the original linear DNA molecule;

b) 2-LTR circles, formed by non-homologous end joining of the linear DNA extremities;

c) auto-integration products, resulting from a suicidal, intramolecular integration of the viral DNA.

Among the others, autointegration products are the sole requiring integrase catalytic activity for their formation. The cellular protein BAF (identified as, and named after, the barrier-to-autointegration factor in Mo-MLV infection[18]) was demonstrated to participate in the regulation of autointegration product formation, as an inhibitor of suicide integration and a promoter of efficient intermolecular DNA recombination once a suitable chromosomal target is identified. The role and mechanism of BAF action are well-established for Mo-MLV[19], but a similar strategy could be reasonably attributed to HIV, whose PICs have been confirmed to contain, and depend on, BAF for integration activity[20-22].

Kinetics of the integration reaction and by-product formation can be followed by *Alu*-PCR technique, a quantitative Taqman PCR carried out with primers annealing to the retroviral LTR and to chromosomal *Alu* repeats. The strategy exploits the high frequency and random distribution of *Alu* elements in primate genomic DNA (5% of the mass of the human genome, distributed roughly 5,000 bp apart, randomly oriented). Since retroviral integration occurs at many locations in the human genome, each provirus will have a unique distance to the nearest *Alu* sequence, thus generating amplification products of different lengths. With such a technique it was possible to measure the relative amount of linear HIV cDNA product with respect to integrated provirus and 2-LTR circles[23-25]. It came out that total HIV cDNA accumulates quickly after infection, reaching a maximum abundance after 12 hours, and then declining over the next 50-60 hours. The 2-LTR circles peak in abundance

24 hours post infection and decline thereafter; integrated proviruses become detectable by 24 hours, but reach a plateau only after 48 hours. The final number of proviruses per cell is typically considerably lower than the total number of cDNA copies measured at 12 hours (up to 20-fold), indicating that only a small fraction of retrotranscribed molecules is integrated in the host genome.

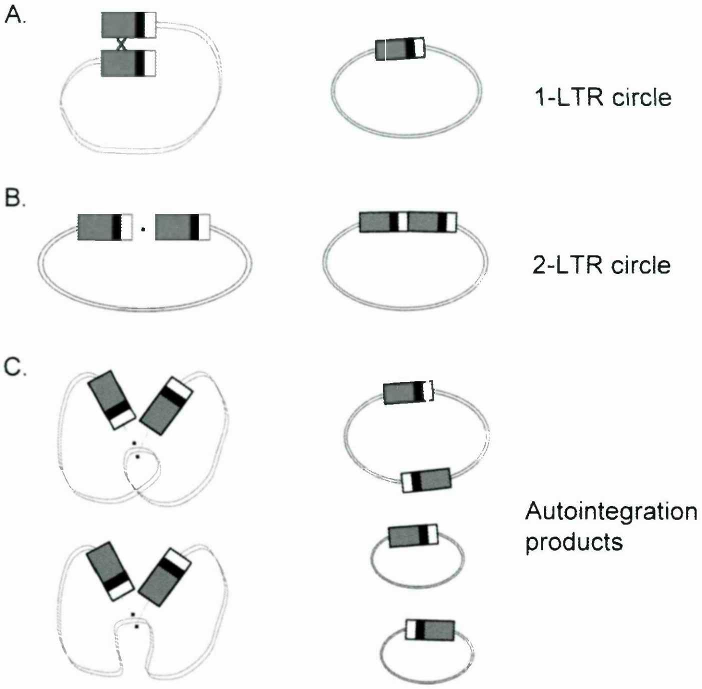


Figure 6. Unintegrated viral DNA products. Dead-end by-products deriving from viral cDNA molecules that are non-productively integrated in the host cell genome. **(A)** 1-LTR circle originating from homologous recombination of LTRs. **(B)** 2-LTR circle form by non-homologous end joining between viral DNA ends. **(C)** Suicide intramolecular integration of viral cDNA results in a single circle containing 2 LTRs or in a pair of 1-LTR circles.

2.1.4 Preintegration complex composition

Retroviral integration is mediated by the preintegration complex (PIC), a large nucleoprotein structure containing the fully reverse transcribed viral DNA associated to proteins of both viral and cellular origin. Composition and organization of PICs have been studied more extensively for HIV-1[26-30] than for MLV[31,32], but in both cases many aspects remain poorly understood. PICs isolated from both MLV- and HIV-1-infected cells contain reverse transcriptase (RT), integrase (IN) and the already described host protein BAF (see 2.1.3). Only MLV PICs retain capsid (CA) proteins, which are found only in traces in HIV-1 PICs that instead contain matrix (MA) and Vpr (viral protein R) proteins. No role in the integration process has been demonstrated so far for RT and CA proteins; MA, Vpr and IN proteins instead have been all proposed as karyophilic agents facilitating the nuclear import of HIV-1 PICs[33].

Several cellular proteins have been reported to bind HIV-1 PICs; for some of them the association occurs via direct interaction with viral IN. Among these are members of the DNA repair machinery, constitutive chromatin components, and chromatin remodelling complexes. hRad18 (the human homolog of *Saccharomyces cerevisiae* Rad18 protein) participates in the DNA post-replication/translesion repair and was shown to bind HIV-1 IN and protect it from accelerated degradation[34]. Other components of the DNA repair machinery, such as DNA-dependent protein kinase (DNA-PK)[35] and poly(ADP-ribose) polymerase-1 (PARP-1)[36], both activated upon DNA strand breaks, are also required for efficient HIV-1 integration, but no direct association with PICs was ever demonstrated for them.

Kalpana et al.[37] used a two-hybrid system to isolate a previously unknown human protein interacting with HIV-1 IN, therefore called Ini-1 (for integrase interactor 1). Ini-1 is part of the SNF/SWI chromatin-remodelling complex, a global transcriptional co-activator; interaction with viral IN was shown to stimulate its DNA-joining activity.

Another chromatin remodelling protein, this time associated with gene silencing and transcriptional repression, has been identified as an HIV-1 IN interactor; this is the human EED, member of the *Polycomb* group proteins, which showed an apparent positive effect on IN-mediated DNA integration reaction *in vitro*, in a dose-dependent manner[38].

HMG I(Y) (high mobility group) is a further example of a nonhistone chromatin-associated protein that is required for HIV-1 PIC function[39,40]. HMG I(Y) is involved in transcriptional control and chromosomal architecture and was able to restore intermolecular integration activity from salt-stripped PICs. Attempts to demonstrate binding between HMG I(Y) and purified IN have been unsuccessful, and it has been therefore proposed that the protein acts simply by binding to the HIV-1 cDNA via A/T-rich sequences. Like Ini-1, HMG I(Y) at least promotes the covalent strand transfer step of the integration reaction. A role for HMG I(Y) was also proposed in the MLV integration process[41], even if physical association with MLV PICs was never demonstrated.

LEDGF/p75 (lens epithelium-derived growth factor) is undoubtedly the best-characterized cellular cofactor of HIV-1 IN[42-46]. This transcriptional co-activator significantly stimulates IN enzymatic activity both *in vitro* and *in vivo* and might also function as a chromatin acceptor for HIV-1 PICs. In fact LEDGF/p75 is intimately associated with chromatin, through an N-terminal PWWP domain and AT-hook DNA-binding motifs and both structural features are required for HIV-1 efficient infection. This suggests a "bridging" role for LEDGF/p75, which would favour harbouring of PICs to the host DNA by binding chromatin on one side and IN on the other (further discussed in section 3.3.3).

2.1.5 Regulation of proviral transcription

A productive integration event results in the formation of a provirus, a DNA viral intermediate stably inserted into the host-cell genome. At this stage the virus mimics a cellular gene and relies almost entirely on the host-cell machinery for gene expression. This strategy, unique among animal viruses, implies that the viral genome contains a large array of *cis*-acting control elements regulating transcriptional initiation from eukaryotic promoters. Most of these elements are transcription factor binding sites lying in the LTRs of the proviral DNA, particularly enriched in the U3 region upstream of the transcription start site (first nucleotide of the R region).

Retroviral transcription is operated by the host-cell RNA polymerase II, which synthesizes cellular messenger RNAs and some small nuclear RNAs. In eukaryotic cells, the minimum requirement for RNA polymerase II transcription initiation is the assembly of a basal transcription complex onto gene promoters. For most promoters, including retroviral ones, the TATA box is the core element that directs RNA polymerase II recruitment; this is achieved through binding of the TFIID multiprotein complex, composed of a TATA-binding protein (TBP) and several TBP-associated factors (TAFs). TFIID recruitment, in turn, promotes the association of other basal factors and, finally, of RNA polymerase II. Transcription is initiated when the carboxy-terminal tail of the large subunit of RNA polymerase II is phosphorylated and the enzyme is released from the core promoter. As the transcript is elongated, the basal transcription machinery is partly disassembled, while elongation proceeds under the control of specific elongation factors.

Transcription rates are finely tuned by regulatory *cis*-acting sequences. These regions are still considered promoter elements when located in the immediate vicinity of the basal promoter. However, they are often situated at considerable distance from the promoter they modulate; in this case, they are regarded as distinct elements and termed enhancers or silencers, depending on their mediating a positive or negative effect on basal promoter activity. In retroviral LTRs, the spacing between the transcription start site and the enhancer/silencer motifs is reduced, usually less than 1 kb. Transcription factors bind these control elements in a sequence-specific manner and act as transcriptional activators as well as repressors. This is often achieved in collaboration with coactivators or corepressors, recruited to the transcription site by protein-protein interactions. Transcription factors are grouped into structural families defined by common DNA-binding motifs, implying that related proteins can bind similar or even identical binding sites. Determining which member of a given family functions on a particular element becomes therefore a challenge, and cannot be assessed but experimentally. This is also the case when looking at retroviral LTRs, where regulatory transcription factor families are readily inferred by the presence of their consensus sequence, but experimental data supporting the involvement of specific members are often lacking or controversial.

Although different retroviruses share many essential features in their gene regulation, each retrovirus has evolved unique solutions to replicate in specific cell types. Complex retroviruses also employ virally encoded *trans*-activators that act in conjunction with cellular proteins to control viral gene expression.

In the next paragraphs the two examples of Mo-MLV and HIV-1 regulation are presented as prototypes for transcriptional regulation strategies employed by simple and complex retroviruses.

2.1.5.1 Transcriptional regulation of Mo-MLV

The LTR of Mo-MLV is a paradigm for the transcriptional control machinery of simple retroviruses. The vast majority of *cis*-acting elements are located in the LTR U3 region, which includes a basal promoter and an upstream enhancer (**Figure 7**). The core promoter contains a TATA box and a CCAAT box motif; the latter is bound by the C/EBPs (CCAAT/enhancer binding proteins), a six-member family of transcription factors sharing a highly conserved, basic-leucine zipper domain involved in dimerization and DNA binding. C/EBP family members have pivotal roles in the control of cellular proliferation and differentiation, metabolism and inflammation, particularly in hepatocytes, adipocytes and hematopoietic cells[47], the natural target of Mo-MLV.

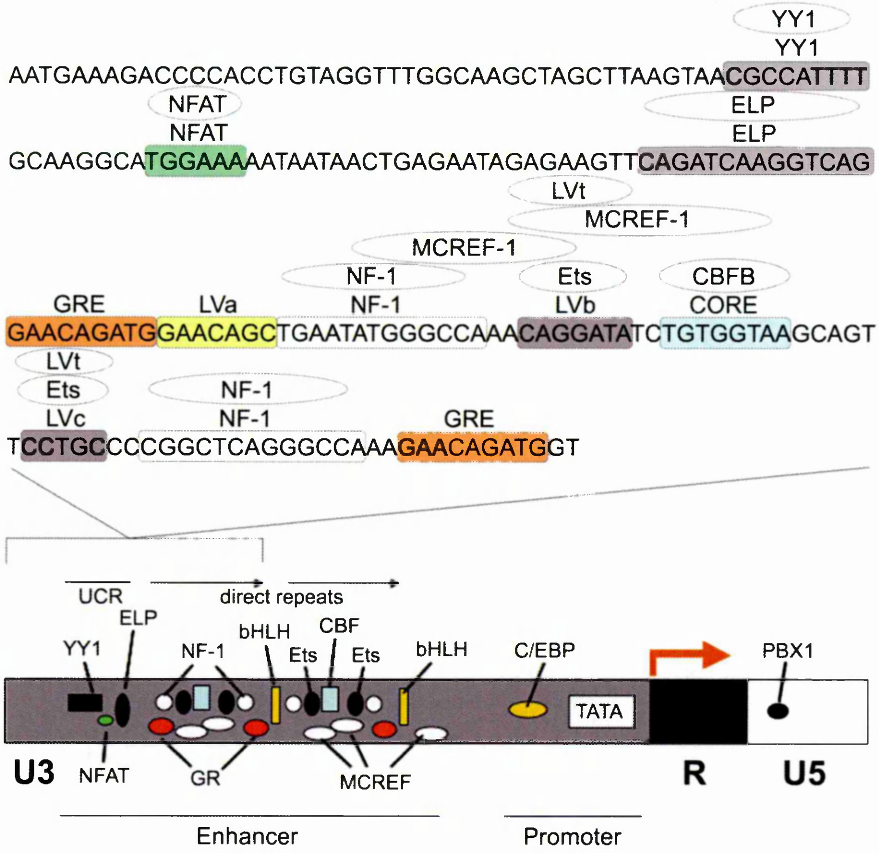

YY1
YY1
AATGAAAGACCCCACCTGTAGGTTTGGCAAGCTAGCTTAAGTAACGCCATTTT
NFAT
NFAT
ELP
ELP
GCAAGGCATGGAAAAATAATAACTGAGAATAGAGAAGTTCAGATCAAGGTCAG
LVt
MCREF-1
MCREF-1
NF-1
Ets
CBFB
GRE
LVa
NF-1
LVb
CORE
GAACAGATGGAACAGCTGAATATGGCCAAACAGGATATCTGTGGTAAGCAGT
LVt
Ets
NF-1
LVc
NF-1
GRE
TCCTGCCCCGGCTCAGGGCCAAAGAACAGATGGT
UCR
direct repeats
ELP
YY1
NF-1
bHLH
CBF
Ets
Ets
bHLH
C/EBP
PBX1
TATA
U3
NFAT
GR
MCREF
R
U5
Enhancer
Promoter

Figure 7. Structural organization of the Mo-MLV LTR. The scheme shows transcriptional control elements of Mo-MLV LTRs specifying which cellular factors recognize them. U3 sequence up to the first 75-bp direct repeat is also shown in detail, with transcription factor binding sites highlighted by colored boxes and nuclear factors known to bind them in ovals. The basal promoter includes a CAAT box (recruiting C/EBP factors) and a TATA box. UCR: upstream conserved region, containing YY1, NFAT and ELP motifs. PBX1 consensus element is the only regulatory motif identified within the U5 region up to date.

The enhancer structure has been extensively characterized and is composed of a set of 5' unique motifs, the so-called upstream conserved region (UCR), followed by two direct repeats of approximately 75 bp each, containing binding sites for multiple nuclear proteins, closely packed and partially overlapping. The UCR is a particularly well-conserved region shared among different gamma-retroviruses[48] (Moloney MLV, spleen focus forming virus, myeloproliferative sarcoma virus and Friend MLV); the region was initially identified as a negative control region in Mo-MLV LTR, since it contains two potentially inhibitory motifs. One is a target for the embryonal long terminal repeat-binding protein (ELP), a mammalian homolog of the Drosophila Fushi-Tarazu transcriptional repressor that binds to, and suppresses transcription of, the MLV LTR in undifferentiated murine embryonal carcinoma cells[49]. The second inhibitory sequence is recognized by the bifunctional Ying Yang 1 (YY1) protein, originally described as the UCR binding protein-I (UCRBP-I)[50]. YY1 is a ubiquitously expressed factor and can act either as a transcriptional repressor or as an activator, in both cellular and viral enhancers. Despite YY1 being identified at first as a negative regulator of Mo-MLV LTR[50], subsequent analysis in cells of hematopoietic and non-hematopoietic origin revealed that deletion of the UCR results in a significant reduction of enhancer activity[48]. The decrease in expression levels was accounted for partly by the YY1 motif deletion and partly by deletion of a third binding site within the UCR, identified as an NFAT (nuclear factor in activated T cells) motif. The NFAT family comprises five members expressed in most immune-system cells, where they play a substantial role in the transcription of cytokine genes and other genes critical for the immune response[51].

At least eight sites for protein binding have been mapped to each copy of the 75 bp direct repeats. The glucocorticoid responsive element (GRE) was demonstrated in rat fibroblastoma cell lines to bind the glucocorticoid receptor in the context of Moloney murine sarcoma virus (Mo-MSV)[52,53] LTR, a virus strictly related to Mo-

MLV. However, *in vivo* footprinting of Mo-MLV LTR in murine fibroblasts and T cells failed to show occupancy of the GRE sites[54], suggesting that glucocorticoid responsiveness of Mo-MLV LTR could be cell-context dependent.

Overlapping to the GRE is the LVa (leukemia virus factor a) motif[55], an Ephrussi box (E-box) element recognized by several transcription factors from the basic helix-loop-helix (bHLX) structural family.

NF-1 (nuclear factor I/X) is a CCAAT-binding transcription factor which binds to two sites in each enhancer repeat[56]. Like GRE motif, the occupation of NF-1 sites varies among cell types; *in vivo* footprinting experiments revealed binding of NF-1 in Mo-MLV-infected fibroblasts but not in lymphoid cells[54].

The LVb (leukemia virus factor b) site has been shown to bind many proteins of the Ets transcription family, including Ets-1 and Ets-2[57], LVt[58], GABP and Fli-1. Ets proteins are a family of helix-loop-helix transcription factors regulating the expression of a myriad of genes involved in the development and differentiation of a variety of tissues and cell types. This functional versatility emerges from their interactions with other structurally unrelated transcription factors[59].

The CORE motif is recognized by the core binding factor (CBF)[60], a heterodimeric protein whose alpha subunit (AML1, acute myeloid leukemia 1) interacts directly with DNA, while the beta subunit (CBFB) increases the stability of CBF-DNA complex. The complex plays a major role as a transcriptional activator in hematopoiesis. There are evidences that Ets and CBF cooperate *in vivo* to regulate transcription from the Mo-MLV enhancer by concerted binding to the LVb and CORE sites[61].

Overlapping to the LVb and CORE sites is the binding motif recognized by MCREF-1 (mammalian type-C retrovirus enhancer factor-1), a nuclear protein only partially characterized[58,62].

Recent work identified an additional regulatory sequence in the U5 region of Mo-MLV LTR, perfectly conserved in 14 other murine retroviruses. This is the PBX consensus element (PCE) recognized by heterodimers of the homeodomain proteins PBX1 (pre-B-cell leukaemia transcription factor 1) and PREP1 (PBX regulating protein 1)[63]. Both mutations of the PCE and inhibition of PBX1 protein synthesis by antisense oligonucleotides and siRNA strategies significantly diminish viral transcription, whereas PBX1 and PREP1 over-expression enhances MLV transcriptional levels.

Although the exact identity of each cellular protein functioning at a specific site is still under investigation, mutagenesis studies of the enhancers and promoter indicate that all identified binding sites correspond to positive-acting elements within Mo-MLV LTR and are therefore necessary for high-level LTR transcriptional activity. Indeed, such a promiscuous array of binding sites for tissue-specific as well as ubiquitously expressed transcription factors allows sustained Mo-MLV expression in most mammalian cell types, of hematopoietic as well as non-hematopoietic origin (*e.g.* deriving from neural, epithelial and muscular tissues).

2.1.5.2 Transcriptional regulation of HIV-1

HIV-1 transcription is regulated by *cis*-acting elements spread over U3 and R regions of the LTRs (**Figure 8**). A TATA box defines the basal promoter; immediately upstream is a strong enhancer element, composed of two NF-kB and three Sp1 consensus sites. NF-kB proteins are transcriptional activators encoded by the NF-kB/Rel gene family, functioning in a variety of homo and heterodimeric configurations. Activation of Nf-kB proteins is induced upon T-cell and monocyte activating signals but also in response to cytokine stimulation. NF-kB proteins were shown to be important for HIV-1 transcription in a series of independent studies[64-66]. Individual tests of different family members have shown that the various NF-kB homo and heterodimers may exert differential effects on HIV gene expression, the most common always being a potent activation of LTR transcription. This is apparently achieved in cooperation with the constitutive Sp1 transcription factor, whose interaction with NF-kB family member RelA was demonstrated to augment binding to and transactivation of the HIV LTR[67]. Consistently with this observation, mutation of both the NF-kB and the adjacent Sp1 sites is necessary to severely reduce viral replication, entailing that the highly conserved arrangement of the two motifs enhance the efficiency of these factors in activating HIV transcription.

Upstream the NF-kB and Sp1 positively acting motifs is the so-called negative response element (NRE), exhibiting both negative and positive regulatory properties. Among repressor proteins binding the NRE are COUP-TFs (chicken ovalbumine upstream promoter transcription factors), members of the steroid/thyroid hormone receptor superfamily; mutation of COUP site resulted in an increase of LTR-directed transcriptional activation[68]. The negative effect on HIV transcription mediated by NFAT-1 (nuclear factor in activated T cells 1) binding site is much more controversial; while deletion of NFAT-1 consensus from the HIV LTR resulted in the production of viruses replicating more rapidly than parental ones in T cell cultures, the same motif was not able to modulate the expression levels of an HIV LTR-driven heterologous gene, neither positively nor negatively[69,70].

Similarly unclear is the role of TCF-1 sites, recognized by a T-cell-specific transcription factor that activates the T-cell receptor C alpha enhancer[71].

The NRE also contains *cis*-elements with a stimulatory effect on HIV transcription; these are two immediately adjacent E-box and Ets binding sites located -130 to -166 bp upstream of the transcription start site. Cooperative DNA binding of the helix-loop-helix protein USF-1 (upstream stimulatory factor-1) and of Ets-1 protein was demonstrated on these motifs in T cells. The two proteins were also shown to interact directly, forming a transactivation complex required for full transcriptional activity of the HIV-1 LTR[72]. Beside the E-box located in the distal enhancer, USF-1 can also bind to two initiator-type elements near the transcription start site of the HIV-LTR (-3 to +20; +35 to +60), again with a stimulatory effect. The upstream initiator site partially overlap with a -17 to +27 region recognized by three other factors (YY-1[73] in cooperation with LBP-1[74], and TDP-43[75]), all acting as transcriptional repressors. LBP-1 (also known as upstream binding protein-1, UBP-1) recognizes three sites within this region and has an additional low-affinity binding site overlapping the TATA-box; when interacting with this element, LBP-1 specifically represses HIV-1 transcription by preventing the recruitment of the general initiator factor TFIID to the core promoter[76].

Like other complex retroviruses, HIV-1 has evolved a regulatory mechanism relying upon a virally encoded transcriptional activator, the product of the *tat* gene. Tat protein alone is able to enhance LTR-directed transcription by hundreds to thousands of fold, and mutations of the *tat* gene result in complete abolishment of HIV replication[77,78]. The Tat-responsive region (TAR) is located at the 5' end of viral RNAs (+1 to +59); as soon as it is transcribed, TAR forms a stable stem-loop secondary structure that is recognized and bound by Tat protein. Once bound, Tat is able to increase the processivity of RNA polymerase II by recruiting various transcription factors such as the TBP, the general transcription factor TFIIB and the positive transcription elongation factor B (P-TEFB). This leads to the formation of very active elongating transcription complexes that hyperphosphorylate RNA polymerase II C-terminal domain[79], thus ensuring continuous and rapid reinitiation of transcription to the benefit of the viral promoter strength. This scenario favors the notion that Tat acts by interacting with multiple viral and cellular partners at a time. In accordance with the view that Tat is multi-functional it has been shown that Tat also regulates cotranscriptional mRNA capping[80] and splicing[81].

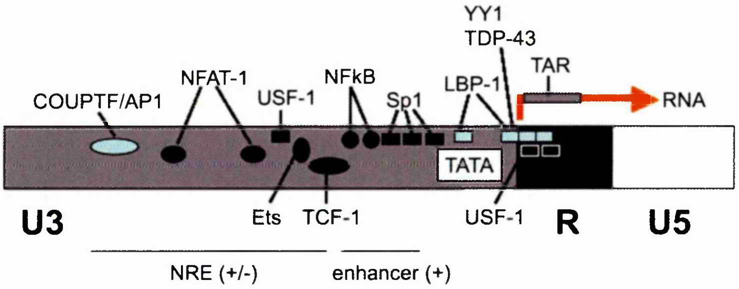


Figure 8. **Structural organization of the HIV-1 LTR**. Schematic representation of *cis*-acting regulatory elements in HIV-1 LTR; transcription factors known to bind transcriptional control elements are specified. HIV-1 LTR includes a distal negative response element (NRE), exhibiting both positive and negative regulatory properties, and a proximal enhancer, promoting proviral transcription. TAR: Tat response element, localized within the R region of nascent RNA transcripts.

2.2 Retroviral vectors and gene therapy

In the last decade it has been clearly established that the transfer of a therapeutic gene into somatic cells (gene therapy) has an enormous potential for the management of many diseases, both inherited and acquired. The ability of retroviruses to integrate efficiently into the genomic DNA of animal cells and be stably replicated and transmitted to all their progeny was a strong incentive for the development of retroviral gene transfer vectors. From many studies it was clear that retroviral genomes could accommodate extensive alterations, and, even though these changes often resulted in replication defects, altered viruses could be propagated in the presence of a replication-competent, "helper" virus[82,83]. Such vector preparations were necessarily contaminated by the helper virus, spreading after infection of target cells, which rendered the procedure unacceptable for human gene therapy purposes. A major advance in retroviral vector design for gene therapy applications came with the development of retroviral packaging cells that provide all of the retroviral proteins in *trans* but did not produce replication-competent virus[84,85]. Many packaging cells of the first generation still produced helper virus as a result of recombination events, but evolution in design has enormously reduced this frequency. In the last generation retroviral vectors only the minimal viral elements required for high efficiency transfer are retained, while the remaining viral coding regions are either eliminated or supplied in *trans*. This is possible because the early steps of the retroviral replication cycle, from viral entry to integration, are completely independent of viral protein synthesis, but instead rely on viral proteins packaged within the virions (RT, IN, protease) and on *cis*-acting elements included in the viral genome. These are a promoter and a polyadenylation signal for viral genome production in the packaging cell, a packaging signal for incorporation of vector RNA into virions, signals required for reverse transcription and short repeats at the termini of viral LTRs necessary for integration. All the intervening genomic material can be replaced with the sequence of interest, to accommodate up to 10 kb of heterologous DNA. To further reduce the risk of replication-competent recombinants, gamma-retroviral and lentiviral vectors are often engineered to become self-inactivating (SIN), meaning that they lose the transcriptional capacity of their LTR once transferred to target cells[86-88]. This is achieved by deleting the transcriptional enhancers or the enhancers and promoter in the U3 region of the 3' LTR from the DNA used to produce the vector RNA. During the first cycle of reverse transcription, occurring upon target cell infection, this deletion is transferred to the 5' LTR, generating a transcriptionally inactive provirus (**Figure 9**). However, any promoter internal to the LTRs in such vectors will still be active. Besides minimizing the frequency of replication-competent recombinants, this strategy also reduces transcriptional interference between LTRs and internal promoter/s, and eventually transactivation effects on adjacent cellular genes once the provirus is integrated in the host genome.

Packaging systems also allow production of transfer vectors with heterologous envelope proteins, so that the viral tropism can be modified or extended at wish. For instance, pseudotyping vectors with the surface protein of the vesicular stomatitis virus (VSV-G) expands viral host range to include insect, mammalian, fish and amphibian cells; moreover, being VSV-G mechanically more stable than other envelope proteins, it is possible to concentrate VSV-G-pseudotyped particles by ultracentrifugation, collect high-titer vector stocks, and store them for long-term periods[89,90].

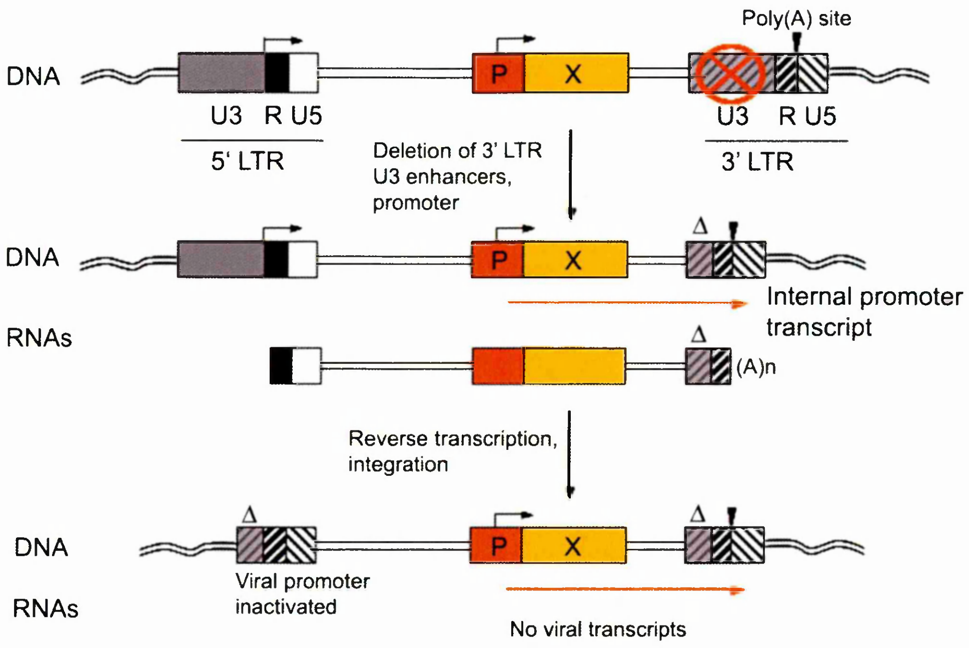


Figure 9. Construction of self-inactivating vectors. The U3 region of the 3' LTR is partially deleted (Δ) to remove enhancers and/or promoter from the DNA used to produce vector RNA. Deletion is transferred to 5' LTR upon target cell infection and reverse transcription. Black arrows indicate promoter transcriptional activity; long red arrows represent transcripts from the internal expression cassette. P, internal promoter; X, gene of interest.

Due to these features, retroviral vectors are among the most widely used tools for gene delivery in general and for human gene therapy in particular (**Figure 10**). As a matter of fact, for some problematic but extremely valuable therapeutic targets, such as human stem cells, retroviral vectors represent the only available strategy to transfer therapeutic genes with efficiency compatible with clinical applications. Indeed, the transplantation of autologous, genetically modified stem cells is a promising therapeutic approach for a variety of genetic disorders of hematopoietic, epithelial or neural cells. These include severe combined immunodeficiencies (SCIDs)[91], thalassemias[92], lysosomal storage disorders[93-95], skin adhesion defects[96] and hemophilia[97,98]. Gamma-retroviral vectors derived from murine leukemia viruses (RV) have been used in hundreds of gene therapy trials since 1991. However, for a number of clinical applications RV vectors are highly likely to be replaced by lentiviral vectors (LV) derived from human or animal immunodeficiency viruses. In fact, LV vectors transduce a wide variety of human cells *ex vivo* and *in vivo*, achieving high-level and long-term expression of transgene(s). Most importantly, due to the active nuclear transport of the PICs, LV vectors can transduce both dividing and non-dividing cells, a clear advantage when targeting quiescent or rarely dividing stem cells. Several years of research have improved the efficacy and safety of the LV vector technology to such an extent that the first clinical trials using HIV-1-derived vectors have been recently approved and started[99-101].

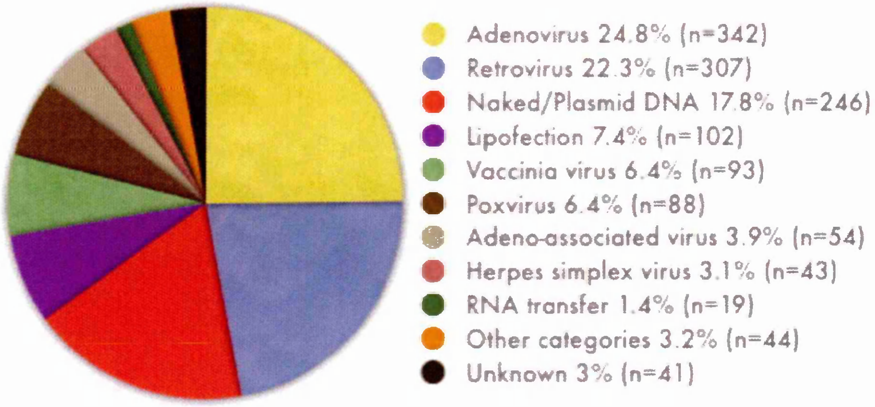


Figure 10. Vector used in gene therapy clinical trials. The chart shows in what proportion different gene delivery systems are used in all the approved, ongoing or completed human gene therapy clinical trials worldwide. *n* indicates the number of trials conducted with each vector type. Data are obtained from The Journal of Gene Medicine clinical trial site[102].

3. Retroviral integration features and mechanisms: state of the art

3.1 Insertional mutagenesis as a gene therapy adverse event

Replication-defective retroviral vectors are excellent gene therapy tools, efficiently delivering therapeutic genes to a variety of cell types. Thanks to the integration reaction, retroviral DNA is stably inserted into the host cell chromatin, providing long-lasting transgene expression and permanent transmission to the host cell progeny.

Due to its easy accessibility, blood is one of the organs in the human body that is of special interest for gene therapy interventions. The blood system reveals a hierarchical structure, with a relatively limited number of hematopoietic stem cells (HSCs) being the origin of any mature blood cell. Thus, modification of a small number of long-term repopulating stem cells might be sufficient to achieve therapeutic efficacy in the entire blood system.

Because they reach high expression levels in the hematopoietic system, Mo-MLV-based vectors (RV) carrying wild type LTRs have been largely, and in some cases successfully, used in gene therapy for blood disorders since 1991. These vectors were considered relatively safe, because the integration events were believed to be random, and the chance of accidentally disrupting or activating a gene remote. *In vitro* integration models had identified some factors enhancing or reducing insertion efficiency, such as nucleosomal assembly, presence or absence of DNA-binding proteins[103], and DNA physical structure[104]; however, these observations could not even hint at a risk related to vector insertion in the human genome.

Nevertheless, the oncogenic potential of murine RV has been known for decades and even largely exploited to identify genes involved in murine and possibly in human cancers (in the so-called "retroviral tagging" approach[105]). In fact, administration of replication-competent RV to susceptible mouse strains often lead to tumor development, as a result of insertional deregulation of growth-controlling genes followed by clonal expansion of cells hosting such integrations. Replication-defective RV vectors were also reported to cause insertional oncogenesis in mice[106], but the risk of mutagenesis of cellular genes promoting a malignant phenotype was estimated to be low (10^{-7} per insertion), again assuming that retroviral integration occurs randomly over the genome. Such assumption was readily reconsidered when a lymphoproliferative disorder was reported in one patient treated for X-linked severe combined immunodeficiency (X-SCID) with MLV-transduced HSCs[5]. Mapping of RV integrations in the predominant T-cell clone revealed a single proviral insertion within the LMO-2 locus, associated with upregulation of transcript and protein levels. Aberrant expression of the LMO-2 protein had been already reported in spontaneous cases of acute lymphoblastic leukemia, resulting from the chromosomal translocations t(11:14) and t(7:11). These observations lead to the conclusion that the leukemia-like disease was a consequence of an insertional mutagenesis event, and that a reassessment of the potential risk of retrovirally mediated gene therapy was necessary. This became obvious as a similar complication was reported in three more patients enrolled in the same clinical study[6,8] and also in one patient recruited in an independent X-SCID trial[2,7]. The five adverse events have remarkable features in common: all but one malignant clones hosted at least one RV insertion nearby the proto-oncogene LMO2, always resulting in LMO2 protein over-expression, and all leukemias developed 2 to 5 years after gene therapy treatment. These facts suggest that the leukemogenesis mechanism is likely to be the same. The product of LMO2 gene (LIM-only protein 2) acts as a bridging molecule in transcription factor complexes, thanks to several zinc-binding finger-like motifs; the protein is expressed early in hematopoiesis, and it is down-regulated during commitment in all except the erythroid lineage[107]. In T cells, down-regulation of the protein appears to be crucial, since mice constitutively expressing Lmo2 in the thymus develop T-cell leukemia, preceded by an accumulation of immature T cells[108,109]. This indicates that LMO2 deregulation could increase susceptibility to leukemia by blocking T cell differentiation. It was also suggested that an additional role in the X-SCID adverse events was played by the transgene delivered to HSCs, the IL2Rγc gene. The gene encodes a signaling subunit common to several interleukin receptors, all of which promote T-cell proliferation upon ligand binding. If LTR-driven LMO2 over-expression blocks T cell development at a stage in which one of IL2Rγc partners is present, a complete interleukin receptor may assemble, rendering the cells hypersensitive to growth factors and inducing their proliferation. According to this model, cooperation between LMO2 and IL2Rγc, together with secondary mutations, would give rise to the observed clonal T cell leukemia[110]. In fact, IL2Rγc role as cooperative oncogene in the human gene therapy setting is still controversial; recent reports using murine models have suggested that the IL2Rγc itself could contribute to leukemic transformation[111,112], whereas functional assays performed in human CD34$^+$ HSCs showed no effects of IL2Rγc over-expression on T cell development and proliferation[113]. As a matter of fact, no clonal lymphoproliferation has been reported to date in patients treated for ADA deficiency[1], despite the observation of a high frequency of integration near LMO2 and other T-cell proto-oncogenes[114], indicating that either the therapeutic transgene or the X-SCID background[115], or both, might have been critical factors for tumor onset.

Whatever the mechanism of leukemia development, the striking outcome of gene therapy of X-SCID is that 5 out of 19 patients successfully treated in two independent clinical trials developed a malignancy due to insertional mutagenesis, 4 of which even at the same genomic locus. This observation led the scientific community to necessarily reconsider both the assumption of random distribution of retroviral integration in the genome and the risks associated to retroviral gene transfer in human beings.

3.2 Non-random integration pattern of retroviral vectors in mammalian genomes

Since a concrete risk of developing tumors by insertional mutagenesis was assessed in the X-SCID trial[7,8], understanding the mechanisms that dictate retroviral target-site selection in the human genome has become a major safety issue in the field of gene therapy. A deeper investigation of retroviral insertion preferences was also necessary to explain the basic virology underlying the integration process, which is still far from being completely understood.

Before completion of genome sequencing projects, it was impossible to obtain an accurate global picture of retroviral integration events. Early studies using *in vitro* integration models identified several factors relevant to integration site selection, such as DNA bending induced by nucleosomal assembly, steric hindrance to target DNA due to DNA binding proteins[103], and DNA intrinsic structure[104]. However, target site selection *in vivo* remained poorly understood. Pioneering *in vivo* studies on Mo-MLV and ASLV (avian sarcoma leukosis virus) integration pattern produced conflicting results, with some reporting that transcriptionally active regions favor retroviral integration[116,117] and others suggesting the opposite[118].

As soon as almost complete sequences were available for several vertebrate genomes, genome-wide approaches were used to analyze integration targeting in a statistically rigorous manner. Large-scale, high-throughput methods were designed to clone and sequence the junctions between proviral and host-cell DNA. The position of integration sites in the genome was then correlated with other annotated features, such as presence of genes, transcriptional activity, centromeric regions[119], fragile sites[120], CpG islands, hypersensitive sites[121] and, very recently, epigenetic modifications[122]. This was done in a variety of cell types derived from different species (bird, human, non-human primate, murine primary cells and/or cell lines) after acute infection with different retroviruses or retroviral vectors (among others HIV-1, SIV, Mo-MLV, ASLV, extensively reviewed by Bushman *et al.*[123]).

Considering the common assumption of random distribution of retroviral integrations in the genome, the results of these large-scale surveys were almost astonishing. Not only did they uncover genomic features systematically and specifically associated with retroviral insertions, but they also pointed out that each retrovirus has a unique, characteristic pattern of integration within the human genome.

3.2.1 Primary DNA sequence and integration site selection

One of the first genomic features to be investigated for a role in target site selection was the primary DNA sequence at the target site. In fact, while integrase has strict sequence requirements for the viral DNA ends (the dinucleotide CA, invariably located 2 bp from both ends of the viral termini, and sequences up to 15 bp upstream of the CA), target site sequences are very diverse. A recent study re-analyzing integration sites from HIV-1, Mo-MLV, ASLV and SIV-infected cells found a weak statistical palindromic consensus, centered on the virus-specific duplicated target site sequence[124]. The consensus was weakly conserved but distinguishable between different retroviruses, as later confirmed by other larger surveys[122,125]. The same consensus was also found around integration sites in naked genomic DNA catalyzed *in vitro* by PICs, suggesting that the observed preferences are due to the integration machinery itself and not to host factors. Apart from the primary sequence, DNA structural properties such as A-philicity, DNA bendability, protein-induced deformability, and hydrogen bond potential patterns were also investigated for a positive or negative correlation with target site selection. All of these structural properties were found favored at the integration sites of one or another retrovirus. By the author's own admission, it is difficult to think of the consensus as the most favorite sequence at each base, but instead it might be better to consider certain bases being excluded at certain positions to meet the spatial or energy requirements of the integration complex.

Given that target site selection is only weakly sequence specific, other genomic features were explored.

3.2.2 Retroviral integration and genes

Having in mind the transactivation effect of Mo-MLV LTRs on the LMO2 gene in the X-SCID patients, the correlation between integration sites and transcriptional units was promptly investigated. Different retroviruses show distinct target site preferences. Considering the well-characterized RefSeq genes as the reference category, around 30% of the human genome consists of genes[126]. HIV-1 and SIV integrations are found inside genes with frequencies ranging from 60 to 85%, depending on the cell type, while the frequency for Mo-MLV and ASLV is between 40 and 60%[11,12,121,127-131]. Transcription units are therefore preferential targets for retroviral and especially for lentiviral integration, this also entailing an insertion bias towards gene-rich regions on chromosomes.

The next step was to investigate whether there were any preferences in the location of integration sites along the transcription unit. Remarkably, no biases were found for HIV-1, SIV or ASLV, while a strong preference for promoter-proximal regions was reproducibly observed for Mo-MLV. Indeed, up to 20% of Mo-MLV integrations landed within 5 kb upstream or downstream of a transcription start site (TSS). Accordingly, a strong association was found between Mo-MLV insertion sites and CpG islands within 1 kb. CpG islands are chromosomal regions enriched in the rare CpG dinucleotide that often correspond to gene-regulatory regions, and therefore promoters, containing clusters of transcription factor binding sites. ASLV integrations are only slightly biased towards CpG islands, while HIV-1 integration is even disfavored. The main determinant of MLV promoter preference is the viral integrase, presumably through a direct tethering interaction with transcription factors and/or other proteins bound at TSSs. This was elegantly demonstrated in HeLa cells using a chimeric HIV virus packaged with a Mo-MLV integrase (HIVmIN)[121]. Such a vector recapitulated most of the Mo-MLV integration biases, showing the typical clustering of insertion sites around the TSS and the well-documented MLV preferences for CpG islands and DNaseI hypersensitive sites.

These findings imply a profoundly different integration mechanism for MLV with respect to other retroviruses, which cannot but affect its application as a gene therapy vector. Preferential integration near the TSS of host genes, where LTR transactivation can be more effective, undoubtedly confers to RV vectors an increased genotoxic potential compared to other vectors.

3.2.3 Retroviral integration and gene activity

Once the preference for genes was established, transcriptional profiling analyses were performed on host cells to assess the influence of transcriptional activity on integration site selection. The median expression level of genes targeted by HIV-1 and Mo-MLV integration events was consistently higher than the median expression level of all genes assayed in the microarray[9,11,13,114,127]. Only a weak bias in favor of active genes was instead observed for ASLV[11].

Since transcriptional activity favors integration of Mo-MLV and HIV-1 and that different cell types show unique transcriptional profiles, the effects of cell-type-specific transcription on integration pattern was assessed, at least for HIV-1[11]. As largely expected, genes that were relatively more active in a given cell type were more likely to be targeted by HIV integration. However, the bias was quantitatively modest, probably because most of the cellular program of gene activity is overlapping among many cell types.

3.2.4 Retroviral integration and transcription factor binding sites

Given the MLV propensity to integrate nearby promoters of active genes, one would expect to observe an enrichment of transcription-factor binding sites (TFBSs) near the integration sites of this virus. By now the idea has been pursued by a single group of researchers, and results were reported in the same paper describing the role of MLV integrase in directing PICs to TSSs[121]. A collection of 531 positional weight matrices (representing a collection of transcription factor binding sites) was used to annotate ± 1 kb-regions surrounding the integration sites of wild type MLV and HIV vectors, and of chimeric HIV vectors packaged with an MLV integrase (HIVmIN) and/or an MLV Gag protein (HIVmGagmIN and HIVmGAG, respectively). The results were then compared to matched random control sites to assess statistically significant enrichments. wt-MLV, HIVmIN, and HIVmGag-mIN data sets showed the highest numbers of significantly enriched TFBSs, many of which were in common to all groups or shared between two out of three groups. wt-HIV and HIVmGag returned far fewer TFBSs, with no motifs in common at all. However, few of the sites associated to MLV integration were still found enriched when promoter sequences were used as controls instead of randomly chosen genomic sites. This suggested that some general features of promoters attract MLV integration, more than specific interactions with TFs. Nevertheless, a regression analysis indicated that the presence of a nearby promoter could not fully account for the favorable effect of TFBSs on MLV integration frequency, leading the authors to admit a possible effect of TFBSs on MLV integration targeting beyond just marking promoters.

3.3 Proposed mechanisms for integration site selection

The *genus*-specific integration patterns of HIV-1, Mo-MLV and ASLV imply a distinct molecular mechanism directing integration site selection for each retrovirus. Target DNA accessibility can explain some common characteristics, like preference for active genes and avoidance of centromeric heterochromatin, but other peculiar features, like MLV preference for promoter-proximal regions, require a different, more complex model.

3.3.1 Ty retrotransposons: a paradigm for tethered integration

Studies of retrotransposons in yeast provide an interesting candidate mechanism. Ty elements are well-studied yeast retrotransposons that replicate by cycles of transcription, reverse transcription and integration similar to retroviruses, except for the fact that all the steps occur inside a single cell. This life-style poses special problems. Yeast genome is 60 to 70% coding and a randomly integrating element is at high risk of committing suicide by insertional inactivation of host gene. Probably for this reason, Ty elements evolved strategies to actively select targets where insertion would not compromise host fitness. There appear to be at least three distinct mechanisms to avoid host genes, exemplified by the Ty1, Ty3 and Ty5 elements. Both Ty1 and Ty3 integrate at the 5' ends of RNA polymerase III-transcribed genes, in regions that can tolerate insertions with no adverse events, while Ty5 favours integration at telomeres. Ty3 element targets tRNA genes with extraordinary precision, inserting within few base pairs of the TSS. This is probably mediated by local tethering of PIC to the TFIIIB component of the basal transcription machinery[132]. The Ty1 element integrates with less selectivity, in a window of about 750 bp upstream of RNA polymerase III TSSs; the histone deacetylase, Hos2, and the Trithorax-group protein, Set3, both components of the Set3 complex, have been recently proposed to tether Ty1 to tRNA genes[133]. The Ty5 element shows a further integration specificity, with 95% of insertions found either at telomeres or at silent mating loci; in this case, the heterochromatin protein silent information regulator 4 (Sirp4) is involved in specifying integration sites, through direct interaction with the Ty5 encoded integrase[134,135].

In each of these cases, there is evidence that Ty integration complexes are tethered to their preferred sites by interaction with specific cellular proteins, mediating local integration. It is reasonable to suppose that such a tethering mechanism might also operate for retroviruses, with a targeting strategy opportunely suited to promote their evolutionary persistence. As discussed above, intracellular Ty retrotransposons evolved to direct their integration outside transcription units, thus minimizing the risk of host gene perturbation. On the contrary, acutely infecting retroviruses need to maximize the production of progeny virions by producing the largest number possible in the shortest time, and integration in transcriptionally active regions may facilitate high-level transcription. The retroviral integration machinery probably evolved accordingly, allowing interactions with host nuclear proteins enriched in active chromatin regions.

3.3.2 Tethering models for retroviral integration

A proof of principle that retroviral PICs can be tethered to integration sites by cellular interactors is provided by several *in vitro* studies performed with engineered integrases fused to sequence-specific DNA binding domains. Such hybrid integrases are capable of targeted integration *in vitro*, demonstrating that tethering of integrase protein to target sites can constrain integration site selection. Different combinations were tested, all with encouraging results. HIV-1 integrase (IN) was fused to the

DNA-binding domain of the phage lambda repressor protein λR[136], of the *Escherichia coli* LexA repressor[137], and of the zinc finger proteins E2C[138]and zif268[139]; ASLV integrase was also linked to *E. coli* LexA protein DNA-binding domain[140]. All these engineered integrases programmed integration near the binding site specified by the fused DNA-binding domain *in vitro*. A certain level of efficiency was also observed *in vivo*, where the HIV IN/E2C fusion protein was demonstrated to re-direct integration into a unique E2C-binding site within the 5' untranslated region of erbB-2 gene on human chromosome 17, with seven to tenfold higher preference when compared to a wild type IN (from 0,15% to 1-1.5% of the total integrated proviruses)[141]. Off-target integration was still largely predominant, but the study represented the proof of concept that tethering can affect integration targeting, and that IN-DNA interactions might be engineered to constrain integration site selection.

If tethering is indeed involved in retroviral integration site selection, the main challenge becomes the identification of chromosomal ligands for the retroviral integration machinery and of their counterparts within the PICs. In principle, any viral or cellular component of the PIC could act as a binding partner in a tethering interaction. Several cellular proteins have been isolated as physically bound to viral PICs (hRad18, Ini-1, EED, HMGI(Y), LEDGF/p75, described in section 2.1.4); for some of them a direct interaction with viral IN was also demonstrated. Among these, the best characterized and deeply studied by now is the lentiviral integrase interactor PSIP1/LEDGF/p75.

3.3.3 LEDGF/p75: a candidate for lentiviral integration tethering

Despite its name, the lens epithelium-derived growth factor (LEDGF/p75) is a ubiquitously expressed nuclear protein, tightly associated with chromatin throughout the cell cycle[142]. The protein came to the attention of retrovirologists when it was identified in affinity-based screens for its tight binding to HIV-1 IN and it was observed that it was capable of stimulating IN catalytic function *in vitro*[42,143,144]. LEDGF/p75 is a member of the hepatoma-derived growth factor (HDGF) related protein (HRP) family, characterized by a conserved N-terminal PWWP domain, found in a variety of nuclear proteins[145,146]. Of the six described HRP family members[147,148], only LEDGF/p75 and its highly homologous HRP2 contain a second conserved domain at the C-terminus, thereafter termed IBD (integrase binding domain), that allows their interaction with different lentiviral INs[149]. The PWWP domain, together with a nuclear localization signal and a double copy of an AT-hook DNA-binding domain mediate LEDGF/p75 association with chromatin, with no apparent sequence specificity[150,151]. The cellular functions of LEDGF/p75 remain largely unknown, although a role in transcriptional activation has been proposed right after the protein was identified[152,153]. Nevertheless, the role played by LEDGF/p75 in HIV infectivity was deeply investigated. The most robust results came from studies on human cells with RNA interference knockdowns of LEDGF/p75[43,44,154-156] and on murine cells with homozygous gene-trap mutations in the LEDGF/p75 locus[156,157]. When LEDGF/p75 protein is depleted, the first effect is a re-localization of the IN enzyme to the cell cytoplasm, with loss of chromosomal association and even an increased proteosomal degradation of the viral protein. A second, important consequence is an overall reduction of HIV-1 infectivity, due to a severe impairment in the integration process. Residual integration sites were analyzed, to find a decrease in the HIV typical preference for transcription units, and an increase in insertion nearby CpG islands and promoter regions, classical targets of other retroviruses. Integration did not become random, however, and transcription units were still favored. Therefore it is still plausible that cell factors other than LEDGF/p75 participate in PIC tethering to chromosomes, although LEDGF/p75 remains the dominant cellular binding partner of HIV-1 IN, required for efficient integration and replication of the virus.

Overall, these observations suggest a model where one domain of LEDGF/p75 binds chromatin at active transcription units and the other acts as a receptor for incoming PICs; enhancing IN DNA strand transfer activity, LEDGF7p75 would then direct integration to a nearby genomic locus. Such a tethering model predicts that LEDGF/p75 should accumulate on active transcription units, but this has not been experimentally demonstrated so far. It is not even known how LEDGF/p75 might recognize active transcription units. A recent genome-wide study found a positive correlation between HIV-1 insertion sites and certain post-translational histone modifications[122]; accordingly, one possible model for LEDGF/p75 recognition of active transcription units would be via the histone modifications specifically marking them.

3.4 In vivo clonal expansion of MLV-transduced human and murine hematopoietic cells

When the first case of leukemia was observed in the X-SCID gene therapy trial, all former experience in animal models and human gene therapy studies was thoroughly reviewed to determine the incidence, if any, of neoplastic transformation in transduced cells[158,159]. At that time there had been only one additional report of malignancy arising from transduced cells after insertional activation of a proto-oncogene in an animal model[106]. Using a replication-defective Mo-MLV-based vector, these authors introduced a clinically used reporter gene (ΔLNGFR, a truncated form of the nerve growth factor receptor) into murine bone marrow cells and transplanted them into irradiated mice. Hematopoietic disorders were observed only after secondary and tertiary transplants, arising within 22 and 16 weeks, respectively. All diseased mice carried the same leukemic clone, with a single vector copy integrated into and transactivating the murine gene Evi1 (ecotropic viral integration site-1) from both LTRs. The authors speculated that the insertional activation of the Evi1 transcription factor could have induced a preleukemic state, followed by a second cooperating event common to all subclones, and suggested a role for the reporter transgene in the leukemogenesis. Such role, however, was never confirmed in the clinical setting[160]. Except for this report, no other evidence of clonal dominance or neoplasia was uncovered at that time, probably due to a lack of systematic long-term follow-up in the murine studies and in low or non-persistent levels of gene transfer in the human clinical trials.

The X-SCID adverse event boosted a series of studies to evaluate Mo-MLV vector genotoxicity, both in the murine and in the human setting.

Systematic analysis of Mo-MLV integration pattern in natural or experimentally induced leukemias/lymphomas identified insertion sites recurrently associated with a malignant phenotype, *i.e.*, loci that are targets of retroviral integration in more than one tumor[161]. These were called "common retroviral integration sites" (CISs) and occurred in the vicinity of proto-oncogenes or other genes associated with cell growth and proliferation, the activation or deregulation of which is likely involved in the establishment and/or progression of neoplasia. To manage all data coming from multiple high-throughput insertional mutagenesis screening projects, a comprehensive Retroviral Tagged Cancer Gene Database (RTCGD) was created, containing the genomic position of each retroviral integration site cloned from a mouse tumor, the distance between it and the nearest candidate disease gene(s) and its orientation with respect to the candidate gene(s)[162]. The database became soon the standard reference in the field, allowing users to search both for CIS genes and unique viral integration sites or to compare the integration sites cloned by different laboratories using different models.

Some of the CISs included in the RTCGD have been also identified at relatively high frequency in the progeny of MLV-transduced hematopoietic cells in mice, nonhuman primates and humans. In most cases the CISs marked few dominant clones, more often with a non-malignant phenotype, which mainly contributed to the hematopoietic reconstitution. This suggests a "clonal dominance" model, where retroviral insertion into certain genes confers some growth and/or survival advantage to transduced progenitors, resulting in their *in vivo* amplification; such induced clonal expansion does not necessarily lead to malignant transformation of the affected cell clones.

The first suggestion of clonal dominance of hematopoietic stem cells triggered by retroviral gene marking was observed in cohorts of healthy mice in which a single or very few clones dominated hematopoiesis after serial bone marrow transplantation[15]. In both primary and secondary transplant recipients, dominating clones hosted insertions nearby CISs, proto-oncogenes or other signaling genes. Transcriptional deregulation by retroviral LTR was observed in all insertion sites analyzed. Mdsl/Evil locus was recovered several times both in primary and secondary recipients. The authors conclude suggesting a selection process by which preferential survival of long-term repopulating clones is triggered by insertional deregulation of genes that enhance their "fitness", without necessarily resulting in malignant transformation.

A high frequency of integrations within the Mdsl/Evil locus was also retrieved from non-human primate hematopoietic cells[16]. The authors analyzed 702 integration sites in Rhesus Macaques that underwent transplantation with autologous CD34$^+$ HSCs transduced with amphotropic Mo-MLV-derived retroviral vectors. Insertion in Mdsl/Evil region was detected 14 times in 9 animals, primarily in circulating granulocytes. All 9 animals had normal blood counts, with no evidence of leukemia, and a polyclonal hematopoiesis. The findings suggested again that, although insertion into the Mdsl/Evil locus as a single event impacted on engraftment or survival of primitive progenitors, it did not result in abnormal proliferation or differentiation.

Shortly thereafter, the first case of retroviral vector-associated neoplasia in a non-human primate was reported[163]. A Rhesus Macaque transplanted with MLV-transduced CD34$^+$ cells to express a reporter gene and a drug-resistant variant of the dihydropholate reductase gene developed an acute myeloid leukemia, five years after treatment. Tumor cells contained two vector insertions, one of which located in the first intron of the anti-apoptotic gene BCL2A1. The same two integrations were previously identified as dominant during the first year after transplantation, before becoming undetectable for the subsequent four years, and then re-emerging in the dominant clone contributing to myeloid hematopoiesis and to the fatal myeloid sarcoma. Out of 82 large animals treated and followed long-term, this was the only documented case of malignancy, but still it raised a note of caution that the vector-mediated insertional mutagenesis contribution to a neoplastic process may not be limited to the context of X-SCID gene therapy.

In vivo expansion of cell clones containing insertionally activated growth-promoting genes was also observed in the clinics, in two adults infused with genetically modified cells for the treatment of X-linked chronic granulomatous disease (CGD). The risk of insertional mutagenesis in this trial was estimated to be low, because the therapeutic gene (gp91^{phox}) was not expected to provide a survival or growth advantage to transduced cells, unlike the IL2Rγc gene delivered in the X-SCID trial. The distribution of gene-modified cells in the two subjects was studied over time, and became increasingly non-random in both subjects; the myeloid compartment was mainly affected. Integrations in three genetic loci emerged as predominant after 5 months, and increased up to > 80% of insertions retrieved from circulating transduced cells; these were the well-known MDS1/EVI1 locus, hosting 91 integrations, the related gene PRMD16 (36 insertions) and SETBP1 (7 hits). Both PRMD16 and SETBP1 were first identified as involved in leukemogenesis. PRMD16 (also known as MEL1, for MDS1/EVI1-like gene 1) is a PR domain-containing transcription factor highly related to EVI1; it was found in t(1;3)(p36;q21)-positive acute myeloid leukemia as a transcriptionally activated gene near the chromosomal breakpoint[164]. SETBP1 (SET binding protein) was identified as a novel protein binding to the acute undifferentiated leukemia-associated protein SET; SETBP1 could play a role in the mechanism of SET-related leukemogenesis and tumorigenesis, perhaps by suppressing SET function[165]. Over-expression of MDS1-EVI1 transcript was revealed in both patients, while PRMD16 and SETBP1 mRNAs were deregulated in one or the other subject. Notably, myeloid cell proliferation and differentiation was normal, suggesting that the expanded cells retained nearly physiological properties.

Genome-wide analysis of integration sites retrieved from ADA-SCID[114] and X-SCID[166,167] patients in three independent gene therapy trials revealed a substantially different scenario. Despite a clustering of integrations was observed in the proximity of CISs or near potentially oncogenic loci (among the others LMO2, RUNX1, BCL2, CCND2), no clonal outgrowth was detected *in vivo*. Likewise, there was no sign of clonal dominance in > 45 patients cumulatively treated with $>10^{11}$ retroviral vector-transduced T cells, although one fifth of the retroviral insertions affected the expression of neighboring genes[13].

Pursuing the idea that clonal dominance arises *in vivo* by amplification of those cells that host retroviral integrations conferring them a growth/survival advantage, Kustikova *et al.* have recently compiled an insertion dominance database (IDDb)[168]. Summarizing data from several laboratories, they developed a database of retroviral insertion sites detected in dominant clones contributing to phenotypically intact, mildly dysplastic and overtly malignant hematopoiesis of serially bone marrow transplanted mice. As reasonably expected, genes belonging to the IDDb were involved in proliferation, apoptosis and transcription regulatory networks, and some of them were already implicated in HSC biology.

4. Aim of the study

The aim of this thesis project is to characterize the integration patterns of gamma-retroviral and lentiviral vectors in the genome of human hematopoietic cells, and investigate the viral and cellular determinants of their target site selection. In this context, I first compared the integration patterns of Mo-MLV- and HIV-1-based vectors, and then I evaluated the role of viral LTRs, and of the transcriptional complexes binding to them, in targeting viral PICs to the cell chromatin. I chose an experimental setting strictly resembling the clinical standards for the gene therapy of monogenic blood disorders, *i.e.*, acute infection of $CD34^+$ hematopoietic stem/progenitor cells with Mo-MLV- and HIV-1-based vectors. The use of clinically relevant target cells transduced with the same vectors employed in ongoing trials is mandatory to transfer knowledge from the bench to the clinical practice, with the specific aim of assessing risks associated to gene transfer technologies and improving accordingly their safety and efficacy.

As reviewed in the previous sections, the occurrence of leukemia-like diseases in gene therapy patients treated with Mo-MLV-based retroviral vectors has raised safety concerns for the genotoxic risk associated to retroviral insertion into the human genome, especially in the context of long-living, self-renewing stem cells. Therefore several groups performed large-scale studies aimed firstly at describing the integration characteristics of different retroviruses in mammalian genomes and then possibly at understanding the molecular mechanisms underlying them. However, the most comprehensive studies, analyzing hundreds of integration sites at once, were mainly performed with lentiviral vectors, both in cell lines (SupT1[127], HeLa[12,121], H9[12], IMR-90[130], CEM[124], Jurkat[169]) and primary cells (peripheral blood mononuclear cells, PBMCs[11]) natural targets of HIV infection. Until recently, a single *in vitro* study collected a large number of insertions from HeLa cells infected with a Mo-MLV-derived vector[12]. In most other cases, MLV integration sites were retrieved *ex-vivo*, from the peripheral blood of human[4,114,166,167] and non-human primates[9,10,16,129] or from the bone marrow of mice[10,15,168], several weeks after transplantation of transduced hematopoietic stem cells. Aim of these studies was to evaluate the contribution of retrovirally-marked stem cells in the bone marrow repopulating dynamics and to assess the genotoxic risk associated to the gene therapy approach. Indeed, many of these studies showed the existence of MLV recurrent insertion sites near proto-oncogenes or other genes associated with cell growth and proliferation, deregulation of which may cause clonal amplification and/or malignant transformation of transduced progenitors *in vivo*. However, pre-transplant, MLV-infected hematopoietic cells were analyzed neither in mice nor in nonhuman primates and poorly characterized in humans (100 to 250 insertion sites analyzed). Hence, from these studies it was not possible to establish whether clonal dominance was entirely the result of *in vivo* selection, or if it was favored by the existence of highly preferred regions of retroviral integration that make clonal amplification more likely to occur. To answer this question, large numbers of integration sites must be retrieved from MLV-infected CD34$^+$ stem/progenitor cells after short-term culture periods, when clonal dominance induced by retroviral insertion cannot appear. The same can be tested for HIV vectors, whose integration pattern in the specific setting of human hematopoietic cells has not been deeply investigated so far. Given that lentiviral vectors are likely to replace gamma-retroviral vectors for a number of clinical applications, an assessment of their integration characteristics in the relevant cell context appears highly desiderable.

Apart from comparing the distribution of Mo-MLV and HIV-1 integrations in unselected human hematopoietic cells, a second question addressed in this thesis is whether transcriptional regulatory elements contained in viral LTRs exert any role in the integration site selection of both gamma-retroviral and lentiviral vectors. The rationale of such a question is that transcription and integration are intimately linked aspects of retroviral life cycle, and that each viral family has evolved a molecular strategy to target its integration in order to maximize the likelihood of survival and propagation to target cells. In case of acutely infectious gamma-retroviruses, this somehow involves integration in the proximity of gene regulatory elements and promoters, a strategy that probably ensures a productive interaction of the viral transcription unit with actively transcribed chromatin regions. In the case of lentiviruses, integration into active genes, but at a higher distance from transcription start sites, may be more permissive for the latent phase of the viral life cycle[§]. Following this idea, I have investigated whether viral PICs bind host transcription factors through their enhancers and regulatory elements, and whether these factors play a role in tethering PICs to active chromatin regions. The hypothesis was explored from both sides, by analyzing viral genetic determinants as well as the arrangement of transcription factor binding sites in the genomic regions flanking the retrieved integration sites. To my knowledge, there is no evidence so far rigorously documenting a role for viral LTRs and LTR interactors in the integration process.

[§] Such interpretation of lentiviral integration preferences implies that lentiviruses deliberately use latency as a survival strategy; the work by Siliciano seems to suggest that latency is rather an accident of infecting a CD4$^+$ T cell that is returning to a resting state170. Persaud D, Zhou Y, Siliciano JM, Siliciano RF. Latency in human immunodeficiency virus type 1 infection: no easy answers. J Virol. 2003;77:1659-1665.. Indeed, the persistence of HIV-1 is not dependent on latency, since the virus replicates continuously, and relentlessly evolves to escape from immune response. At present it remains controversial whether latency is a strategy for survival or whether it is not, and thereafter if the reactivation from latency is deliberate or if it is just a failure of cell silencing of invading genetic elements.

A single study (reviewed in section 3.2.4) has suggested a function for cellular transcription factors in the integration targeting of MLV-based retroviral vectors[121]. A significant enrichment of TFBSs was observed in the proximity of MLV insertion sites but not nearby HIV integrations when they where compared to matched random controls. However, the number of over-represented TFBSs was strongly reduced in MLV data sets when promoter sequences were used as controls instead of randomly generated genomic sites. The author concluded that general features of promoter regions, rather than specific TFs, act as tethering factors for MLV PICs, even though they do not exclude a possible effect of TFBSs beyond just marking promoters. In fact, the issue was not investigated deeply enough in this study to ascertain a role for transcription factors independently of promoters.

In most other cases, classical proteomic approaches, based on biochemical assays or genetic screenings, have been extensively used to identify host factors associated to viral PICs. These studies led to the identification of several proteins potentially affecting retroviral integration reaction, but only for one of them an unequivocal role was established; the ubiquitous co-activator LEDGF/p75 was demonstrated to act as a tethering factor for HIV-1 PICs to active chromatin regions, via direct binding of lentiviral integrase (see section 3.3.3 for details). For other cellular components of PICs, like hRad18, Ini-1, EED, BAF and HMGI(Y) (see section 2.1.4), such a tethering activity was not established. Most importantly, LEDGF/p75 activity is restricted solely to lentiviruses, while much less is known for the integrase of murine gamma-retroviruses, and the genetic and/or epigenetic determinants of their target site selection remain poorly understood. Recurrent MLV integration sites found in clones dominating the hematopoiesis of humans[4], primates[16] and mice[15,168] identify “stemness” pathways[158], further suggesting a link between integration site selection and transcription.

A deeper understanding of the mechanisms underlying target site selection by PICs would contribute both to the basic virology and to the gene therapy clinical practice, where the main interest is to develop viral vectors with the safest integration profiles.

5. Materials and methods

5.1 Retroviral vectors

MLV-derived gamma-retroviral vectors containing a green fluorescent protein (GFP) gene, an adenosine deaminase (ADA) cDNA or an IL2 receptor γ chain cDNA under the control of a wild type MLV LTR were the previously described LGSΔN[171], GIADA[1], and MFG-γ_c[2] vectors, respectively (designated in **Figure 11**, Section 6.1.1, as MLVa, MLVb, and MLVc). LGSΔN and GIADA vectors also contained a simian virus-40 (SV-40) internal promoter, driving the expression of a truncated nerve growth factor receptor (ΔLNGFR) or a neomycin resistance gene, respectively. The ΔU3 vector carried a GFP gene under the control of an U3-deleted (-413 to -62) LTR, and the same internal cassette of the LGSΔN vector, and was previously described as LGSΔN-ΔCAAT[171]. The SFFV-MLV vector expressed the GFP marker under the control of the spleen focus forming virus (SFFV) LTR, in the previously described pSF91 MLV vector backbone (a gift from C. Baum, Hannover)[172]. HIV vectors with wild type LTRs were the pHR2pptCMV-GFPwpre and the pHR2pptGSΔN LV vectors, retaining HIV-1 wild-type LTRs and driving the expression of GFP or ΔLNGFR under internal CMV or SV40 promoters. To generate the pHR2pptCMVGFPwpre construct, a pptCMVGFPwpre fragment from the pRRLsin-18.ppt.CMV-GFPwpre[173] vector was cloned into *Cla*I-*EcoR*I sites of pHR2MD-NGFR[174]. To obtain the pHR2pptGSΔN LV construct, the pHR2pptCMVGFPwpre vector was digested with *Bam*HI/*Eco*RI and ligated to a GFP-SV40ΔLNGFR cassette. The ΔU3-HIV[CMV] vector carried -418 to -18 deletion in the U3 region and an internal GFP expression cassette driven by the cytomegalovirus (CMV) immediate-early promoter, and was previously described as pRRLsin-18.ppt.CMV-GFPwpre[173]. The ΔU3-HIV[MLV] vector carried a -418 to -40 U3 deletion and was constructed by inserting an internal ΔLNGFR expression cassette driven by the full MLV-LTR into the pRRL.sin-40.GFP vector[174]. The MLV-HIV vector was built by inserting the PCR-amplified -413 to -62 fragment of the MLV U3 region at position -40 in the HIV LTR of the pRRL.sin-40.GFP vector[174], and adding an internal SV40-driven ΔLNGFR expression cassette.

RV vector supernatants were produced by transient transfection of the amphotropic Phoenix packaging cell line. Infectious particle titer was determined on K562 cells. The SFFV-MLV vector was VSV.G pseudotyped by transient co-transfection of 293T cells with an MLV *gag/pol* expression plasmid (a gift from C. Baum) and a VSV-G expression plasmid. Infectious particle titer was determined on 293T cells. The ADA γ_c receptor RV vectors were produced as amphotropic or GaLV envelope-pseudotyped particles from stable packaging cell lines, and titered as previously described[1,2]. VSV-G pseudotyped LV particles were prepared by transient co-transfection of 293T cells, collected and concentrated as described[175], and titrated on 293T cells or SupT1 cells.

5.2 Transduction of target cells

CB CD34$^+$ HSCs were purified from umbilical cord blood by magnetic sorting. Blood was harvested from the umbilical artery with heparised syringes, diluted 1:3 to 1:4 in phosphate-buffered saline (PBS), layered in 50 ml conical tubes above 15 ml of Ficoll (LymphoprepTM; Axis-Shield PoC, Oslo, Norway), and centrifuged (1,800 rpm, 30' at 4°C, brake off). Buffy coats containing mono-nuclear cells were collected and washed twice in cold PBS-BSA-EDTA buffer (PBS, 0,5% bovine serum albumin, 2mM EDTA, degassed). Red blood cells were lysed 10' in ice in ACK solution (NH_4Cl 0.15M, $KHCO_3$ 1mM, Na_2EDTA 0.1mM), and lysis is blocked by addition of medium containing fetal bovine serum (FBS). Cells were then incubated with an anti-CD34 antibody conjugated with magnetic beads and separated by positive selection with the CD34-MiniMACS cell separation kit (Milthenyi, Auburn, CA), following the manifacturer instructions. The phenotype of isolated cells was checked by flow cytometry analysis after staining with an RPE-conjugated anti-human CD34 antibody (Beckton Dickinson).

Before transduction with retroviral vectors, CD34$^+$ cells were stimulated for 24-48 hours at a density of 1 x 10^6 cells/ml in serum-free Iscove's modified Dulbecco's medium (IMDM) supplemented with 20% BIT serum substitute (Stem Cell Technologies; Vancouver, BC), 20 ng/ml human thrombopoietin, 100 ng/ml Flt-3 ligand (PeproTech; Rocky Hill, NJ), 20 ng/ml interleukin-6, and 100 ng/ml stem cell factor (R&D Systems; Minneapolis, MN). Cytokines are required to induce proliferation of HSCs and maintain their "stemness" throughout the infection period. Transduction with RV vectors was performed by spinoculation (3 rounds at 1,500 rpm for 45 min) in the presence of retroviral supernatants and 4-μg/ml polybrene. Transduction with LV vectors was performed by over-night incubation of CD34$^+$ cells with vector stocks at a multiplicity of infection (MOI) of 200 in the presence of 4-μg/ml polybrene. Transduction efficiency was evaluated by analysis of EGFP and/or ΔLNGFR expression by flow cytometry using a mouse anti-human NGFR antibody (Beckton Dickinson). Transduced cells were collected 5-12 days after infection.

BM- or PB-derived CD34$^+$ cells were purified from normal donors or SCID patients again by magnetic sorting, pre-stimulated for 24 hours in IMDM containing human serum, or serum-free X-Vivo10 medium, and a cytokine cocktail (FLT3-

ligand, SCF, TPO, IL-3), and transduced by three cycle-exposure to the GIADA1 or the γ_c receptor RV vector supernatant as previously described[1,2].

SupT1 cells were grown in RPMI 1640 (BioWhittaker) supplemented with 10% fetal bovine serum, and transduced with MLV-HIV viral stocks at an MOI of 25, in the presence of 8-µg/ml polybrene. After virus addition, cells were spinoculated for 1 hour (1,800 rpm, 4°C) and left at 4°C for another hour to ensure a synchronous infection. Cells were then transferred to a 37°C incubator and collected after 4 to 10 hours to analyze pre-integration complexes, or left in culture for 2 additional weeks for the analyses on integrated proviruses.

5.3 Sequencing, mapping and annotation of retroviral integration sites

Integration sites were cloned by linker-mediated PCR (LM-PCR) or linear amplification-mediated PCR (LAM-PCR), as described[12,176,177]. Briefly, genomic DNA was extracted from 0.5-5 x 10^6 infected cells and digested with *Mse*I and a second enzyme to prevent amplification of internal 5' LTR fragments (*Pst*I for RV vectors and *Sac*I/*Nar*I for LV vectors). An *Mse*I double-stranded linker was then ligated and LM-PCR performed with the following nested primers specific for the linker and the 3' LTR:

MLV: 5'-GACTTGTGGTCTCGCTGTTCCTTGG-3'

MLV nested: 5'-GGTCTCCTCTGAGTGATTGACTACC-3'

HIV: 5'-AGTGCTTCAAGTAGTGTGTGCC-3'

HIV nested: 5'-GTCTGTTGTGTGACTCTGGTAAC-3'.

PCR products were shotgun-cloned into the pCR2.1 TOPO vector (TOPO TA cloning kit, Invitrogen; Carlsbad, CA) and transformed into TOP10 competent cells, to generate bacterial libraries of integration junctions. Single white colonies were picked, inoculated into Luria Broth (LB) medium and grown at 37°C over-night. DNA was then extracted (NucleoSpin Plasmid kit, Macherey-Nagel; Düren, Germany) and sequenced using the M13rev primer (5'-CAGGAAACAGCTATGACC-3'; Primm srl DNA sequencing service, Milan, Italy). A valid integration contained the MLV or HIV nested primer, the entire MLV or HIV genome up to a CA dinucleotide and the linker nested primer. Sequences between the 3' LTR and the linker primers were mapped onto the human genome (UCSC Human Genome Project Working Draft, hg17) using Blat sequence alignment tool[178], requiring a ≥ 95% identity over the entire sequence length and selecting the best hit. The absolute genomic coordinates of the integration sites where defined as a result of the combination of genomic alignment and vector relative orientation data. Random genomic sequences originated by LM-PCR (genomic *Mse*I-*Mse*I, *Pst*I-*Mse*I, *Nar*I-*Mse*I or *Sac*I-*Mse*I fragments) were mapped by the same criteria, and used as experimental controls.

Insertion sites and experimental control sequences were annotated according to two different criteria. In the first annotation criteria (used for the entire section 6.1), sequences were classified as intergenic when occurring at an arbitrarily chosen distance of > 30 kb from any Known Gene (UCSC definition), perigenic when ≤ 30 kb upstream or downstream of a Known Gene, and intragenic when within the transcribed portion of at least one Known Gene. According to the second annotation criteria (used for the entire TFBS analysis, section 6.2), insertion sites were classified as "TSS-proximal" when occurring at a distance of ±5 kb from the TSS of any Known Gene, "intragenic" when occurring within the transcribed portion of at least one Known Gene > 5 kb from the TSS, and "intergenic" in all other cases. In both annotation criteria, whenever multiple transcript variants exist, the most represented and/or the longest isoform was chosen. Integration sites retrieved from published data sets[12,121] were re-mapped and annotated according to the same criteria.

Gene density analysis was performed using the Table Browser tool of the UCSC genome browser. For each integration, the number of Known Genes (a single isoform in case of multiple variants) contained in a range of 1 Megabase (Mb) around the insertion site was scored. For comparison, I also calculated the gene density of the entire genome, virtually dividing each chromosome in 1 Mb consecutive segments and computing the number of Known Genes contained in each fragment.

A genomic region was defined as an "hot spot" for retroviral integration according to criteria developed for defining cancer-related common insertion sites (CIS), with minor modifications[14,161]. To include borderline integrations, cutoff values were set at 36 kb for 2 insertions, 56 kb for 3 insertions and 104 kb for 4 or more insertions.

For all pairwise comparisons, I applied a two-sample test for equality of proportions with continuity correction (Rweb 1.03).

5.4 Gene expression profiling

The expression profile of CD34$^+$ cells was determined by microarray analysis. RNA was isolated from 1 to 2 x 10^6 CB- and BM-derived CD34$^+$ cells stimulated with cytokines according to the same protocols used for RV (CB- and BM-derived cells) or LV (CB-derived cells) vector transduction, transcribed into biotinylated cRNA, hybridized to Affymetrix HG-U133A Gene Chip arrays (Santa Clara, CA) and analyzed as previously described[13]. To correlate retroviral integration and gene activity, expression values from the CD34$^+$ cell microarrays were divided into 4 classes (*i.e.*, absent, low, below the 25th percentile in a normalized distribution, intermediate between the 25th and the 75th percentiles, and high above the 75th percentile).

5.5 Functional clustering analysis

Functional cluster analysis of genes targeted by retroviral integrations and from control sequences was performed using the DAVID 2.1 Functional Annotation Tool[179,180] (http://david.abcc.ncifcrf.gov). In the DAVID annotation system, a Fisher exact test corrected for multiple comparisons (DAVID's EASE score) is adopted to measure the level of gene-enrichment in Gene Ontology (GO) annotation terms with respect to a background population, and GO categories considered over-represented when yielding an EASE score < 0.05. A list of 417 cancer-associated CIS was obtained from the Mouse Retrovirus Tagged Cancer Gene Database[181], where murine genes were replaced with human homologs. Genes were also analyzed by the network-based Ingenuity Pathways Analysis tool (Ingenuity® Systems, www.ingenuity.com), to search for the most relevant molecular interactions, functions and pathways linking them. Gene identifiers were uploaded into the application, and mapped to their corresponding Focus Gene in the Ingenuity Pathways Knowledge Base, a structured and context-rich knowledge base manually compiled from scientific literature. Gene networks were algorithmically generated based on the direct or indirect interaction between Focus Genes. The Functional Analysis of each network identified the biological functions and/or diseases that were most significant to the genes in the network (Fischer's exact test).

A list of 417 cancer-associated CISs was obtained from the Mouse Retrovirus Tagged Cancer Gene Database[181], where murine genes were replaced with human homologs. Two different sources were used to define a list of 596 human proto-oncogenes; the UNSW Embryology DNA-Tumor Suppressor and Oncogene

Database[182] contains genes that are classified as tumor suppressors or oncogenes in the Online Mendelian Inheritance in Man (OMIM) database; the Tumor Gene Database[183] is a broad directory of genes mutated in cancers, proto-oncogenes and tumor suppressor genes.

5.6 Transcription factor binding site analysis

TFBS analysis was carried out on genomic sequences encompassing each integration site with ±1.0 kb of sequence length. Based on the TSS-proximal/intragenic/intergenic annotation of each integration site, we grouped data sets that did not significantly differ from each other (two-sided test on equal proportion) into seven groups of integration preferences, and generated the same number of random weighted control groups of sequences reproducing, in proportion, the specific integration preference of each vector. Each fitted background was composed of 10,000 sequences of 2.0 kb in length derived from 100,000 randomly generated integration sites throughout the genome (**Table 5**). TFBS enrichment analysis was performed with Clover[184], with dinucleotide randomization and motif *p*-value threshold set to 0.05. Clover program is able to screen a set of DNA sequences against a precompiled library of motifs and assess which, if any, of the motifs are statistically over- or under-represented in your data sets when compared to a background group of sequences. A precompiled library of 123 TFBSs, described as positional-weight matrices, were here obtained from the Jaspar Core 2005 database of experimentally validated motifs[185]. The appropriate weighted background was paired with each sequence set. TFBSs having a global *p*-value < 0.05 were considered as significantly enriched in the test sequences and selected for analysis. Motif likelihood ratio was used for cluster analysis and PCA. The number of over-represented TFBSs per sequence was plotted as a boxplot to display differences between the data sets without making any assumptions of the underlying statistical distribution. The spacings between the different parts of the box indicate the degree of dispersion and skewness in the data, and identify outliers (these were omitted in **Figures 23** and **29** for better graphical visualization). Each box is built starting from five numbers: the minimum (smallest observation), the first quartile (which cuts off the lowest 25% of the data), the median (middle value), the third quartile (which cuts off the highest 25% of the data), and the maximum (largest observation).

Pattern discovery among and within different groups was performed with a two-way hierachical cluster analysis on motif likelihood values, using the Euclidean distance as a similarity measure between clusters. Before analysis data were scaled on motif columns. To add robustness to the dendrogram analysis and reduce test bias, we applied an approximately unbiased (AU) test on column dendrograms, sampling them with 10,000 multiscale bootstrap replicates[186]. Nodes having an Approximately Unbiased (AU) p-value > 0.95 were scored as significant and stable nodes.

As an additional tool to find patterns of TFBSs within our data sets we performed a Principal Components Analysis (PCA). Data were again scaled on motif columns, *i.e.*, Jaspar enriched motifs were considered as vectors, assuming a given likelihood value for each sequence (the analysis was unsupervised, *i.e.*, motifs coming from different data sets were not distinguished). A correlation matrix was built calculating the covariance between all possible pairs of motifs, and eigenvectors and eigenvalues[§] for this matrix were calculated. Eigenvectors were

[§] In mathematics, given a squared matrix, an eigenvector of that matrix is a nonzero vector which, when multiplied by the matrix, changes in length, but not in direction. The amount by which the original vector is scaled after the multiplication represents the eigenvalue for that eigenvector. Eigenvectors can only be found for square matrices, in a number equal to the number of rows and columns of that matrix. All then ordered by eigenvalue, highest to lowest, obtaining the components in order of significance (the component with the highest eigenvalue explaining the greatest percentage of variance in the system). The principal components were then used as new spatial coordinates to plot the original data sets, to obtain the plots of **Figures 25, 31** and **34**. For each bidimensional plane considered, only motif vectors having a projection on it longer than cos ($\pi/4$) ~ 0.707 were considered as relevant. Having all vectors a length = 1 in poly-dimensional space, if their projection is longer than 0.707, the angle between the motif vector and the plane is less than $\pi/4$ (45 degrees), meaning strong correlation between the motif and the plane of that principal component. Relevant motifs were also plotted (loadings plots of **Figures 25, 31** and **34**).

Analysis of conserved TFBSs was performed using the TFBS Conserved Track at UCSC Genome Browser, which includes binding sites conserved between the human and mouse or rat genome alignment (188 human matrices from the TRANSFAC Matrix Database v 7.0). After determination of the total count of matrices that match in each 2.0 kb test sequence, random and matched fitted backgrounds, a Fisher exact test (two-sided, confidence level = 0.95) was used to determine statistical significance. The STAMP tool-kit[188] was used to match JASPAR and TRANSFAC matrices using default parameters.

All statistical analyses were performed using the R language and environment for statistical computing and graphics version 2.6.2 (http://www.R-project.org) and several contributed packages. Hierarchical clustering used the *pvclust* package; PCA analysis used *ade4*; parallel processing was implemented using the *snow* package. Stats package was used for the other analyses.

the eigenvectors of a matrix are perpendicular, *i.e.*, at right angles to each others, no matter how many dimensions you have.

5.7 Southern and Western blot analysis

Southern Blot analysis was performed on cytoplasmic and nuclear DNA extracted from MLV-HIV infected SupT1 cells 4-7-10 hours, and 14 days after transduction. For each time point, 1 x 10^6 cells were lysed 10 minutes on ice in 200 μl of 5 mM Pipes, pH 8.0, 85 mM KCl, and 0.5% Nonidet P-40 (ChIP cell lysis buffer). Lysates were centrifuged 10' at 13,000 rpm, 4°C, and supernatants were saved as "cytoplasmic fractions". Pelletted nuclei were washed once in ice-cold PBS and resuspended in 200 μl PBS. DNA was then extracted from cytoplasmic fractions and nuclei by the QIAamp DNA Blood Mini Kit (QIAGEN), eluted in 60 μl of DNase-free water and loaded, undigested, on a 0.8% agarose gel. After an over-night run, the gel was transferred to a nylon membrane (Hybond-N, Amersham) by Southern capillary transfer, probed over-night with 2 x 10^7 dpm of a ^{32}P-labeled GFP probe, and exposed for 72 hours at -80°C to X-ray film.

For Western Blot analysis, SupT1 cells were lysed in buffer I (10 mM Hepes pH 7.9, 10 mM KCl, 0.1 mM EDTA, 1 mM DTT, and protease inhibitors) on ice, at a concentration of 100 x 10^6 cells/ml. After 15' incubation, Nonidet P-40 was added to a final concentration of 0.5%. Lysates were vortexed for 10'', kept on ice for other 10', and centrifuged 30'' at 13,000 rpm, 4°C. Supernatants were saved as cytoplasmic protein extracts. Pelletted nuclei were resuspended in the same volume of buffer II (10 mM Hepes pH 7.9, 0.6 M NaCl, 1.5 mM $MgCl_2$, 0.1 mM EDTA, 0.5 mM DTT, 5% glycerol and protease inhibitors), incubated on ice 30' and vortexed several times. Nuclear lysates were cleared by centrifugation at 13,000 rpm, 30', 4°C. Proteins from cytoplasmic and nuclear fractions were diluted in Bradford reagent and quantified by spectrophotometer analysis. 50 μg of proteins were run on 8% SDS-polyacrylamide gel and transferred to a nitrocellulose membrane (Hybond-

ECL, Amersham). Aspecific sites on the membrane were blocked by 1-hour incubation at room temperature in 5% milk-TBST (0.1 M Tris pH 8.0, 150 mM NaCl, 0.1% Tween-20) and incubated for 2 hours at room temperature with the primary antibody, diluted in 5% milk-TBST (from 1:100 to 1:2000, depending on the antibody). After several washes in TBST, the appropriate peroxidase-labeled secondary antibody was added, again diluted in 5% milk-TBST, and incubated at room temperature for 45'. Following 2' of ECL detection (Hyperfilm, Amersham), membranes were exposed to films for 1 to 15 minutes, depending on the primary antibody. Primary antibodies used were rabbit or goat polyclonal IgGs directed against AML-1 (sc-286799), CBFB (sc-10779), C/EBPα (sc-9314), C/EBPβ (sc-150), C/EBPδ (sc-636), Ets-1/Ets-2 (c-112), NF-1 (sc-870), and YY-1 (sc-281); secondary antibodies were donkey anti-rabbit (sc-2077) or anti-goat (sc-2020) HRP-conjugated IgGs, all from Santa Cruz Biotechnology.

5.8 RNA extraction and RT-PCR analysis

Total cellular RNA was extracted from 5 x 10^6 SupT1 cells, 10 hours after infection with the MLV-HIV vector, using the QIAamp RNA Blood Mini Kit (QIAGEN). 500 ng of extracted RNA were loaded on a denaturing 1% agarose gel to check for RNA integrity. As a positive control for RT-PCR analysis, MLV-HIV genomic RNA was isolated from ~10^7 virions using the NucleoSpin RNA Virus kit (Macherey-Nagel). Both cell- and virion-isolated RNAs were then digested with 20 μg/ml DNaseI (bovine pancreatic deoxyribonuclease I, SIGMA), in the presence of 0.1 mM DTT and 20 mM $MgCl_2$, 1 hour at 37°C. The enzyme was heat-inactivated at 65°C, 5', and the RNA samples used to set up the retrotranscription reaction (SuperScript III kit, Invitrogen). A specific oligo annealing on the 3' end of the GFP mRNA was used to prime retrotranscription instead of the classical random examers/oligo-dTs, to reduce aspecific amplifications in the following PCR. Samples with no RT enzyme were processed in parallel with real samples to control for residual DNA contamination. One tenth of the RT reactions were then subjected to PCR, with GFP for and rev primers internal to the oligo used for the RT reaction. PCR products were finally run on a 1% agarose gel and stained with ethidium bromide for visualization.

GFP RT-oligo: 5'-GTTACTTGTACAGCTCGTCCATGCC-3'

GFP for: 5'-CACATGAAGCAGCACGACTT-3'

GFP rev: 5'-TGCTCAGGTAGTGGTTGTCG-3'

5.9 Chromatin immunoprecipitation assay

ChIP assays were performed essentially as already described[189]. Chromatin was prepared from 30-50 x 10^6 SupT1 cells transduced with the MLV-HIV vector (MOI = 25) 10 hours or 14 days after infection. Cells were cross-linked with 1% formaldehyde-containing medium, 10 minutes at room temperature. Cross-linking was blocked by addition of PBS-glycine to a final concentration of 0.125 M. Cells were washed twice with ice-cold PBS, centrifuged at 4500 rpm for 10 minutes at 4 °C, and resuspended in cell lysis buffer (5 mM Pipes, pH 8.0, 85 mM KCl, and 0.5% Nonidet P-40) containing protease inhibitors (10 μg/ml aprotinin, 10 μg/ml leupeptin, and 1 mM PMSF) and kept on ice for 15'. Lysates were then homogenized several times with a Dounce homogenizer (tight pestle), and the resultant homogenates were centrifuged at 4500 rpm for 10' at 4 °C to pellet the nuclei. Nuclear pellets were resuspended in sonication buffer (50 mM Tris-HCl, pH 8.1, 10 mM EDTA, 0.1% SDS) containing protease inhibitors and PMSF and kept on ice for 20 min. Nuclear extracts were sonicated to obtain DNA fragments ranging from 200 to 1,500 bp in length and centrifuged at 13,000 rpm for 10 min at 4 °C.

The equivalent of 1-2 x 10^6 cells was immunoprecipitated over-night with 4 ug of rabbit anti-AML1, anti CBF-B, anti-Ets1/2, and anti YY1 antibodies (sc-28679, sc-10779, sc-636, and sc-281, respectively, from Santa Cruz Biotechnology) in RIPA buffer (10 mM Tris-HCl pH 8, 1 mM EDTA pH 8, 0.5 mM EGTA, 1% Triton X-100, 0.1% SDS, 0.1% Na-deoxicholate, 140 mM NaCl). Immunoprecipitations with rabbit anti-HA-probe (Y-11) (sc-805, Santa Cruz Biotechnology) and with no antibody were included as controls. Supernatant from the no-antibody sample was saved as the total input chromatin. Immunoprecipitated DNA was analyzed by PCR with primers amplifying the entire U3 region of the MLV-HIV LTR (for: 5'-CTGGAAGGGCTAATTCACTCC-3'; rev: 5'-CCCAGTACAAGCAAAAAGCA-3'). A 0.1% dilution of the total input was amplified to evaluate the relative enrichment of a specific antibody with respect to the control antibody and the no-antibody samples.

6. Results

6.1 Integration preferences of Mo-MLV and HIV-1-based retroviral vectors in human CD34$^+$ HSCs

The first part of my PhD project was designed to investigate the general integration properties of those retroviral vectors that are used for the gene therapy of human hematopoietic disorders. These are classical Mo-MLV-based vectors, expressing the gene of interest under an intact LTR, or the new generation of self-inactivating vectors, both with Mo-MLV- and HIV-1-derived backbones, where U3 regulatory elements have been deleted to abolish transcription initiation from proviral LTRs. The latter are likely to replace the wild type LTR vectors to reduce the genotoxic risks associated to insertional deregulation of tumor-related genes. With the specific intent to compare the integration patterns of MLV-based and HIV-based vectors in hematopoietic cells on a genome-wide scale, no distinction was made in this first part between vectors with wild type and deleted LTRs, and their integration sites were unified and analyzed as a whole. This allowed me to collect sufficiently large numbers of insertion sites and perform a series of otherwise impossible and/or statistically unreliable analyses, such as the characterization of integration hot spots.

6.1.1 Genome-wide analysis of retroviral integration preferences in human CD34$^+$ HSCs

Human CD34$^+$ HSCs were isolated from umbilical cord blood (CB) pools, bone marrow (BM) from patients with ADA-SCID and X-SCID, or peripheral blood (PB) from a healthy donor. After 24 to 48-hour pre-activation with cytokines, CB

CD34$^+$ cells were transduced with Mo-MLV-derived gamma-retroviral (RV) or HIV-1-derived lentiviral (LV) vectors carrying a green fluorescent protein (GFP) reporter gene and either a wild type or an U3-deleted (ΔU3) LTR. BM CD34$^+$ cells were transduced with RV therapeutic vectors expressing either the adenosine deaminase (ADA) enzyme[1] or the IL-2 receptor γ chain[2] under the control of a wild type LTR. PB CD34$^+$ cells were again transduced with the IL2Rγc therapeutic vector. Transduction efficiency ranged from 15% (ΔU3-MLV) to more than 90% (ΔU3-HIV), depending on the vector and target cell type, and remained stable throughout the culture period, as assessed by flow citometry analysis. DNA was extracted 1 to 12 days after infection, from cells that underwent 1 (BM and PB samples) to 5-6 (CB samples) cell doublings in culture. The short-term culture period was a fundamental requirement to exclude clonal outgrowth and selection of cells harboring insertions activating growth-promoting genes. Vector-genome junctions were amplified by linker-mediated (LM-) or linear amplification-mediated (LAM-) PCR approaches adapted to different vector types, and cloned into bacterial libraries that were then sequenced to saturation. Sequences between the 3' LTR and the linker primers were mapped onto the human genome (UCSC Human Genome Project Working Draft, hg17) using Blat[178], requiring a $\geq$ 95% identity over the entire sequence length and selecting the best hit. Cumulatively, I mapped 1,030 RV and 849 LV integrations in CB- PB- or BM-derived CD34$^+$ cells. A total of 595 RV integrations were retrieved from CB cells transduced with wild type (395) or ΔU3 (200) LTR vectors, both expressing the GFP from viral LTRs and a truncated form of the nerve growth factor receptor (ΔLNGFR) reporter gene under the control of an internal simian virus 40 (SV40) promoter (MLVa and ΔU3-MLV vectors in **Figure 11**). GFP expression from the ΔU3-MLV LTR was in fact barely detectable, due to a very low residual activity of the TATA box, which is retained in the vector configuration. 435 RV integrations were obtained from BM cells transduced with wild-type LTR vectors expressing ADA or IL2Rγc (MLVb and MLVc vectors, respectively, in **Figure 11**) therapeutic genes.

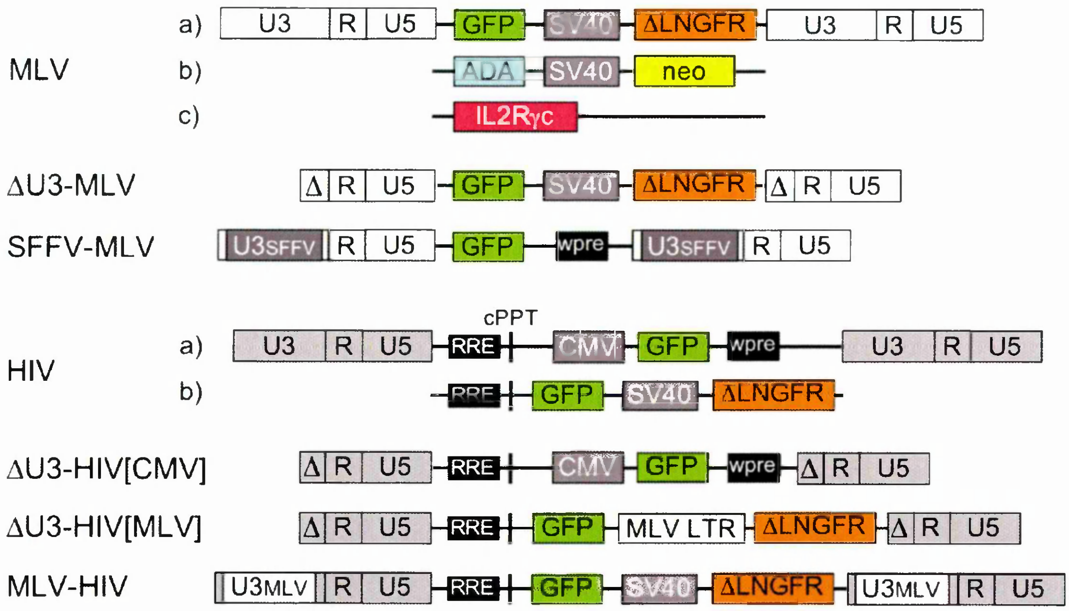


Figure 11. Schematic representation of RV and LV vectors with wild type or modified LTRs. For each of the vectors used for CD34$^+$ HSCs transduction, the LTR composition and the internal structure are depicted. RV LTRs are indicated by white boxes, LV LTRs by grey boxes. U3, R and U5 regions are specified for each LTR. Δ stands for partial deletion of the U3 element. $U3_{SFFV}$ and $U3_{MLV}$ are the U3 elements of the spleen focus-forming virus and of the Mo-MLV respectively. Internal expression cassettes, when present, are also schematized, consisting of an internal promoter (CMV: cytomegalovirus immediate-early promoter; SV40: simian virus 40 promoter; MLV LTR: internal Mo-MLV complete LTR) driving the expression of a marker gene (GFP: green fluorescent protein; ΔLNGFR: truncated nerve growth factor receptor; neo: neomycin resistance gene). The three RV vectors cumulatively called 'MLV' (a, b and c) differ in terms of transgene and internal structure but possess identical, Mo-MLV wild type LTRs, and were therefore considered as a single vector when performing integration site analysis. MLVb and MLVc vectors are the therapeutic vectors used for the gene therapy of ADA[1] and X-SCID[2], respectively. Similarly, the two LV vectors cumulatively called 'HIV' (a and b) carry a different internal cassette but have identical, HIV-1 wild type LTRs, and were again considered as a single vector for integration analysis. Some characterizing elements of lentiviral vectors are also depicted (RRE: Rev-responsive element; cPPT: central polypurine tract). The woodchuck hepatitis virus post-transcriptional regulatory element (wpre) was inserted in some vectors to augment viral titers.

All LV integrations were obtained from CB cells transduced with wild type (404) or ΔU3 (445) LTR vectors, expressing GFP and/or ΔLNGFR from internal SV40 or cytomegalovirus (CMV) promoters (HIVa-b and ΔU3-HIV[CMV] vectors in **Figure 11**).

Of the 1,030 RV integrations, 16.7% were found in an intergenic position, 55.0% within the transcribed portion of at least one gene and 28.3% at a distance of 30 kb or less upstream or downstream of one or more genes (**Table 1**; the complete list of sequences has been deposited at GenBank, with the accession numbers ER916114 to ER918350). Among LV integrations, 148 (17.4%) were in an intergenic position, up to 609 (71.7%) in an intragenic position and 92 (10.9%) in a perigenic position. Conversely, a collection of 798 control sequences randomly cloned by LM-PCR contained 369 (46.2%) intergenic, 308 (38.6%) intragenic, and 121 (15.2%) perigenic sequences. Compared to controls, RV vectors showed a preference for intragenic and perigenic integration, while LV vectors showed a much higher preference for intragenic positions. All differences were statistically significant ($p < 0.001$, 2-sample test for equality of proportions with continuity correction). RV general integration preferences were similar in CD34$^+$ and HeLa cells, as indicated by the re-analysis of 869 insertions retrieved from a previous published collection[12] (**Table 1**).

I then assessed the position of integrated proviruses with respect to all genes (UCSC track of Known Genes) found in an interval of 30 kb around each insertion site ("vector-gene interactions" in **Figure 12**). Compared with randomly cloned control sequences, a significant clustering around transcription start sites (TSSs) was observed for RV but not for LV vectors. The validity of the experimentally generated control sequences was confirmed comparing their distribution with that of

65,000 computer-generated random sequences[9]; the two distributions resulted almost indistinguishable. Overall, approximately 30% of the total RV vector-gene interactions were within 10 kb from the TSS of Known Genes, compared with 16.1% for LV vectors (p < 0.001; **Table 1**; **Figure 12**). The RV general integration preferences were similar in CD34$^+$ and HeLa cells, as indicated by parallel analysis of 869 insertions from a previously published collection[12] (**Table 1**).

Table 1. Retroviral integration site distribution in human CD34$^+$ HSCs

	Intergenic (%)	**Intragenic (%)**	**Perigenic (%)**	**Total hits**	**±10 kb from TSS (%)**	**Vector/gene interactions***
CD34$^+$ cells						
RV all	16.7	55.0	28.3	1,030	29.3	1,517
LV all	17.4	71.7	10.9	849	16.1	1,241
Controls	46.2	38.6	15.2	798	9.1	902
RV hot spots	16.0	56.6	27.4	219	22.2	302
LV hot spots	8.6	81.4	10.0	70	13.2	114
Control hot spots	36.4	59.1	4.5	22	13.0	23
HeLa cells						
RV	18.8	48.1	25.5	869	26.1	1,219
RV hot spots	16.5	53.2	30.3	109	27.3	165

Distribution of RV and LV integration sites unambiguously mapped in unselected CB- and BM-derived CD34$^+$ HSCs, and RV integrations in HeLa cells from a previously published collection[12]. Integrations (total hits) were distributed as inside (intragenic), outside (intergenic), or at a distance of <30 kb upstream or dowstream (perigenic) from Known Genes (UCSC annotation). Insertions at a distance of ±10 kb from transcription start sites (TSS) are indicated as percentage of the total vector/gene interactions. Control sequences were obtained from a randomly cloned library of SacI/NarI/PstI/MseI-restricted, LM-PCR-amplified human CD34$^+$ cell DNA.

*Total number of genes within 30 kb from individual hits + intergenic hits.

In CD34$^+$ cells, RV integrations showed a significant preference for gene-dense regions: more than 60% of proviruses were found in genomic regions containing 6 to 20 Known Genes per megabase (Mb) with a peak of 35% at a density of 6 to 10 genes/Mb. Conversely, more than 60% of control sequences mapped to regions with a gene density of less than 5 genes/Mb ($p < 0.001$, **Figure 13A**). On the contrary, LV integrations followed a distribution within regions of different gene density more similar to that of the control sequences and of the human genome, and different from that of RV ($p < 0.001$, **Figure 13B**).

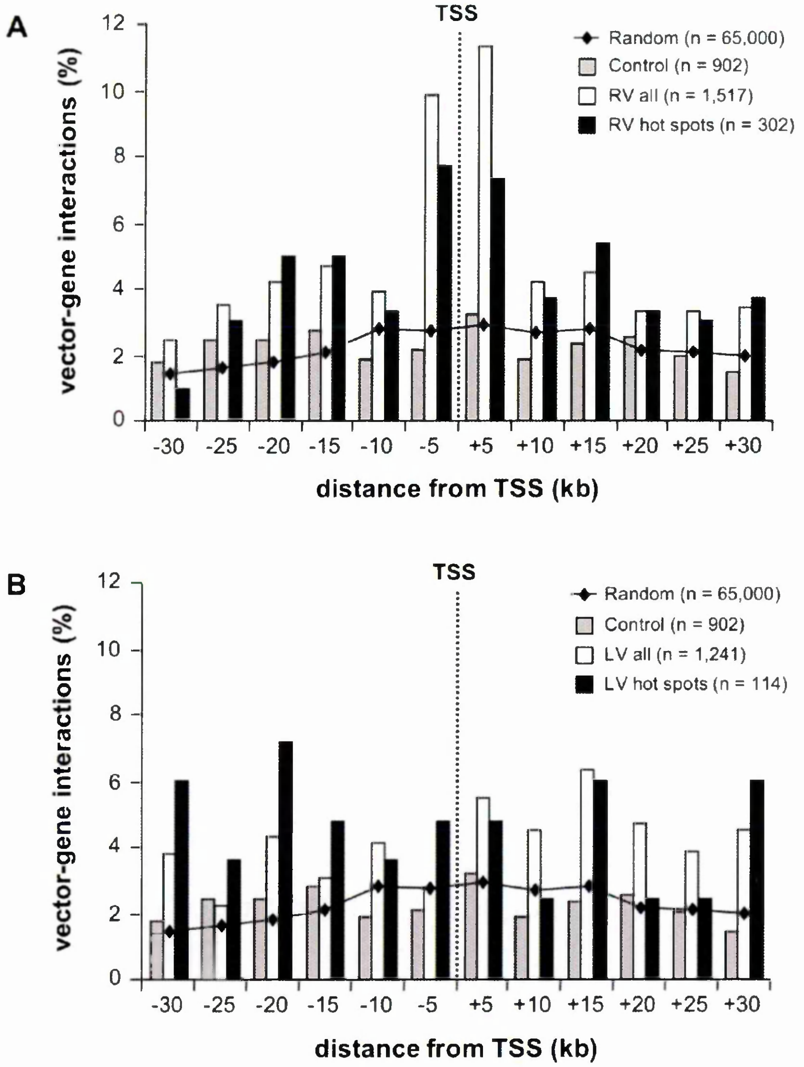


Figure 12. Retroviral integration and transcription start sites. Distribution of RV (**A**) and LV (**B**) integration sites in human CD34$^+$ HSCs within an interval of 30 kb upstream or downstream of the transcription start site (TSS) of Known Genes (UCSC track, considering only 1 isoform/gene). The bars show the percentage of distribution in each 5-kb interval of retroviral insertions, insertion hot spots, and control sequences. The line shows the distribution of 65,000 *in silico*-generated random insertion sites[9]. *n* values indicate vector-gene interactions, *i.e.*, the total number of genes within 30 kb from individual insertions plus intergenic insertions.

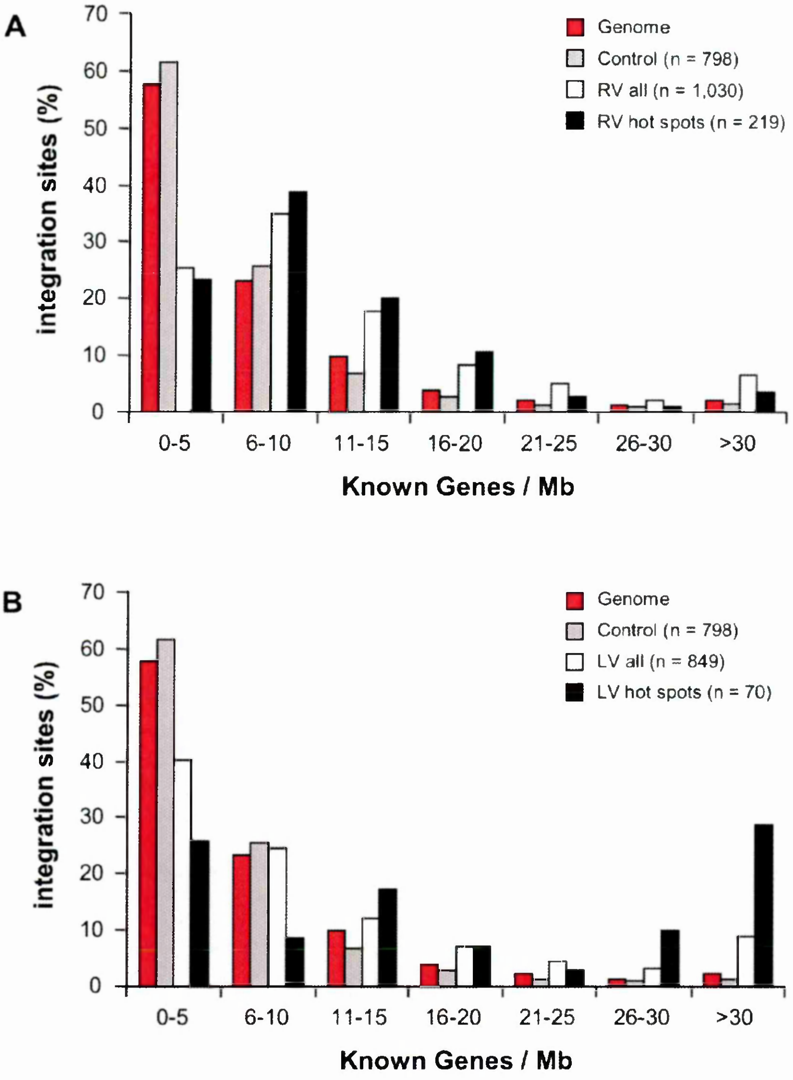


Figure 13. Retroviral integration and gene density. Integration sites and integration hot spots of RV **(A)** and LV **(B)** vectors in CD34$^+$ cells are plotted according to the number of Known Genes contained in a range of 1 Mb around each insertion site, in intervals of 5 genes/Mb. Grey bars indicate the distribution of control sequences. Red bars represent the frequency of 1-Mb segments in the human genome for each gene density interval. *n* values indicate the number of independent hits in each group.

To confirm also in hematopoietic cells the elsewhere observed correlation between gene activity and integration site selection[11,127], I used the results of Affymetrix HG-U133A gene expression arrays already available in my laboratory. These were performed on both CB- and BM-derived CD34$^+$ samples activated in culture with cytokines, the same conditions used for my RV and LV transductions, and therefore virtually represent the transcription profile of CD34$^+$ cells at the moment of PIC entry into the cell nucleus. **Figure 14** shows that approximately 60% out of 1,571 probesets representing 866 genes hit by a RV vector detected a transcript in activated CD34$^+$ cells; among them, 13% were classified as lowly abundant, 30% as intermediately abundant, and 17% as highly abundant. This was significantly different from what observed in the whole microarrays, where 45-47% of all the probesets had a "present" call (percentages were slightly different between CB- and BM-derived cells), with a 11-12%, 23%, and 11-12% breakdown in the 3 abundance classes. With the exception of the lowest expression class, all differences were statistically significant ($p < 0.001$), indicating that RV vectors integrate preferentially into genes active in CD34$^+$ cells at the time of transduction, and particularly in the fraction of genes expressed at higher levels. A similar correlation with gene activity was also observed for genes targeted by LV vectors (**Figure 14B**); 56% of 1,346 probesets associated to 757 hit genes detected a transcript in activated CD34$^+$ cells, with a 13%, 31% and 12% breakdown in the 3 abundance classes. Compared with the whole microarray, the fraction of probesets with a present call was significantly higher (56% *vs.* 46%; $p < 0.001$), but the difference was accounted for essentially by the intermediately abundant transcripts (31% *vs.*23%, $p < 0.001$). This indicates that LV vectors tend to integrate into active genes in CD34$^+$ cells but have no specific preference for genes expressed at high levels when compared with RV vectors ($p < 0.001$).

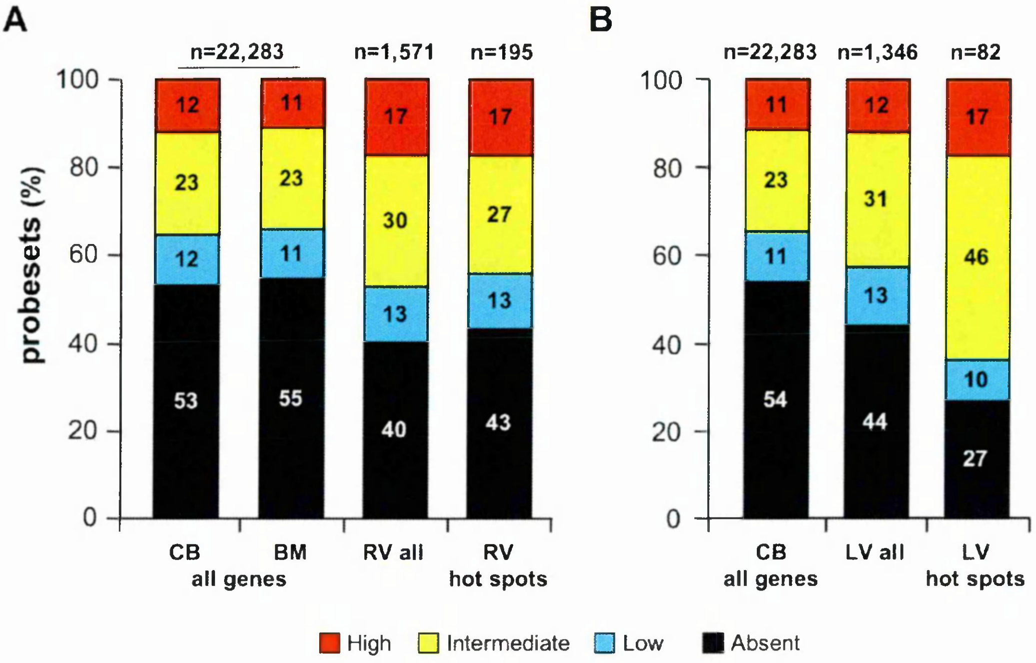


Figure 14. Retroviral integration and gene activity. The bars show the distribution of expression values from Affymetrix HG-U133A microarrays of cytokine-stimulated CD34$^+$ cells. The correlation between retroviral integration and gene activity was performed dividing probeset expression values from the microarray into 4 abundance classes: absent (black), low (below the 25th percentile in a normalized distribution, blue), intermediate (between the 25th and the 75th percentile, yellow), and high (above the 75th percentile, red). **(A)** The first 2 bars (all genes) show the distribution of more than 16,000 genes on the whole microarray of CB- or BM-isolated CD34$^+$ cells activated in the same conditions used for RV transduction; the other 2 bars represent the expression values of the sole genes targeted by RV integrations (all) or by integration hot spots (RV hot spots), obtained from a weighted mean of the CB and BM microarray values. **(B)** The first bar (all genes) shows the distribution of the more than 16,000 genes on the microarray of CB-derived CD34$^+$ cells activated as for LV transduction; the other 2 bars represent the expression values of the sole genes targeted by LV integrations (all) or by integration hot spots (LV hot spots). *n* values indicate the number of probesets associated to each group of genes.

6.1.2 Genes regulating cell growth and proliferation are preferential targets of retroviral integration

To understand which functions were associated to genes hit by retroviral integrations, I performed a classification of target genes following Gene Ontology (GO) criteria. The GO project provides vocabularies and classifications that cover several domains of molecular and cellular biology, freely available for community use in the annotation of genes, gene products and sequences[190]. The functional classification of genes hit by RV and LV vectors in CD34$^+$ cells showed statistically significant biases towards several gene categories (**Figure 15**). In particular, genes involved in the establishment and/or maintenance of chromatin architecture, signal transduction, and cell cycle were significantly more represented in the collection of genes hit by RV integrations compared with their expected frequency in the human genome (EASE score < 0.005). Genes involved in chromatin remodeling and phosphorylation were hit at a higher-than-expected frequency also by LV vectors (EASE score < 0.0005 and < 0.005, respectively), particularly those with serine/threonine kinase and GTPase activity (EASE score < 0.0005). Two additional categories (transcription and apoptosis) were over-represented in genes hit by RV and/or LV vectors, but at less significant levels (EASE score < 0.05).

Similar results were obtained performing a functional annotation of target genes by the network-based Ingenuity pathways analysis (IPA) tool (**Figure 16**). IPA annotation software is based on the Ingenuity Pathways Knowledge Base (IPKB), a database that models functional interactions between genes/gene products, manually compiled from the full text of articles published in peer-reviewed journals. IPA analysis indicated that genes involved in cell signaling, cell growth/proliferation, cell death, cancer, and hematopoietic system development were significantly over-represented in the collection of RV and/or LV integrations with respect to genes annotated in the IPKB software ($0.005 < p < 0.05$). I chose therefore these categories to carry out a direct frequency comparison between RV and LV target genes and our control gene list (the complete lists of genes used for the GO and the IPA analyses are contained in **Appendix 1**). Genes involved in cell signaling, growth/proliferation, and death were over-represented in both RV and LV integrations with respect to control sequences ($p < 0.001$, **Figure 16**), while genes involved in hematopoietic and immune system development, immune response and cancer were specifically over-represented in RV but not LV integrations ($p < 0.001$, **Figure 16**).

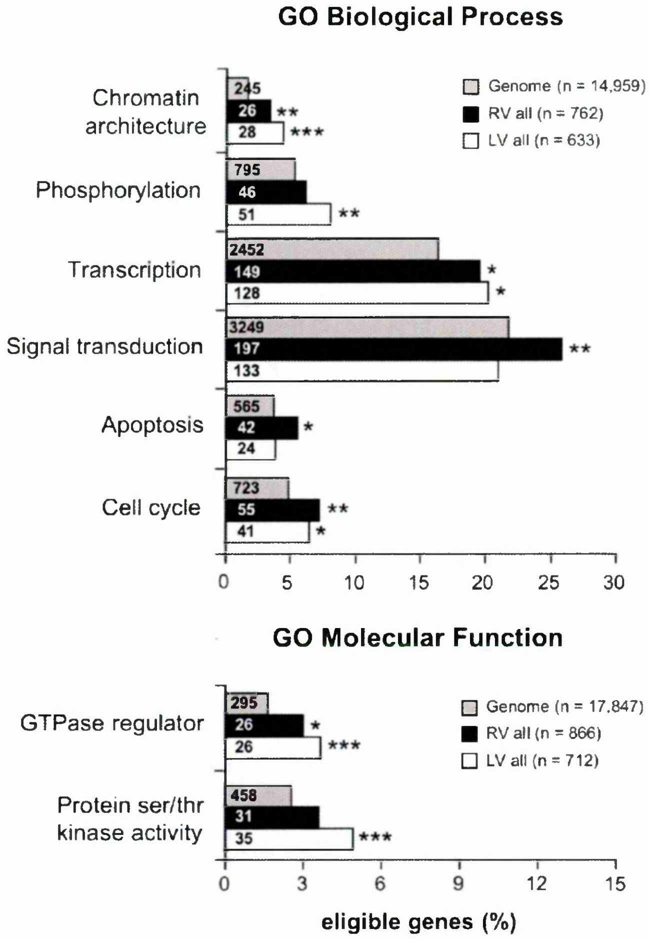


Figure 15. Retroviral integration preferentially targets genes regulating cell growth and proliferation. GO analysis of genes targeted by retroviral integration in CD34$^+$ cells. Genes identified as targets of RV and LV integration were analyzed for significant functional clusters with the DAVID 2.1 software. Functional categories derive from the GO-Biological Process (establishment and/or maintenance of chromatin architecture, phosphorylation, transcription, signal transduction, apoptosis, cell cycle) and the GO-Molecular Function (GTPase regulator activity, serine/threonine kinase activity) classifications. Bars indicate the number of integration target genes annotated within the given category out of *n* genes eligible for each analysis. Asterisks denote the significance level of over-representation of any given category with respect to the human genome, used as background population (***EASE score < 0.0005, **EASE score < 0.005, *EASE score < 0.05). The number of gene identifiers annotated within each functional category is indicated in the bars.

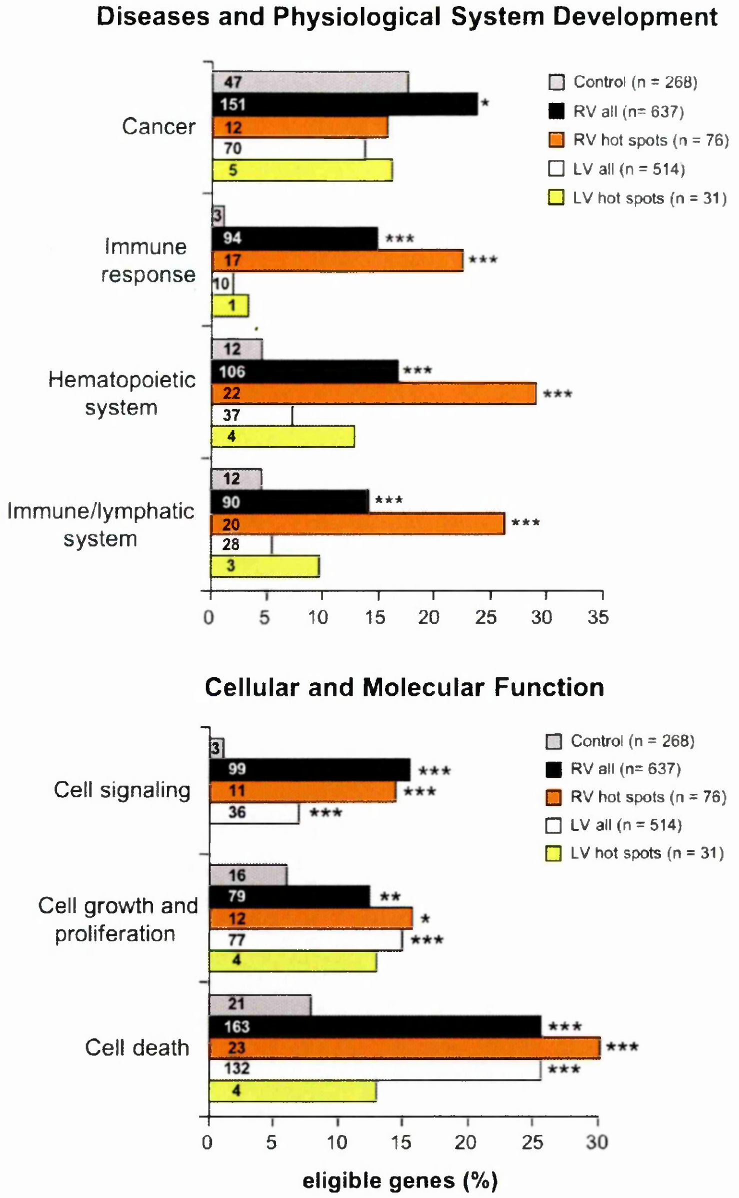


Figure 16. RV vectors preferentially target genes regulating hematopoietic cell growth and differentiation. Functional clustering analysis comparing integration target and control gene lists. Function/disease categories were those significantly over-represented in at least one integration target gene list ($0.005 < p < 0.05$) using the Ingenuity Pathway Knowledge base as background population and the Ingenuity analysis software. Bars represent the percentage of integration target genes belonging to each category among *n* genes eligible for the analysis. Asterisks denote the probability that differences observed between the integration data sets (RV, LV, RV and LV hot spots) and the control data set are due to chance alone (***$p < 0.001$, ** $p < 0.005$, * $p < 0.05$, 2 sample test for equality of proportions with continuity correction). The number of genes annotated within each category is indicated in the bars.

Given the observed preference for RV proviruses to land nearby cancer-associated genes, I performed a further analysis of retroviral-targeted genes using cancer-related databases (see Materials and Methods, section 5.5). RV integrations hit 77 proto-oncogenes and 64 cancer-associated murine common insertion sites (CISs), corresponding to 7.5 and 6.2%, respectively, of the 1,030 integrations (**Figure 17**). Both categories were significantly over-represented ($p < 0.001$) when compared to control sequences (27 proto-oncogenes and 17 CISs out of 798 sequences). On the other hand, LV integrations hit 49 proto-oncogenes and 32 CISs out of 849 integrations, a borderline significant difference in comparison with controls ($p = 0.03$ and 0.07, respectively).

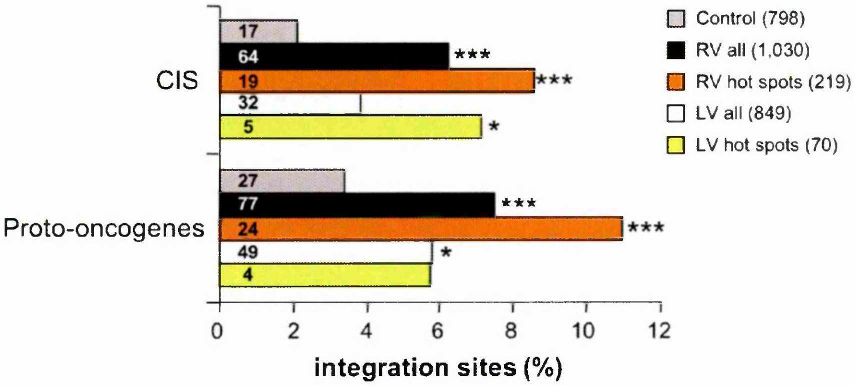


Figure 17. CISs and proto-oncogenes are over-represented in CD34$^+$ RV integrations and integration hot spots. Comparative analysis of the frequency of genes annotated in the CIS and cancer-related gene databases (see Materials and methods, section 7.5, for definitions and data source) between integration target and control gene lists. Bars represent the percentage of RV and LV integrations, RV and LV integration hot spots, and control sequences, targeting at least one proto-oncogene or CIS. The *n* values indicate the number of independent hits in each group. Asterisks denote the level of enrichment with respect to control data set (*** $p < 0.001$, * $p < 0.05$, 2-sample test for equality of proportions with continuity correction).

Overall, these analyses show that both RV and LV vectors have a general tendency to integrate near genes involved in the regulation of cell growth and proliferation, and that RV integration have a specific bias for genes associated with hematopoietic functions and oncogenic transformation. These biases were confirmed when I explored the molecular interactions between integration target genes, using the IPA network generating tool. Ingenuity dynamically computes a large "global" molecular network based on the thousands of direct and indirect physical and functional interactions between orthologous mammalian genes that are annotated in the IPKB. *Ad hoc* algorithms are then applied to select sub-parts of this global network (referred to as "local networks") that are relevant to the gene list of interest. Performing the IPA network analysis on the list of genes targeted by RV and LV integrations (specified in **Appendix 1**), a significant number of those genes resulted functionally linked in molecular networks involved in apoptosis, cell growth/proliferation, signal transduction, transcriptional regulation, and cancer (**Figure 18** and **Appendix 2** for the complete list of networks). Central nodes to both RV and LV networks are genes specifically controlling blood cell proliferation and differentiation, whose deregulation has been related to hematopoietic disorders (among the others EVI1, RUNX1, CBFB, SPP1, ETS1, NOTCH1, CSF1R, FAS[191-198]). This became particularly evident for RV integration target genes when I merged the 5 top RV networks (see **Appendix 2**) into a single, large network, looking for overlapping genes and/or additional relevant functions (**Figure 19**). 59 out of the 155 (38%) genes with an annotated biological function within this network were shared between 2 or more single networks, meaning that local networks are strictly inter-related. The most significant functions associated to the merged network ($10^{-11} < p < 10^{-8}$) were those involved in the hematopoietic system development and function, and in the activation, proliferation and differentiation of blood cells, again pointing out a preferential integration of RV vectors in the vicinity of hematopoietic-specific genes.

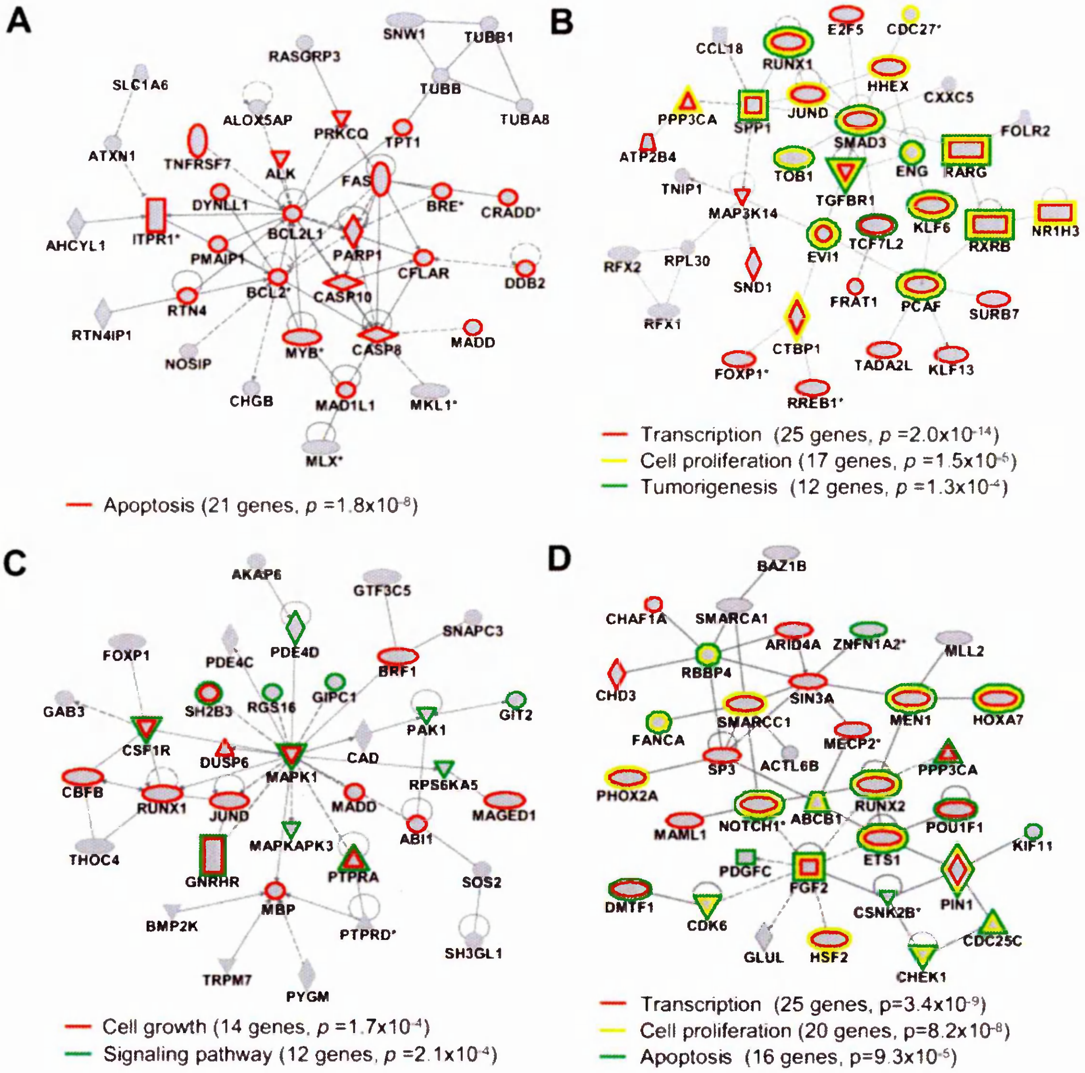
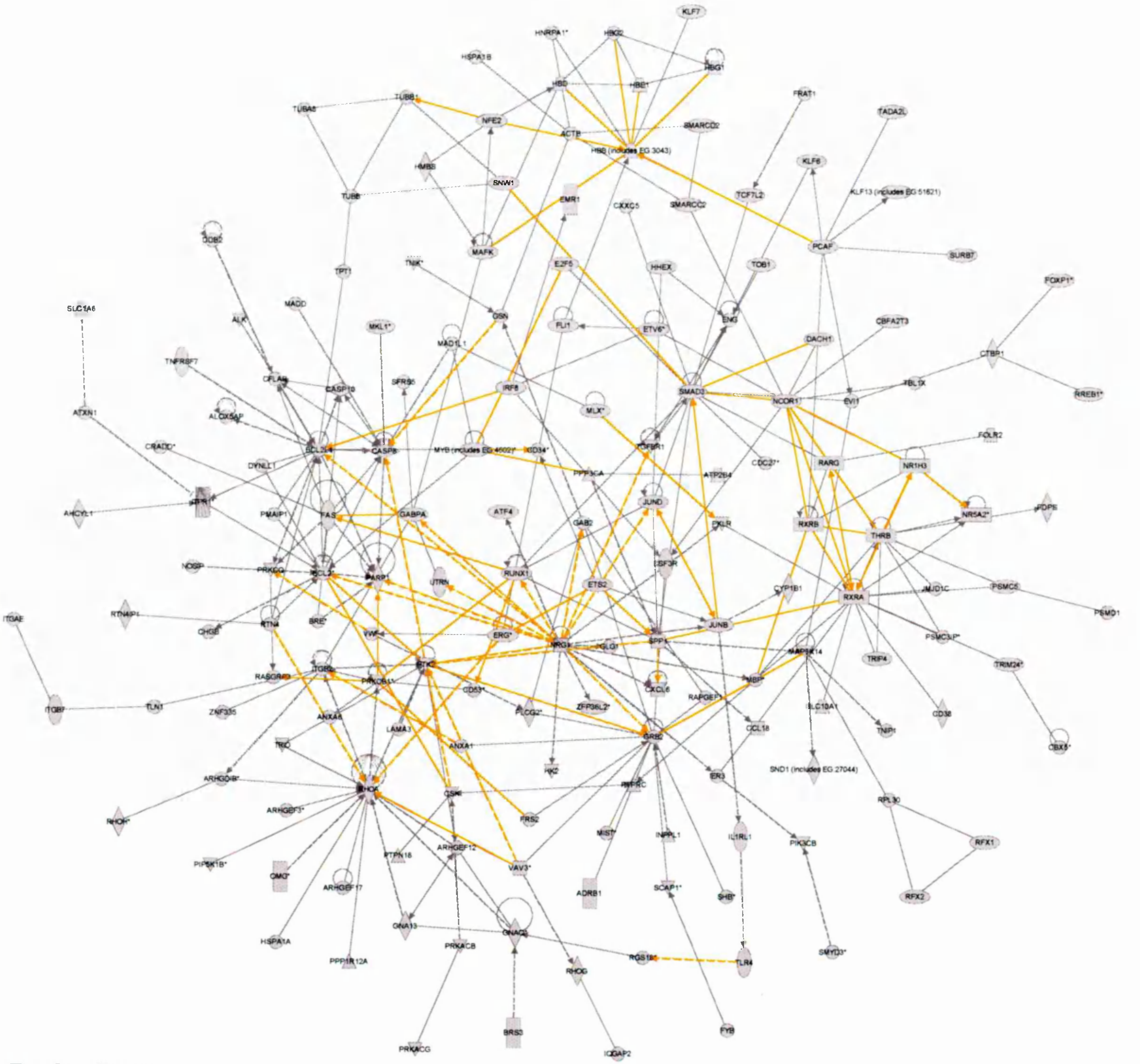


Figure 18. Genes hit by retroviral integration are functionally linked in gene networks. Representative networks originated by Ingenuity analysis of RV (**A** and **B**) and LV (**C** and **D**) target genes (see **Appendix 2** for a complete list). All networks are made of 35 target genes, with an Ingenuity score of 42 or higher. The color code highlights the most significant biological functions associated to each network ($p < 0.001$). Asterisks denote genes hit by at least 2 independent integrations. Shapes and line styles are explained in **Appendix 2**, containing the legend of all symbols used by Ingenuity tool. **(A)** RV network 1; **(B)** RV network 4; **(C)** LV network 1; **(D)** LV network 2 (networks are identified in **Appendix 2**).



Top functions:

- hematological system development and function (66 genes, $p = 6.1 \times 10^{-12}$)
- hematological disease (54 genes, $p = 1.53 \times 10^{-10}$)
- proliferation of blood cells (30 genes, $p = 2.5 \times 10^{-11}$)
- differentiation of blood cells (26 genes, $p = 6.1 \times 10^{-12}$)
- activation of blood cells (22 genes, $p = 2.3 \times 10^{-9}$)

Figure 19. RV networks are functionally inter-related in hematopoietic specific pathways. The network was obtained by combining the 5 RV local networks of **Appendix 2** with the Ingenuity "Merge networks" tool. The orange color highlights direct (continuous lines) and indirect (dotted lines) interactions between genes that are shared among 2 or more local networks. Top functions associated with the merged network are specified, each with the number of genes accounting for that function and the level of over-representation with respect to genes annotated in the IPKB.

Merging of the 4 LV networks showed a good level of overlap between local networks (49 out of 127 genes in common, 38.6%) but no evident biases towards hematopoietic specific functions. The same Ingenuity network analysis performed on the list of genes found nearby control sequences retrieved a single network, therefore impossible to merge (**Appendix 2**).

6.1.3 RV but not LV vectors show a high frequency of integration hot spots

The large number of data I collected, together with the experimental setting I chose, *i.e.*, hematopoietic HSCs analyzed short-term after infection, allowed me to investigate the existence of recurrent sites of RV and LV integration before retrovirally-induced clonal dominance could arise in culture. To visualize how independent integrations clustered together in the genome of CD34$^+$ cells, I started plotting the distribution of the distance between consecutive insertion sites for RV, LV and control sequences (**Figure 20**). Distances between consecutive integrations were plotted individually (upper panels) or grouped into 8 distance intervals (lower panel), for easier comparison between the three data sets (**Appendix 4** for numbers and complete statistics). For up to 16.6% of RV integration sites, the nearest upstream and/or downstream insertions were within 100 kb, while only 4.4% of control sequences and 8.9% of LV insertions were less than 100 kb apart (cumulative frequencies calculated on the first 5 distance intervals, from 1 to 100,000 bp). The same analysis performed with MLV integrations in HeLa cells showed a distribution similar to that of RV vectors in CD34$^+$ cells, even if less accentuated (11.2% of HeLa insertion sites were less than 100 kb apart, compared with 16.6% of RV insertions in CD34$^+$ cells). This was the first, rough indication that RV integration sites were more clustered than LV and control sequences. I then performed a subtler analysis to score for the presence of "true" integration hot spots; I used essentially the same criteria previously applied to the definition of cancer-associated CISs, again based on the distance between two consecutive integrations.

A genomic region was statistically defined a hot spot for integration when containing at least 2 independent insertions in less than 30 kb, 3 in less than 50 kb, and 4 or more in less than 100 kb[14,161]. Overall, 219 (21.3%) of 1,030 RV insertion sites met these criteria, identifying 97 hot spots in the genome of CD34$^+$ cells. A total of 109 (12.5%) of 869 integrations met the same criteria in HeLa cells, defining 52 hot spots. LV vectors showed a significantly lower propensity to integrate at recurrent sites, with only 70 (8.2%) out of 849 integrations meeting the definition criteria, and identifying 33 hot spots (see **Appendix 4** for a complete list of integrations originating hot spots and genes targeted by them). Comparing the 3 collections, a single hot spot region was found in common between RV (4 hits) and LV (3 hits) integrations (chromosome 17 q23.2: 55188652-55285672), while 3 hot spots appeared to be a recurrent insertion site for RV vectors both in CD34$^+$ and in HeLa cells (chromosome 10 q21.2: 63178757-63189469; chromosome 17 q11.2: 22880336-22924624; chromosome X p22.11: 23863173-23925096). Importantly, 22 out of 798 control sequences (2.8%) also met the hot spot definition criteria, defining a background level of false positivity in the LM-PCR analysis. The different subgroups of RV integrations contributed to the hot spot list proportionally to their size, with no apparent bias related to the source of CD34$^+$ cells (CB-, BM- or PB-derived samples), the vectors used for transduction (U3-deleted or wild type LTR vector), or the number of cell doublings undergone in culture before harvesting (**Table 2**). In particular, non-expanded cell populations (those of BM and PB origin), collectively contributing to less than half of the 1,030 total RV integrations, contributed with at least 1 integration to 56 of the 97 (58%) RV hot spots. Such observation confirms that the high percentage of hot spots scored for RV integrations is not due to a clonal selection in culture, but instead it is an intrinsic property of Mo-MLV integration mechanism.

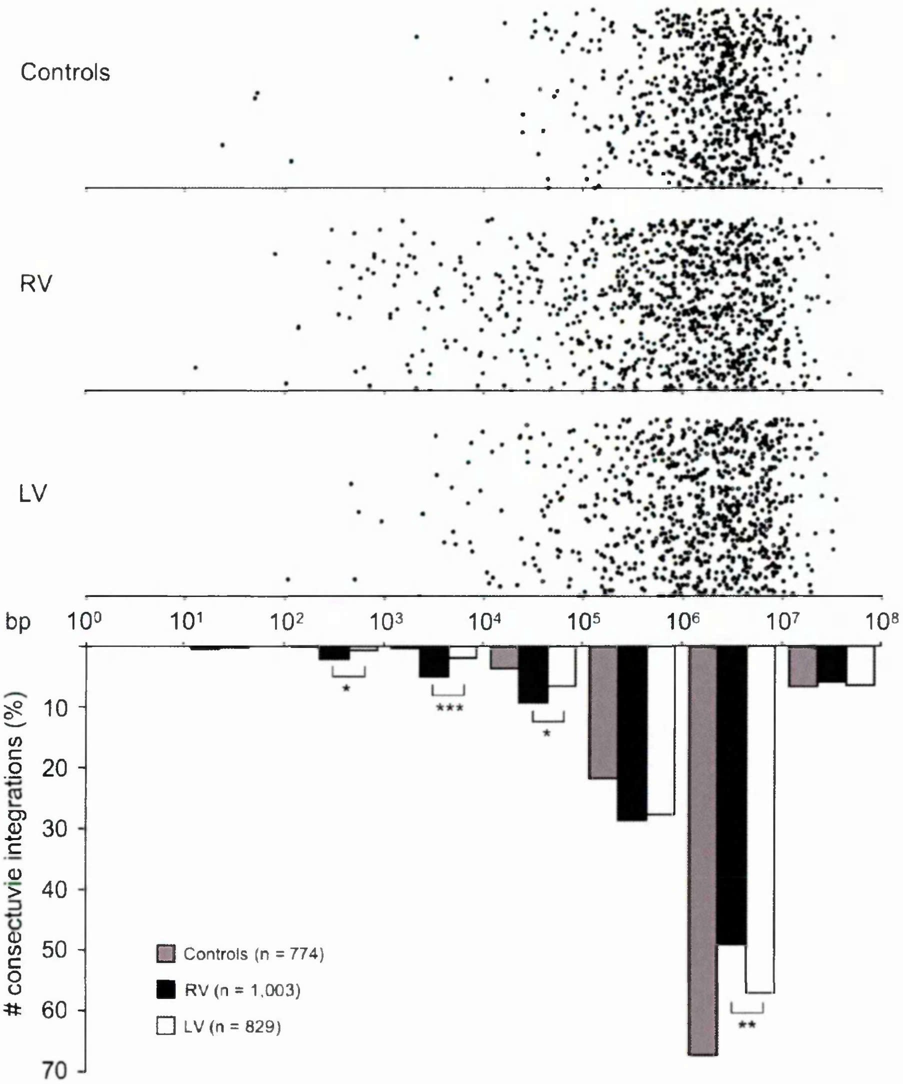


Figure 20. RV integrations are clustered in hot spots. The dot plots on the top represent the distance between pairs of consecutive integrations, plotted along the *x*-axis on a logarithmic scale, computed for RV, LV and control sequence data sets. For a better visualization, dots have been arbitrarily scattered along a virtual *y*-axis, applying a *modulo* function on the distance value (see **Appendix 3** for a detailed description). A quantification of the dot plots is given in the histogram at the bottom, where distances between 2 consecutive integrations are sorted into 8 logarithmic classes (10^0-10^1 bp; 10^1-10^2 bp; 10^2-10^3 bp; 10^3-10^4 bp; 10^4-10^5 bp; 10^6-10^7 bp; 10^7-10^8 bp). The frequency of RV, LV and control sequence consecutive integrations in each distance interval is compared. Asterisks denote statistically significant differences between RV and LV distribution (* $p < 0.05$, ** $p < 0.005$, *** $p < 0.0005$, complete statistics in **Appendix 3**). The *n* values indicate the number of consecutive integration sites for each data set whose distance was plotted.

The position of RV hot spot integrations with respect to Known Genes reflected the RV general integration preferences, with intergenic, perigenic, and gene-dense regions over-represented to the same extent observed in the entire collection of RV integrations, and only a slightly reduced clustering around TSSs ($p = 0.015$; **Table 1**, **Figures 12A** and **13A**). Conversely, LV hot spots showed a higher frequency of integration in intragenic (81.4% *vs.* 71.7%) and gene-dense regions (65.7% *vs.* 35.6% in the "more than 11 genes/Mb" density interval) (**Table 1**, **Figure 13B**). Similarly, RV hot spots occurred in the same proportion of expressed genes as all RV integrations (**Figure 14A**), while LV hot spots contained a significant higher proportion of expressed genes (73.2% *vs.* 55.9%, $p = 0.003$, **Figure 14B**).

Table 2. Contribution of different groups of RV insertions to the integrations generating the RV hot spots.

Data set*	Integrations	% of total (1,030)	Integrations contributing to hot spots	% of total (219)
CB-RV	395	38.3	93	42.5
CB-ΔRV	200	19.4	52	23.7
BM-ADA	190	18.4	33	15.1
BM-X-SCID	120	11.6	18	8.2
PB-ND	125	12.1	23	10.5

* CB-RV: cord blood-derived CD34$^+$ cells transduced with wt-LTR RV

CB-ΔRV: cord blood-derived CD34$^+$ cells transduced with ΔU3-LTR RV

BM-ADA: bone marrow-derived CD34$^+$ cells from ADA-SCID patients transduced with wt-LTR RV

BM-X-SCID: bone marrow-derived CD34$^+$ cells from X-SCID patients transduced with wt-LTR RV

PB-ND: peripheral blood-derived CD34$^+$ cells from normal donor transduced with wt-LTR RV.

Interestingly, the maximum distance between independent integrations defining a hot spot was significantly lower for RV vectors compared with LV vectors and control sequences with hot spot characteristics. Overall, 52% and 67% of the RV hot spots in CD34$^+$ and HeLa cells span less than 10 kb, including those containing 3 or 4 independent hits, compared with 36% and 27% for LV and control sequences, respectively (**Figure 21**). More strikingly, one-fourth (26%) of RV hot spots in CD34$^+$ cells and almost one-half (40%) of those in HeLa cells contained 2 independent integrations in less than 2 kb, compared with only 3% of the LV hot spots. This strengthens what already shown in **Figure 20**, where, comparing the general distribution of all RV and LV integrations, significant clusters of insertion sites were mainly observed for RV but not LV integrations.

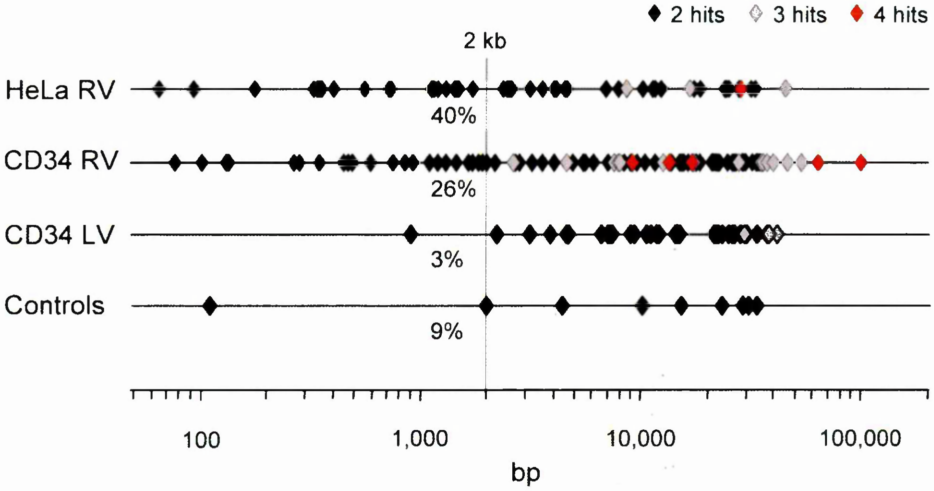


Figure 21. Distribution of the maximum distance between individual hits within RV and LV hot spots. Diamonds represent single hot spots originated from 2 (black), 3 (grey), or 4 (red) hits in the genome of CD34$^+$ HSCs (1,030 RV and 849 LV integrations) and Hela cells (869 RV integrations), plotted according to the maximum distance between individual integrations (in base pairs, on a logarithmic scale). Also shown are "false positive" hot spots generated by applying the definition criteria to a library of LM-amplified random sequences of human CD34$^+$ DNA (798 sequences). A total of 26% of the 97 RV hot spots in CD34$^+$ cells and almost one-half (40.4%) of the 52 RV hot spots in HeLa cells contained 2 independent integrations in less than 2 kb, compared with only 1 of the 33 LV hot spots.

6.1.4 Proto-oncogenes and cancer-associated CISs are hot spots of RV but not LV integration

The list of RV integration hot spots in CD34$^+$ cells included proto-oncogenes, such as LYL1 (lymphoblastic leukemia derived sequence 1) and MYB (v-myb myeloblastosis viral oncogene homolog), cancer-associated CISs, like FLI1 (Friend leukemia virus integration 1), EVI2A (ecotropic viral integration site 2A), EVI2B and NF1 (neurofibromin 1), and genes involved in chromosomal translocations in hematopoietic malignancies, such as the well-known LMO2, MKL1 (megakaryoblastic leukemia translocation 1) and ETV6 (Ets variant gene 6 - TEL oncogene) (**Table 3** for the complete list). All of these genes occured at frequencies significantly higher than expected ($p < 0.001$) and higher than in the overall list of RV integrations (**Figure 17**, red bars). Interestingly, non-expanded cell populations contributed with at least one integration to 9 (53%) of the 17 hot spots targeting a proto-oncogene or a cancer associated CIS, again indicating the absence of any bias related to the number of cell doublings in culture. On the contrary, LV hot spots showed little enrichment for proto-oncogenes or CISs, although in this case low numbers make comparisons poorly significant (**Figure 17**). Moreover, RV but not LV hot spots included a high proportion of genes belonging to the intracellular signaling cascade category (25.3%), which were very significantly over-represented using either the human genome or the total RV integrations as a background population in a GO analysis (EASE score 1.2 e-6 and 2.2 e-4, respectively), despite their relative small number (22). An Ingenuity pathways analysis carried out with the list of genes targeted by RV hot spots showed that genes involved in hematopoietic and immune system development and function and in immune response are further and significantly enriched in RV hot spots with respect to the entire list of RV integrations (**Figure 16**).

Table 3. RV and LV hot spots containing at least one proto-oncogene and/or cancer-related CIS

	Chr	Range (bp)	Hits	Gene symbol	Origin*
RV hot spots	14q24.3	13882	4	C14orf43, **PNMA1**	CB-RV (2) CB-ΔRV (2)
	11p13	48661	3	AF116668, **LMO2**	BM-ADA (1) CB-RV (1) CB-ΔRV (1)
	17q11.2	7827	3	**EVI2A**, **EVI2B**, **NF1**, OMG	BM-X-SCID(1) CB-RV (2)
	10q25.2	1920	2	**ADD3**	BM-ADA (1) CB-ΔRV (1)
	11q23.2	22851	2	**ZBTB16**	BM-ADA (1) CB-RV (1)
	11q24.3	14147	2	**FLI1**	BM-ADA (1) CB-RV (1)
	12p13.2	7360	2	**ETV6**	CB-RV (2)
	16p13.11	18559	2	**ABCC1**	BM-ADA (1)
	19p13.13	137	2	BTBD14B, **LYL1**, **NFIX**, TRMT1	BM-X-SCID(1) CB-RV (1)
	20p12.3	136	2	**PLCB1**	CB-ΔRV (2)
	20q13.12	19100	2	C20orf121, **PKIG**, **SERINC3**	CB-RV (2)
	22q13.1	29588	2	AB051446, **MKL1**, **RUTBC3**	BM-X-SCID (1) CB-ΔRV (1)
	2p11.2	779	2	**CAPG**, LOC284948, RBED1	PB-ND (2)
	2p21	975	2	AK025445, MGC40574, **THADA**, **ZFP36L2**	CB-RV (2)
	4p14	11999	2	N4BP2, **RHOH**	CB-RV (2)
	6q23.3	9422	2	**MYB**	CB-RV (2)
	6p24.3	1991	2	**RREB1**	CB-RV (1) CB-ΔRV (1)
LV hot spots	9q34.3	31043	3	AK130247, C9orf163, INPP5E, **NOTCH1**, PMPCA, DCCAG3	CB-ΔLV (2) CB-LV (1)
	2p21	22106	2	**THADA**	CB-ΔLV (1) CB-LV (1)
	20p12.3	24132	2	**PLCB1**	CB-LV (2)
	17p13.3	25818	2	RUTBC1, **SMG6**, SRR, TSR1	CB-LV (2)
Control hot spots	6q25.1	4561	2	**ESR1**	

Range indicates the maximum distance between hits contained in each hot spot. Proto-oncogenes or CISs are shown in bold. For the complete list of hot spot regions see **Appendix 4**. The number in parentheses indicates the number of hits for each category. * Refer to the legend of **Table 2**.

6.2 Role of LTR and of LTR interactors in the integration site selection of retroviral vectors

The analysis of RV and LV integration sites in human HSCs described herein showed that there is an RV-specific propensity to integrate into hot spots and to target genes involved in the control of growth, differentiation and function of hematopoietic cells. This suggested that the gene expression program of the target cell might be instrumental in directing RV integration, and set the basis for a deeper investigation of the molecular mechanism connecting retroviral integration and transcription. The second part of this thesis was therefore specifically aimed at evaluating the role of transcriptional regulatory networks in directing RV and LV integration. As thoroughly discussed (section 2.1.5), viral LTRs, and in particular the U3 region, contain a large array of *cis*-acting control elements that bind cellular transcription factors (TFs) and regulate transcriptional initiation from eukaryotic promoters. An intriguing hypothesis linking integration and transcription is that cellular TFs sitting on the U3 viral enhancer could cooperate with viral integrase in directing PICs towards regulatory regions actively engaged by the transcriptional machinery. To test such hypothesis, I worked both on the viral and the cellular side, using LTR-modified retroviral vectors and investigating the genomic features surrounding their insertion sites. I designed wild type and LTR-modified (U3-deleted or replaced) RV and LV vectors to infect human HSCs and I collected 200 to 800 integration sites per vector. I then analyzed the effect of LTR modification on RV and LV integration properties by evaluating the arrangement of putative TF binding sites in the genomic regions flanking the retrieved integration sites. Such analysis required specialized skills and deep knowledge of statistics and bioinformatics that I did not possess; therefore I started a close collaboration with the Bioinformatics Core at IFOM-IEO campus (Milan); working together, we defined experimental groups, designed appropriate controls and chose the best approaches to answer our research question.

6.2.1 Collection of integration sites from human hematopoietic cells transduced with LTR-modified retroviral vectors

Human CD34$^+$ HSCs of cord blood, bone marrow or peripheral blood origin were transduced under cytokine stimulation with the Mo-MLV-derived (RV) or HIV-1-derived (LV) vectors schematized in **Figure 11**, carrying wild-type or modified LTRs. RV vectors carried a wild type LTR (MLVa-c), an enhancer-less (ΔU3) LTR, or an LTR from the spleen focus forming RV (SFFV), driving the expression of reporter or therapeutic genes, with or without an internal SV40 promoter-reporter cassette. LV vectors carried a wild type LTR (HIVa-b), a ΔU3 LTR or an LTR containing the Mo-MLV U3 enhancer, and an internal expression cassette driven by different promoters (CMV, SV40 or the entire Mo-MLV LTR). For each vector, 200 to 800 vector-genome junctions were amplified by LM- or LAM-PCR, cloned into bacterial libraries, sequenced and finally mapped onto the human genome. A collection of 795 sequences randomly cloned by LM-PCR was again used as a control group, together with 100,000 computer generated random insertion sites. Integration sites were annotated as TSS-proximal when occurring 5 kb upstream or downstream of the TSS of any Known Gene (UCSC definition), as intragenic when landing into a gene but at a distance > 5 kb from its TSS, and intergenic in all other cases (**Figure 22**). As largely expected, all RV vectors showed a preference for integration around the TSSs, while LV vectors integrated preferentially within genes, as compared to the control sequence set (**Table 4**). Over-representation of TSS-proximal insertions was reduced in the ΔU3-MLV vector data set (12.5% *vs.* 16.6% of MLV), with a concomitant, statistically significant increase in intergenic integrations (47.5% *vs.* 37.0% of MLV, $p < 0.01$, 2-sample test for proportions with continuity correction). LTR modification had no apparent effect on the LV integration preferences in terms of intragenic, intergenic and TSS-proximal distribution.

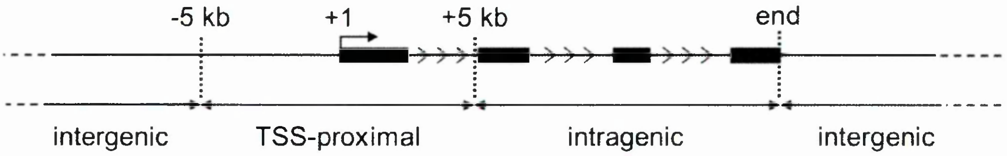


Figure 22. Annotation parameters. Integration sites were annotated as "TSS-proximal" when occurring within a distance of ±5 kb from the TSS of any Known Gene (UCSC definition), as "intragenic" when occurring into a gene at a distance of >5 kb from the TSS, and as "intergenic" in all other cases.

6.2.2 Transcription factor binding sites are over-represented in sequences flanking RV integration sites

To investigate the role of transcription in mediating retroviral target site selection, we evaluated the abundance of transcription factor binding sites (TFBSs) in a ±1,000-bp interval from the integration sites of all RV and LV vectors in human HSCs. To remove from the analysis the possible bias introduced by RV preference for promoter regions, which are enriched in TFBSs by definition, we generated seven weighted control groups of random sequences. These sequences reproduced, in proportion, the integration preferences of each vector set, based on the annotation reported in **Figure 22** (**Table 5**). Such random sequences were then used as pair-weighted background for a TFBS analysis by the Clover program[184], using Jaspar

Core 2005[185] as a database of experimentally validated TFBS motifs. Clover program screens a set of DNA sequences against a precompiled library of motifs and assesses which, if any, of the motifs are statistically over- or under-represented in the sequences when compared to a background set of sequences. Jaspar is an open-access database of annotated, high-quality, matrix-based TFBS profiles for multicellular eukaryotes. The profiles are non-redundant and were derived exclusively from sets of nucleotide sequences experimentally demonstrated to bind TFs, two characteristics that render Jaspar preferable to other more extensive libraries, such as TRANSFAC.

Table 4. Integration distribution of wild type and LTR-modified retroviral vectors in human $CD34^+$ HSCs.

	Intergenic (%)	TSS proximal (%)	Intragenic (%)	Total hits	IN
$CD34^+$ cells					
MLV	37.0	16.6	46.4	829	MLV
ΔU3-MLV	47.5	12.5	40.0	200	MLV
SFFV-MLV	42.0	19.0	39.0	195	MLV
HIV	28.1	8.4	63.5	403	HIV
ΔU3-HIV[CMV]	26.5	7.4	66.1	445	HIV
ΔU3-HIV[MLV]	24.5	9.5	66.0	200	HIV
MLV-HIV	26.0	10.0	64.0	400	HIV
Controls	59.8	4.5	35.7	795	
Hela cells					
MLV[12]	45.0	14.4	40.6	864	MLV
HIV[121]	17.3	5.6	77.1	532	HIV
HIVmIN[121]	50.8	15.7	33.5	325	MLV

Distribution of integration sites of different RV and LV vectors identified by LM- and LAM-PCR in the genome of human $CD34^+$ HSCs and HeLa cells. Control sequences were randomly cloned by LM-PCR from $CD34^+$ DNA samples. See **Figure 11** for the structure of each vector and for the definitions of the annotation parameters. The origin of the integrase (IN) packaged with each vector is indicated in the rightmost column. Insertion sites from HeLa cells were re-analyzed from previously published collections.

Table 5. Definition of weighted backgrounds.

Background group	Intergenic (%)	TSS proximal (%)	Intragenic (%)	Corresponding experimental group
BG1	59.8	4.5	35.7	Controls
BG2	37.0	16.5	46.4	MLV (CD34$^+$)
BG3	28.0	8.4	63.5	HIV (CD34$^+$) ΔU3-HIV[CMV] ΔU3-HIV[MLV] MLV-HIV
BG4	41.6	19.9	38.5	SFFV-MLV
BG5	46.3	13.4	40.3	MLV (Hela) ΔU3-MLV
BG6	17.3	5.6	77.1	HIV (Hela)
BG7	50.8	15.7	33.5	HIVmIN

We randomly generated seven groups of sequences (BG1-7) reproducing, in proportion, the integration preferences of each vector set and we used them as pair-weighted backgrounds for transcription factor binding site analysis by the Clover program. For each background group, the corresponding experimental group/s is/are specified.

Figure 23 shows the number of TFBS motifs that were found enriched by Clover analysis in each group of vectors with respect to its fitted background. The box plots indicate that RV but not LV vectors integrate in genomic regions highly enriched in TFBSs (86.8 and 90.3 average TFBS counts per sequence for MLV and SFFV-MLV respectively *vs.* 27.2 for control sequences, $p < 2.2e\text{-}16$, Wilcoxon rank sum test; for complete statistics refer to **Appendix 5**). The observed enrichment is independent of the position of integration sites with respect to genes and TSSs, since it is present in intergenic as well as in intragenic integrations, with only a slight increase around TSS-proximal insertion sites noticeable in MLV and SFFV-MLV data sets. The RV LTR enhancer appears to play an essential role in this selection, since deletion of the U3 region, but not its replacement with the SFFV enhancer, causes a significant drop in the frequency of TFBSs around the insertion sites (35.4

for ΔU3-MLV *vs.* 86.8 for MLV, $p < 2.2e\text{-}16$). Conversely, sequences around LV vector integration sites show a significantly lower TFBS content compared to control sequences. Interestingly, replacement of the HIV U3 by the MLV U3 enhancer in the HIV LTR (MLV-HIV vector) appears to bias LV integration towards regions with an increased content of TFBSs (from 12.6 TFBSs/sequence of HIV to 29.1 of MLV-HIV). The MLV U3 enhancer plays this role only when placed inside an LTR, since it had no apparent effect in an internal position within the LV vector (compare ΔU3-HIV[MLV] distribution with that of HIV in **Figure 23**).

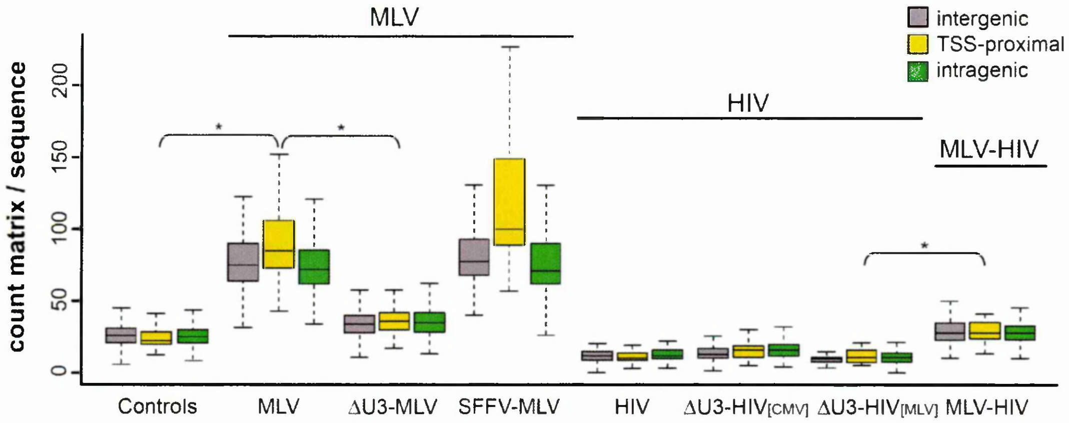


Figure 23. Abundance of TFBSs in genomic sequences flanking retroviral integration sites in human HSCs. Box plot of the frequency of TFBSs (motif count per sequence) in genomic sequences flanking integration sites (±1,000 bp) of different RV and LV vectors, in human HSCs. The plot is broken down into the three annotation categories of intergenic (grey), TSS-proximal (yellow), and intragenic (green) integrations. Statistical significance of differences in TFBS counts among and within groups is reported in **Appendix 5**.

6.2.3 Retroviral integration sites are flanked by unique TFBS motifs.

Given the remarkably different abundance of TFBSs around RV and LV vector integrations, we then moved to the question of which TF motifs were specifically over- or under-represented in each vector when compared to its pair-weighted background. This was visualized by a two-way hierarchical clustering of the likelihood ratio values coming out from the Clover analysis (**Figure 24**). The heatmap shows that each experimental group of sequences is uniquely defined by specific subsets of TFBS motifs, the color code being suggestive of the significance level reached for each motif (blue to red for increasing likelihood values). The row dendrogram on the right of the heatmap shows that RV, control and LV sequences identify three main nodes, from which other branches originate, dictated by the vector LTR configuration. The bootstrapped column dendrogram on the top, instead, splits the data set into two major branches, defining LV and RV vector profiles. The bootstrapping procedure, a resampling technique used to obtain estimates of summary statistics, was here applied to add robustness to the analysis. Only nodes having an Approximately Unbiased (AU) probability value > 0.95 were scored as significant and stable nodes (represented as red branches on the tree of **Figure 24**; the complete analysis is reported in **Appendix 6.1**). A core of four motifs (MA0056, MA0081, MA0026, MA0098, all motifs are listed in **Appendix** 7) is strongly associated (AU = 100) to all RV vectors, independently of their LTR structure. Three of these motifs (MA0081, MA0026, MA0098) are bound by TFs belonging to the ETS family, and one (MA0056) by TFs of the Zn-finger C_2H_2 family. Interestingly, sequences flanking the integration sites of the enhancer-less LTR vector (ΔU3-MLV) lack a set of 12 motifs common to MLV and SFFV sequences, and 5 motifs shared among MLV sequences only. These motifs are therefore associated to an RV or specifically to the MLV U3 enhancer.

The hierarchical cluster analysis shows a strong under-representation of TFBSs in all LV sequences, which shared only one characterizing forkhead motif (MA0032). Although the insertion of the MLV U3 enhancer in the HIV LTR increased the absolute TFBS motif count around integration sites (**Figure 23**), it was not sufficient to change the segregation of the MLV-HIV vector sequences in the cluster analysis. **Figure 24** shows that the MLV-HIV sequences share most of their motif profile with LV sequences, with the notable exception of one Zn-finger motif (MA0021) that is instead in common with the MLV and SFFV-MLV vectors.

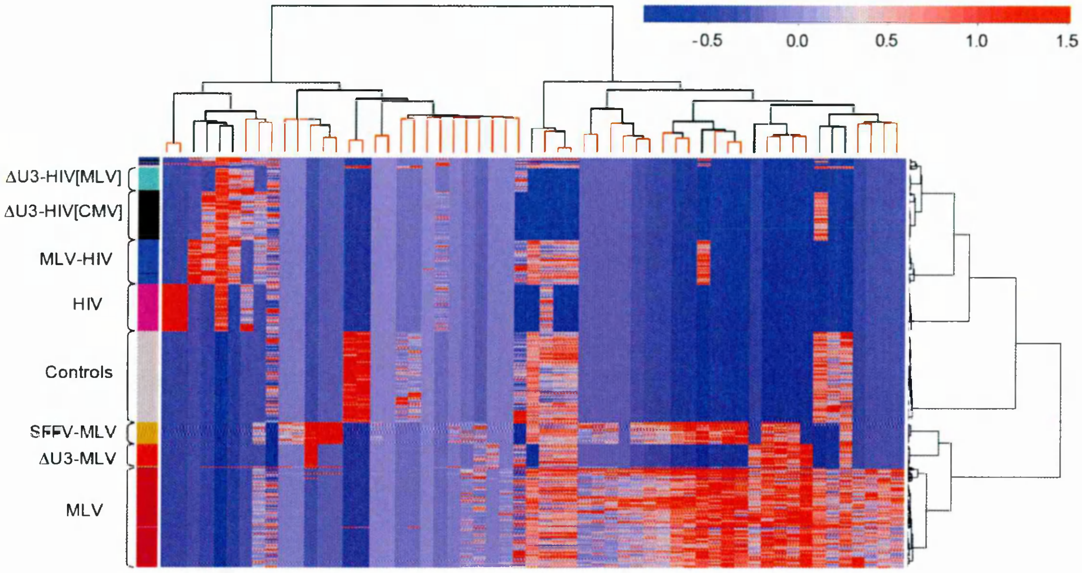


Figure 24. Hierarchical cluster analysis of TFBS motifs around retroviral integration sites in human HSCs. The heatmap defines a specific TFBS motif pattern for each group of sequences (specified on the left). The color code (from blue to red) indicates increasing levels of likelihood values (from under- to over-representation). The row dendrogram on the right shows that RV, control and LV sequences identify three main nodes, from which other branches originate, dictated by the vector LTR configuration. The bootstrapped column dendrogram (top) splits the data set into two main branches, defining LV and RV vector profiles. Red branches on the tree identify stable nodes with an AU *p*-value > 0.95 (detailed dendrogram is in **Appendix 6.1**; Clover analysis results with a complete list of motifs in **Appendix 7**).

To reduce our multivariate data sets to a lower dimension for analysis, while minimizing the loss of information, we chose a Principal Components Analysis (PCA) approach[187]. PCA transforms a number of possibly correlated variables (TFBS motifs, in this case) into a smaller number of uncorrelated variables called pricipal components (PCs). The first PC accounts for as much of the variability in the system as possible, and each succeeding component accounts for as much of the remaining variability as possible. In fact, PCA technique identifies simultaneously all the existing correlations between samples and variables in huge multivariate data, and orders them according to their contribution to the total variance of the system. The most significant relationships between the data dimensions identify major patterns in the data, highlighting the principal similarities and differences among them. Indeed, PCA operations can be thought of as revealing the internal structure of the data in a way which best explains the variance in those data.

When applied to our Jaspar motifs, the PCA confirmed the results of the cluster analysis. A scatter plot of the first two components, accounting together for 31.6% of the total variability, clearly discriminates three main groups: RV sequences (MLV, SFFV-MLV and ΔU3-MLV), LV sequences (HIV, ΔU3-HIV[CMV], ΔU3-HIV[MLV], and the hybrid MLV-HIV), and control sequences (**Figure 25**). The first component differentiates RV from all other sequences, the second one discriminates between LV and control sequences. MLV and HIV groups are oriented along the first component axis but in opposite directions (left panel); the angle between the two is nearly orthogonal, implying an independent behavior. The control group is also independent of RV sequences, and oriented in opposite direction with respect to LV sequences along the second component axis.

The variability within MLV and SFFV-MLV data is higher than in any other group, possibly because of the high number of TFBSs contained in these sequences. Indeed, ΔU3-MLV sequences, which contain a lower number of TFBSs, show a lower variability, although they result still oriented towards the RV group along the axis of the first component. The loadings plot on the right panel shows a high number of TFBSs contributing to the RV group loadings. Among the 19 motif vectors having a length higher than the chosen cutoff (see Materials and methods, section 5.6), one (MA0032) is oriented with the LV group, two (MA0117, MA0089) with the control group, and the remaining ones with the first principal component. Twelve of these vectors are exclusively oriented with the RV group, and belong to different TFBS families; four motifs are recognized by Zn-finger C_2H_2, three by ETS, two by homeodomain-containing, and one by Zn-finger-DOF, HMG, and AP2 transcription factors. The four motifs strongly associated with RV sequences in the cluster analysis (MA0056, MA0081, MA0026, MA0098 of **Figure 24**) are contained in this group.

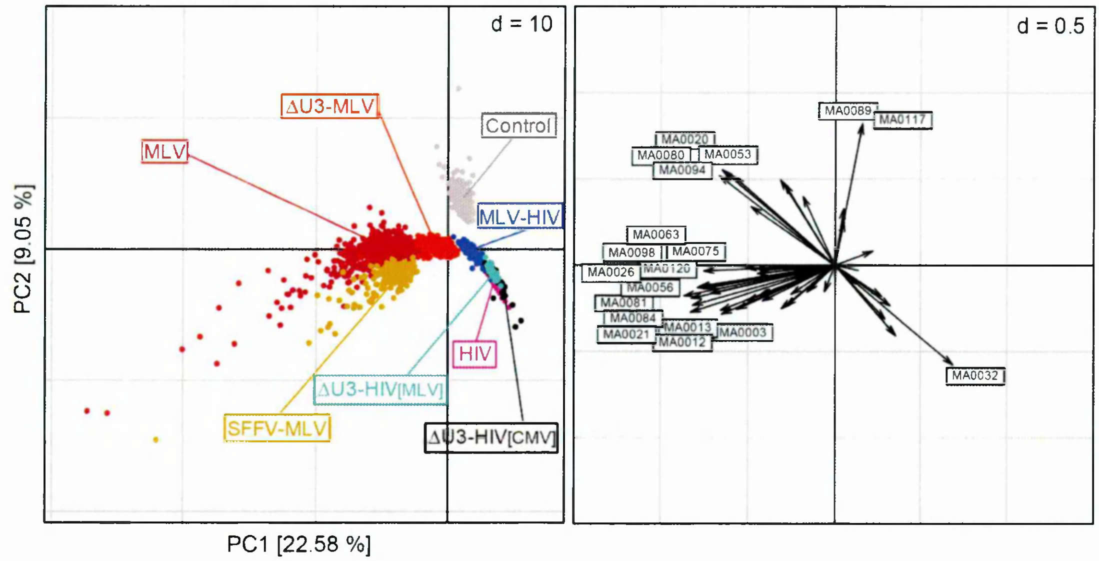


Figure 25. Principal Components Analysis of TFBS motifs enriched around retroviral insertions in human HSCs. The PCA was performed with the likelihood ratio values of the 57 Jaspar matrices that resulted enriched by the Clover program. A scatter plot of the two principal components (PC), accounting together for 31.6% of the total variability (left panel), identifies three main groups: RV sequences (MLV, SFFV-MLV and ΔU3-MLV), LV sequences (HIV, ΔU3-HIV[CMV], ΔU3-HIV[MLV], and the hybrid MLV-HIV), and control sequences. The first component (*x*-axis) discriminates RV from all other sequences, while the second component differentiates LV from control sequences. ΔU3-MLV sequences, containing a lower number of TFBSs, show less variability than the MLV and SFFV-MLV sequences, but are still oriented towards the RV group, along the first component axis. A plot of 19 motif vectors having a length higher than the chosen cutoff (right panel) shows one vector (corresponding to the Jaspar motif MA0032) oriented with the LV group, two (MA0117, MA0089) with the control group, and all the remaining ones with the RV group. The four motifs MA0056, MA0081, MA0026, and MA0098, which were strongly associated with RV sequences (AU *p*-value = 100) in the cluster analysis, are contained within this group.

6.2.4 Evolutionarily conserved TFBSs are enriched in sequences flanking RV integrations.

We next investigated whether an over-representation of TFBSs was still observed around RV integrations when applying more stringent parameters, *i.e.*, considering only evolutionary conserved binding sites. For this analysis, we extracted from the HMR Conserved TFBS table at UCSC 188 motifs belonging to the TRANSFAC Matrix Database (version 7.0) conserved in a human-mouse and/or human-rat alignment. 35.7% and 26.7% of the sequences flanking MLV and SFFV-MLV insertion sites, respectively, contained at least one conserved TFBS (range: 2-30 sites/sequence), a significant difference with respect to their weighted backgrounds and to a random computational control set of 100,000 sequences (17.9%, 18.5% and 14.7%, respectively). Sequences flanking the ΔU3-MLV and all LV integration sites showed no significant enrichment, again with the exception of the MLV-HIV hybrid vector (**Figure 26** upper panel, complete statistics in **Appendix 8.1**).

The same analysis performed on integrations broken down into the three annotation categories of **Figure 22** showed no significant bias towards any of them (**Figure 26** lower panel), meaning that intragenic intergenic and TSS-proximal sequences contributed proportionally to the conserved TFBS over-representation in all samples. A complete list of conserved motifs and their distribution over the different data sets is reported in **Appendix 8.2**. Given the tight constrains in the definition, conserved TFBSs were scored in much smaller numbers than in the Clover analysis.

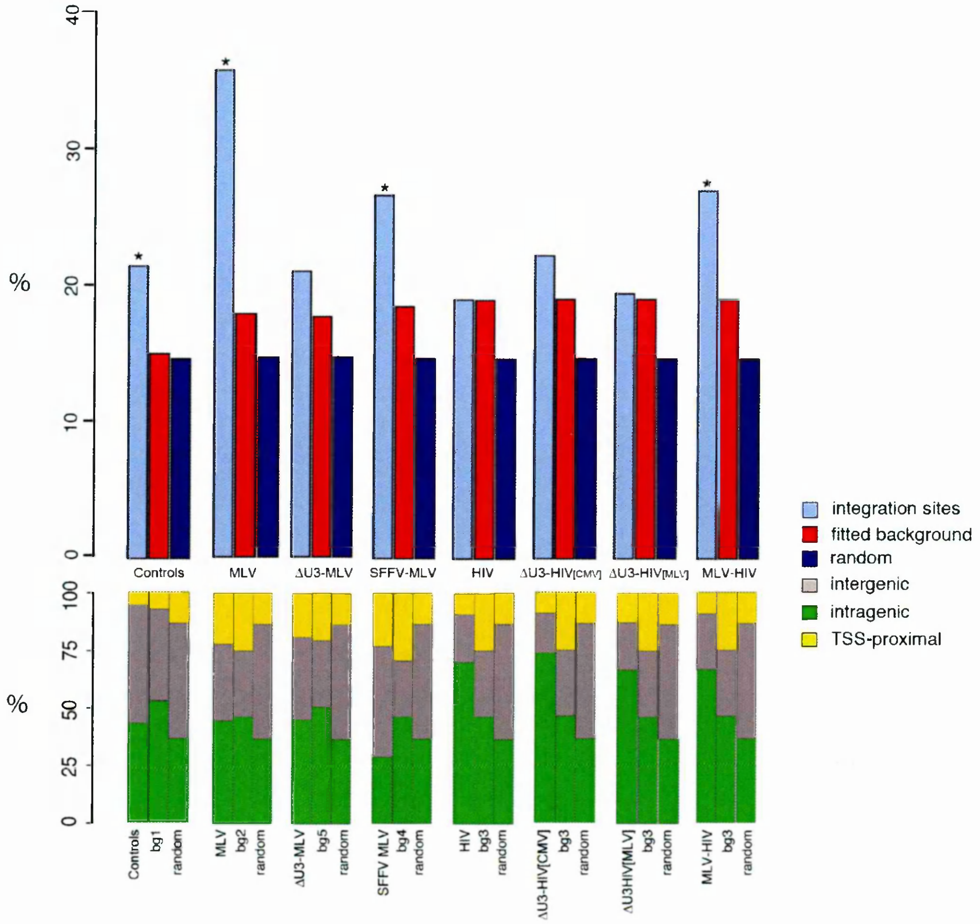


Figure 26. Evolutionarily conserved TFBSs around retroviral integration sites in human HSCs. Analysis of the frequency of evolutionary conserved TFBSs in genomic sequences flanking RV and LV insertion sites in human CD34$^+$ cells, performed on 188 TRANSFAC matrices conserved in a human-mouse and/or -rat alignment (HMR Conserved TFBSs table at UCSC) as a motif database. In the upper panel, data are plotted as percentage of sequences containing at least one conserved TFBS. Each experimental group (light blue bars) is compared to its paired-weighted background ('BG', red bars) and to a random computational control sequence set (blue bars). Asterisks highlight experimental groups that showed a statistical significant enrichment of conserved TFBSs with respect to BG and random sets (one-sided Fisher's exact test, complete statistics in **Appendix 8.1**). In the lower panel, the same frequency data are broken down into three subgroups, according to the integration site annotation (intergenic, TSS-proximal and intragenic). A complete list of conserved motifs and their distribution in each data set are reported in **Appendix 8.2**.

To identify motifs associated with MLV integration by both analyses, we used the STAMP alignment platform and we identified the matrices listed in **Table 6**. Jaspar and TRANSFAC shared motifs are predicted to bind homeodomain, ETS, bZIP, forkhead and Zn-finger proteins, including the cell-type specific growth regulators AML1/RUNX1, FOXO3 and LMO2.

Table 6. Jaspar and TRANSFAC motifs found over-represented around MLV insertion sites.

JASPAR					TRANSFAC (conserved)		
Matrix ID	**Factor**	**Total counts**	**Counts/seq (average)**	**Counts/seq (range)**	**Matrix accession #**	**Factor**	**Total counts**
MA0109	Rush 1α	530	0.63	0-3	M00278	LMO2	18
MA0046	TCF1	871	1.05	0-5	M00132	HNF1	12
MA0002	RUNX1	1,146	1.38	0-4	M00454	MRF2	16
MA0050	IRF-1	1,463	1.76	0-6	M00062	IRF-1	20
MA0012	broad complex_3	1,531	1.84	0-12	M00474	FOXO1	30
MA0123	ABI4	1,726	2.08	0-10	M00515	PPARG	6
MA0026	E74A	1,940	2.34	0-7	M00025	ELK1	4
MA0064	PBF	2,028	2.44	0-9	M00062	IRF-1	20
MA0042	FOXI1	2,217	2.67	0-11	M00289	FOXI1	8
MA0053	MNB1-A	2,246	2.70	0-9	M00062	IRF-1	20
MA0013	broad complex_4	2,297	2.77	0-20	M00477	FOXO3	30
MA0120	Id1	2,553	3.07	0-21	M00258	ISGF3	20
MA0079	Sp1	2,648	3.19	0-10	M00257	RREB1	6
MA0021	dof3	2,902	3.50	0-10	M00062	IRF-1	20
MA0020	dof2	3,201	3.86	0-10	M00062	IRF-1	20

TFBS motifs found significantly enriched in sequences flanking (±1,000 bp) the integration sites of the MLV vector in human HSCs using both the Jaspar and the TRANSFAC conserved motif databases. Frequencies are listed as total counts in the 829 MLV sequences and/or average counts per sequence and range of counts/sequence. Jaspar and TRANSFAC motifs were matched by STAMP[188].

6.2.5 Transcription factors bind retroviral PICs in the cell nucleus.

The association between the MLV U3 enhancer and the over-representation of TFBSs suggested a role for U3-binding proteins in RV target site selection. An intriguing hypothesis is that specific TFs bind the MLV U3 enhancer in the context of nuclear PICs and tether them to genomic regions engaged by the transcriptional machinery. A suggestive observation in this direction is the fact that retroviral LTRs are transcriptionally active prior to integration in acutely infected cells, implying a direct interaction of cellular TFs with viral enhancers and promoter. Transcription from unintegrated MLV LTRs was investigated in the SupT1 human T cell line, after short-term infection with the hybrid MLV-HIV lentiviral vector. In this vector, GFP reporter gene is under the transcriptional control of a hybrid MLV-HIV LTR, containing MLV U3 enhancer elements (**Figure 11** and **26**). Integration kinetics was roughly established by Southern Blot analysis of cytoplasmic and nuclear DNA extracts at different time points after infection (4 hours to 14 days, **Figure 27A**). Nuclear PICs, visible as a ~5,000 bp-band of linear DNA, were barely detectable 4 hours after infection, both in the cytoplasmic and in the nuclear fractions, and peaked at 10 hours, when integrated proviruses were still virtually undetectable. Circular, unintegrated forms were instead already visible 4 hours after infection in the nuclear fraction, but remained at stable levels over time. 14 days after infection integration was complete, with no trace of viral linear or circular cDNA. 10 hours after infection, when linear DNA was the prevalent viral DNA species in the nucleus with no sign of integrated proviruses, was therefore chosen as a reasonable time point to study PIC properties. Transcriptional activity of viral LTRs was measured by GFP expression, evaluating both protein and RNA levels 10 hrs after viral infection (cytofluorimetric and reverse-transcriptase PCR analyses in **Figure 27B** and **C**, respectively). GFP mRNA and protein were readily detectable, demonstrating full LTR activity before proviral integration. Cellular transcription factors responsible for GFP expression from unintegrated LTRs were then investigated by chromatin immunoprecipitation. Several TFs known to interact with the MLV U3 enhancer (schematically represented in **Figure 28A**) were first tested for their expression in SupT1 nuclear and cytoplasmic extracts by Western Blot analysis (**Figure 28B**). YY1 protein, CBF heterodimer (AML1/RUNX1 and CBFB), NF-1 factor, several members of the C/EBP family (α, β, δ), and at least two members of the Ets family

(Ets1 and Ets2), resulted all expressed in SupT1 cells, mainly confined to the nuclear compartment.

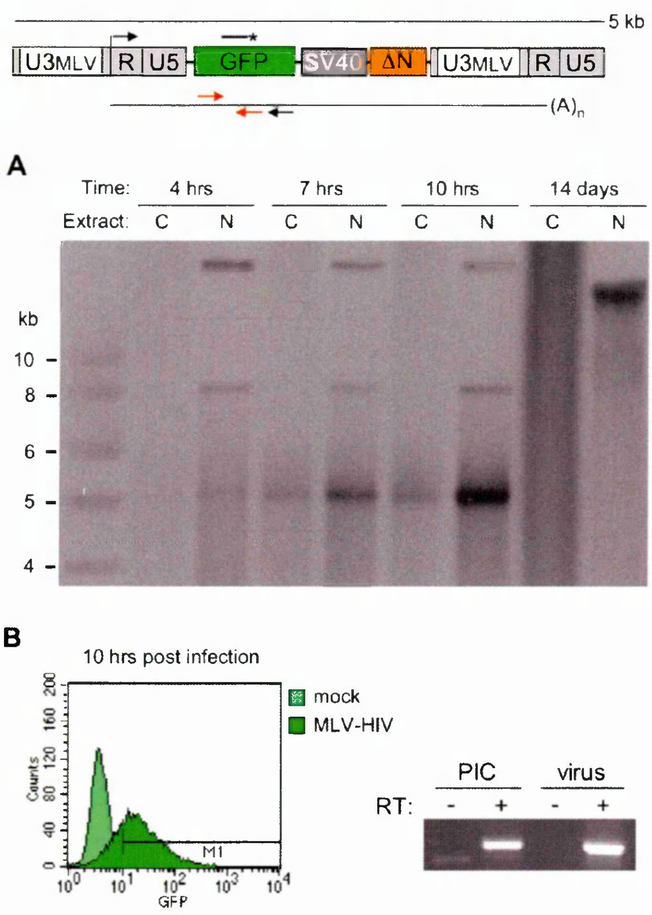


Figure 27. Viral LTRs are transcriptionally competent prior to integration. SupT1 human hematopoietic cells were transduced with MLH-HIV lentiviral vector at an MOI of 25, and samples for DNA, RNA and cytofluorimetric analysis were collected 4 hours to 14 days after infection. Vector structure is schematized on the top, with MLV-HIV hybrid LTR driving the expression of the GFP reporter gene. **(A)** Southern Blot analysis of nuclear and cytoplasmic DNA extracts from MLV-HIV-transduced SupT1 cells 4, 7, 10 hours, and 14 days after infection. For each time-point, DNA was extracted from 1x10^6 cells and run, undigested, on an agarose gel, blotted to a nylon membrane and hybridized to a GFP radiolabeled probe (asterisked line on the vector scheme). Molecular marker sizes are specified on the left (in kb). PIC DNA is the linear molecule of ~5 kb whose levels increase over time, peaking at 10 hrs after infection (proviral DNA, see vector scheme at the top); upper bands are circular, unintegrated viral DNA forms. Signal from integrated proviruses is visible in the last lane as high-molecular weight, undigested genomic DNA. **(B)** GFP expression from MLV-U3 enhancer of unintegrated proviruses. Protein levels were analyzed by cytofluorimetric analysis of MLV-HIV-transduced SupT1 cells (light green) against mock-transduced cells (dark green) 10 hrs post-infection. RNA levels were analyzed by reverse transcriptase-PCR analysis. Total RNA was extracted from MLV-HIV infected SupT1 cells, 10 hours after transduction, and treated with DNaseI. Full-length viral RNAs were specifically retrotranscribed with a GFP-reverse oligo (black arrow on the vector scheme) and cDNAs were amplified by internal GFP forward and reverse primers (red arrows on the vector scheme). Viral RNA genome extracted from pelletted virions ('virus') was used as a positive control for the RT reaction. Negative control reactions containing no reverse transcriptase (RT$^-$) were set up to check for DNA contaminants of RNA samples. RT-PCR products were run on an agarose gel and stained with ethidium bromide for visualization. A GFP transcript was only recovered from RT$^+$ samples.

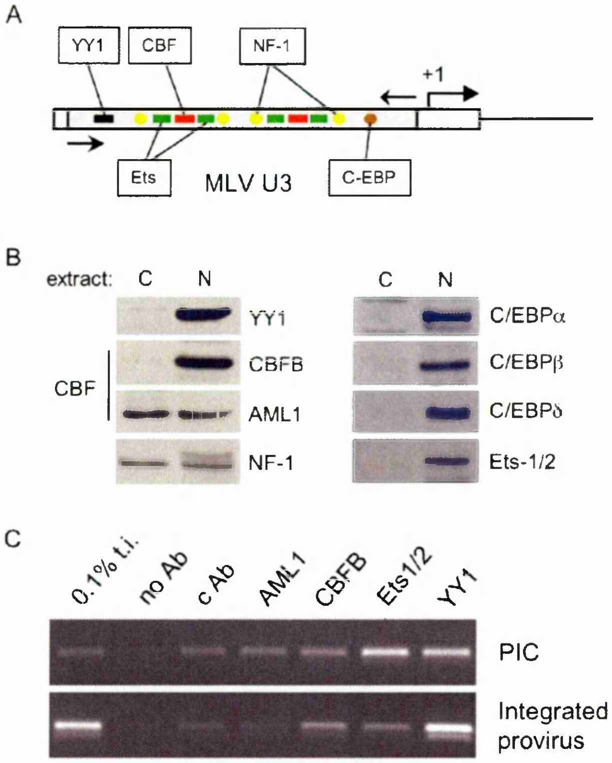


Figure 28. Immunoprecipitation analysis of TFs binding to PICs and integrated proviruses in human hematopoietic cells. (A) Schematic representation of the MLV LTR. Colored bars indicate binding sites for YY1 (black), ETS family members (green), the CBF complex (heterodimer of AML1/RUNX1 and CBFB proteins, red), NF-1 (yellow), and C-EBP proteins (brown) in the U3 enhancer (grey box). +1 indicates the TSS. **(B)** Western Blot analysis of the expression of transcription factors potentially binding MLV U3 enhancer in SupT1 hematopoietic cells. Cytoplasmic and nuclear extracts (50 μg/lane) were run on SDS-polyacrilamide gels and immunostained with anti-YY1, anti-CBFB, anti-AML1, anti-C/EBPα, anti-C/EBPβ, antic/EBPδ, and anti-Ets1/2 polyclonal antibodies. All tested TFs were expressed in SupT1 cells, and, with the exception of NF-1 and AML-1, they were mainly detectable in the nuclear fraction. **(C)** Recruitment of AML1, CBFB, Ets1/2 and YY1 transcription factors on the PICs or the integrated proviruses of the MLV-HIV vector in human SupT1 T-cell line *in vivo*. Cells were cross-linked 10 hours (PIC) or 14 days (integrated provirus) after infection, immunoprecipitated without antibody (no Ab), with a control anti-HA antibody (cAb) or with anti-AML1/RUNX1, anti-CBFB, anti-Ets1/2, and anti-YY1 antibodies, and analyzed by PCR with primers specific for the U3 enhancer (arrows in panel **A**). Amplified fragments were run on agarose gel and stained with ethidium bromide. The first lane corresponds to 0.1% of the total input (t.i.) DNA.

Specific antibodies against these TFs were then used to immunoprecipitate cross-linked DNA isolated from MLV-HIV-infected SupT1 cells 10 hours after transduction. Among all TFs tested, only Ets-1/2 and YY1 showed significant binding within the MLV U3 enhancer in PICs (**Figure 28C**). Interestingly, immunoprecipitation of chromatin from stably transduced cells 14 days after infection showed that the integrated, transcriptionally active U3 enhancer binds Ets-1/2 and YY1, although with different relative intensity, as well as the CBFB component of the CBF heterodimer; binding of the AML1/RUNX1 component was barely detectable (a poor performance of the antibody used for the immunoprecipitation cannot be excluded). These data indicate that specific TFs bind retroviral PICs into the nucleus before integration, although not necessarily in the same configuration required to transcribe the integrated provirus.

6.2.6 Patterns of TFBS motifs flanking retroviral integrations are cell-type specific.

To understand whether the cell context has a role in retroviral integration targeting, we performed a comparative TFBS analysis between sequences flanking MLV and HIV insertion sites in CD34$^+$ cells and sequences retrieved from published collections of retroviral integration sites in the human epithelial cell line HeLa[12,121] (**Table 4**). Also in this cell line, MLV vector integrates in TFBS-rich regions, differently from HIV vector (83.9 *vs.* 29.1 average Jaspar matrices/sequence, **Figure 29,** MLV and HIV box plots).

A two-way hierarchical cluster analysis with both CD34$^+$- and HeLa-derived sequences showed cell-type specific as well as common sets of over-represented motifs (**Figure 30**). The row dendrogram on the right of the heatmap splits the data sets in two branches (MLV and HIV), within which CD34$^+$ and HeLa sequences are clearly separated. The bootstrapped column dendrogram on the top again identifies two main nodes, defining RV and LV distinct patterns (the detailed dendrogram with AU values for each node is reported in **Appendix 6.2**).

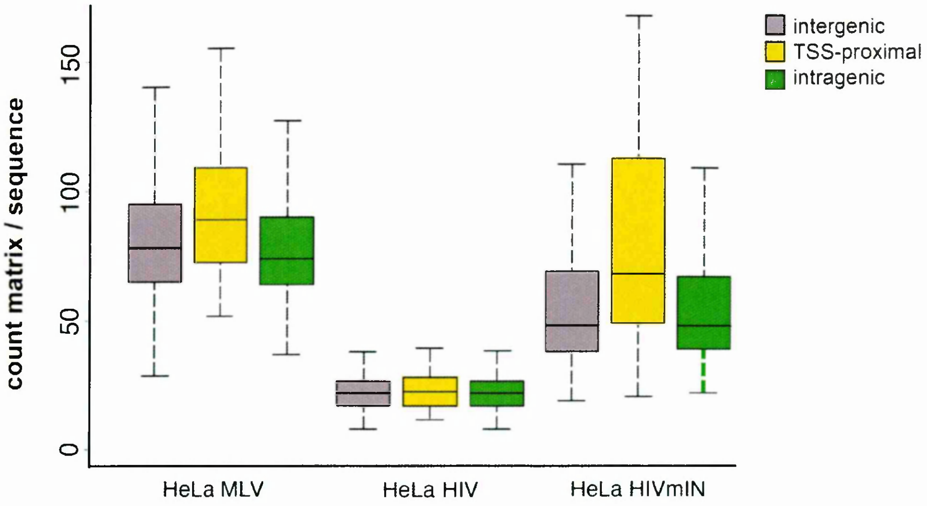


Figure 29. Abundance of TFBSs in genomic sequences flanking retroviral integrations in human HeLa cells. Box plot of the frequency of TFBSs (motif counts/sequence) found ±1,000 bp around intergenic, TSS-proximal, and intragenic insertion sites of an MLV vector, an HIV vector, and an HIV vector with an MLV integrase (HIVmIN) in HeLa cells. Statistical significance of differences in TFBS counts among and within groups is reported in **Appendix 5**.

The cluster analysis shows that three Zn-finger (MA0021, MA0020, MA0053), four ETS (MA0081, M0026, MA0080, MA0098) and two forkhead (MA0041, MA0042) motifs are strongly associated (AU *p*-value > 0.95) with MLV sequences in both cell types. On the other hand, two bHLH-ZIP motifs (MA0058, MA0059) are associated only with HeLa cells and two Zn-finger GATA motifs (MA0075, MA0109) with CD34$^+$ HSCs. Among HIV sequences, three motifs are associated with HSCs (MA0095, MA0027, MA0032), and two (MA0103, MA0117) with HeLa cells.

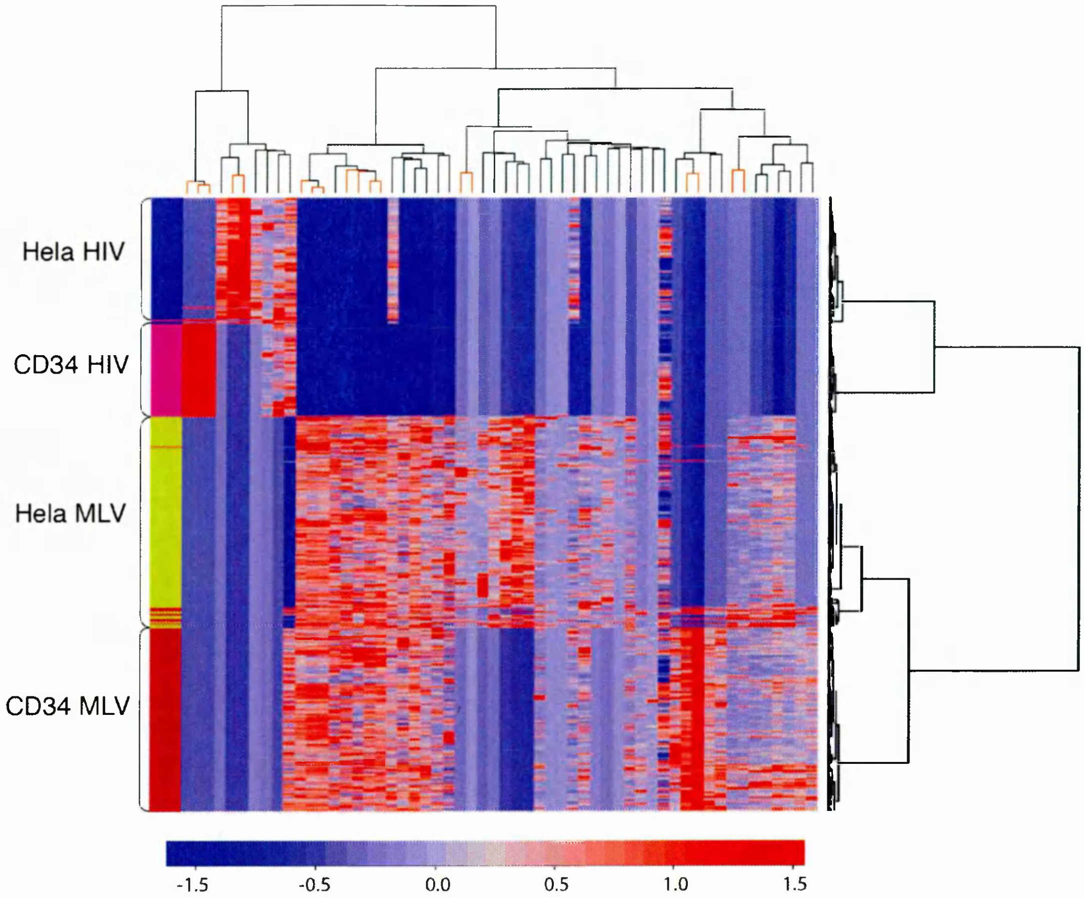


Figure 30. Comparative hierarchical cluster analysis of TFBS motifs around retroviral integrations in human HSCs and HeLa cells. The row dendrogram on the right of the heatmap splits the data set in two braches (MLV and HIV), within which HSC and HeLa sequences are clearly separated. The bootstrapped column dendrogram on the top identifies two main nodes, mainly related to the HIV and the MLV profile (see **Appendix 6.2** for a detailed dendrogram and **Appendix 7** for the complete list of motifs).

A Principal Components Analysis confirmed the results obtained by the cluster analysis. A scatter plot of the first three principal components, accounting together for 41.4% of the total variability, confirms the vector type as the first source of variability (**Figure 31**). The corresponding loadings plots show that motifs that better explain the variability are the same identified in the hierarchical cluster analysis.

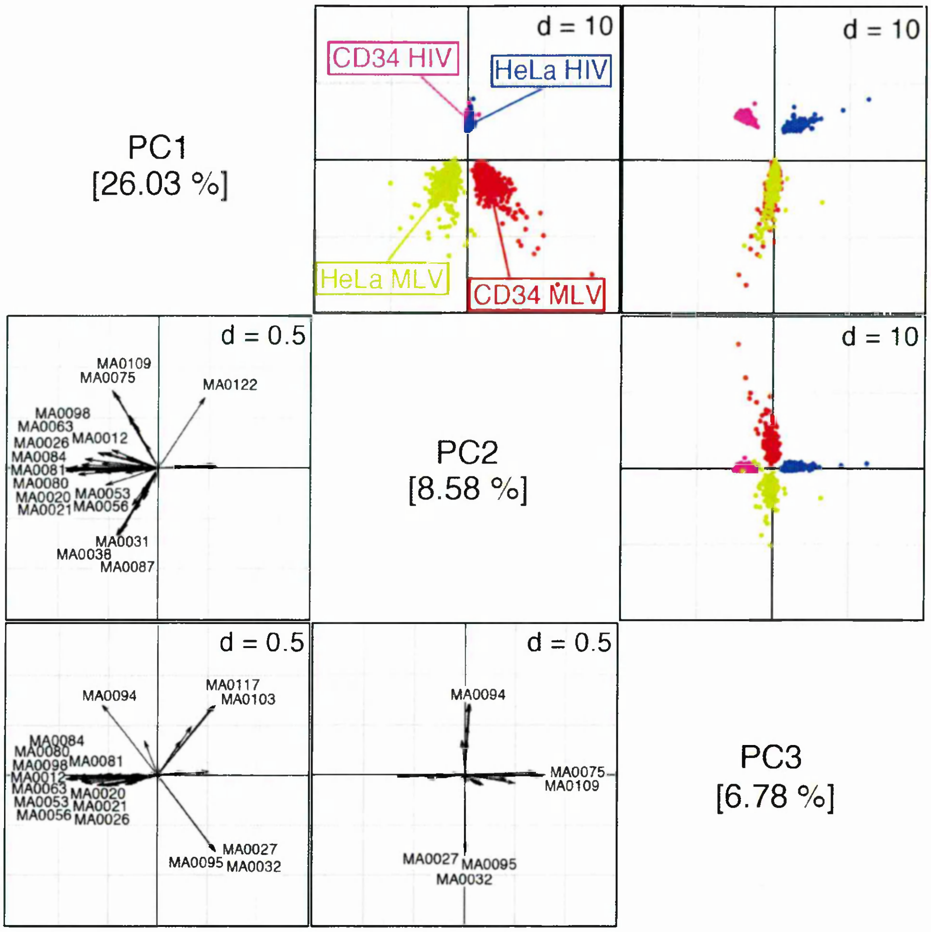


Figure 29. Comparative PCA of TFBS motifs enriched around retroviral insertions in human HSCs and HeLa cells. Principal Components Analysis of likelihood ratio values from the Clover TFBS enrichment analysis. The figure combines the scatter plots (upper -right, colored squares) of the first three principal components, accounting for 41.4% of the total variability, and the corresponding loadings plots (lower-left, black and white squares). On the scatter plots, the first source of variability is the vector type: MLV and HIV sequences distribute in opposite directions along the first component axis. The second and third sources of variability are the cell context within MLV and HIV sequences, respectively. The loadings plots shows that motifs that better explain this specific behavior are the same identified in the hierarchical cluster analysis (refer to **Figure 28** and **Appendix 6.2**).

6.2.7 MLV integrase has a crucial role in directing RV integration in TFBS-rich regions of the genome.

A recent study indicated that the MLV integrase has a crucial role in determining the RV characteristic preference for TSS-proximal regions[121]. To investigate whether the MLV integrase has also a role in directing integration to

TFBS-rich regions, we carried out a comparative analysis of the sequences flanking the insertion sites of an MLV vector[12], an HIV vector, and an HIV vector packaged with an MLV integrase[121], in Hela cells. Sequences were retrieved and re-annotated according to the criteria indicated in **Figure 22**, and analyzed for their Jaspar TFBS content by the Clover program against appropriate pair-weighted backgrounds (**Table 5**). The box plots in **Figure 29** show that MLV sequences are highly enriched in TFBSs when compared to HIV sequences (83.9 *vs.* 29.1, p < 2.2e-16, Wilcoxon rank sum test, complete statistics in **Appendix 5**). Interestingly, the MLV integrase re-directs the integration of an HIV vector (HIVmIN) towards regions significantly enriched in TFBSs, independently of the intergenic, intragenic or TSS-proximal location of the insertion site (p < 2.2e-16). Analysis of evolutionarily conserved TFBSs indicated a similar, statistically significant trend (**Figure 32**).

A two-way hierarchical cluster analysis showed that MLV and HIV sequences are defined by substantially different patterns of over-represented motifs. Both the row (right) and the bootstrapped (top) dendrograms clearly discriminate MLV and HIV sequences. Most importantly, HIVmIN sequences are associated to MLV sequences in the bootstrapped dendrogram, and share most of their characteristic TFBS motifs with them. These include a 7-motif branch (MA0099, MA0003, MA0063, MA0021, MA0026, MA0084, MA0012) that is significantly under-represented in HIV sequences in the column dendrogram (**Figure 33** and **Appendix 6.3**).

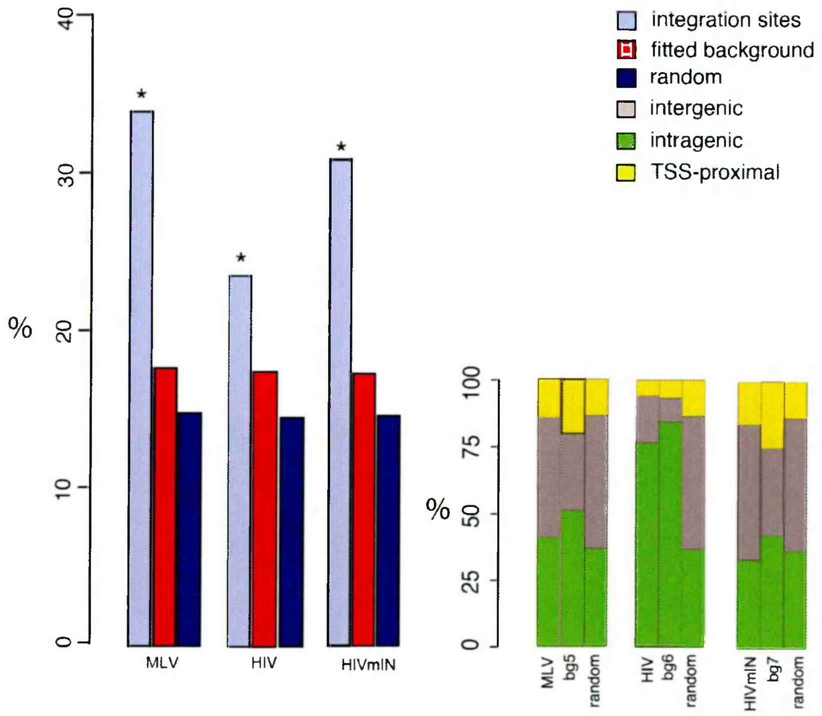


Figure 32. Evolutionarily conserved TFBSs around retroviral integration sites in human HeLa cells. Analysis of the frequency of evolutionarily conserved TFBSs in genomic sequences flanking integration sites of an MLV vector, an HIV vector and an HIV vector packaged with an MLV integrase (HIVmIN) in HeLa cells, using 188 matrices conserved in a human-mouse and/or –rat alignment (HMR Conserved Transcription Factor Binding Sites table at UCSC) as a motif database. In the upper panel, data are plotted as percentage of sequences containing at least one conserved TFBS. Each group of sequences (light blue bars) is compared to a weighted (red bars) and a random (blue bars) computational control sequence set. Asterisks highlight experimental groups that show a significant enrichment of frequency compared to their control sets (one-sided Fisher test, complete statistics in **Appendix 8.1**). In the lower panel, frequency data are broken down in three subgroups according to the integration site annotation, *i.e.*, intergenic (grey bars), TSS-proximal (yellow bars) and intragenic (green bars). The complete list of conserved motifs and their distribution over the different data sets are reported in **Appendix 8.2**.

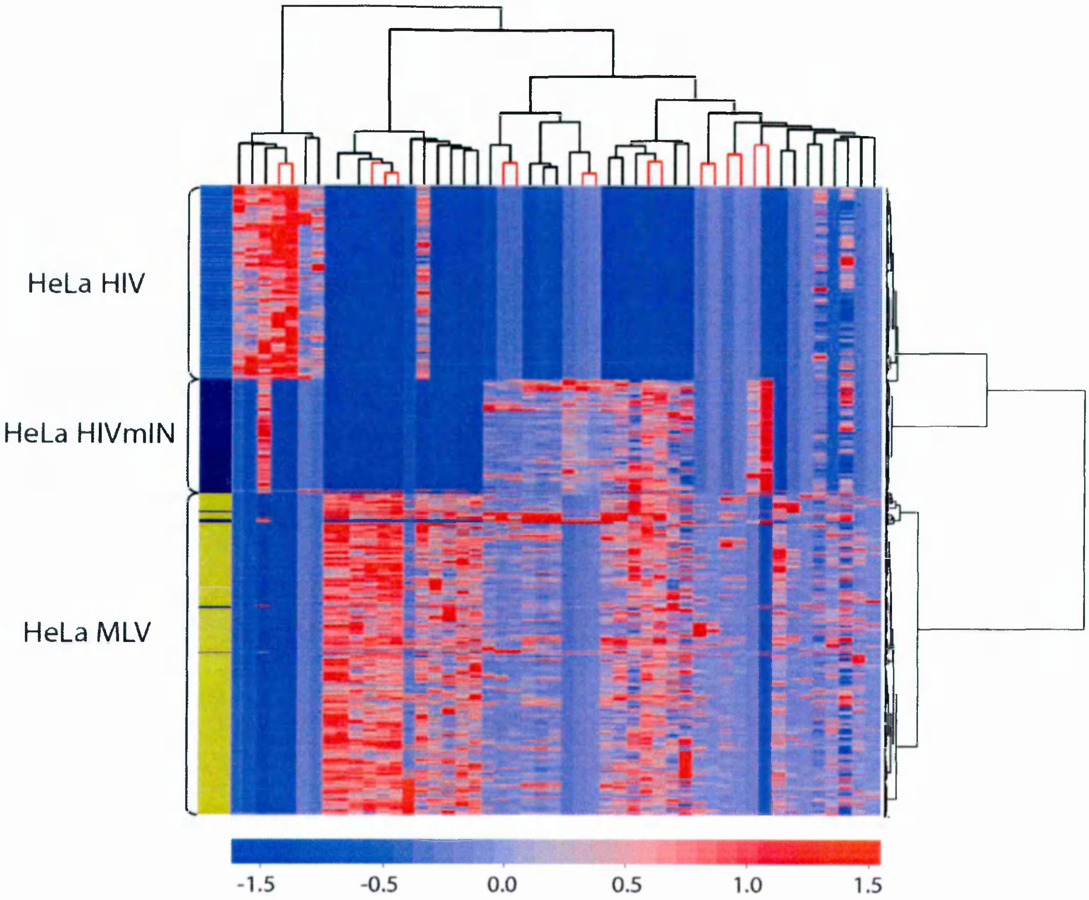


Figure 33. Two-way hierarchical cluster analysis of TFBS motifs around retroviral integrations in human HeLa cells: role of MLV integrase. The row dendrogram on the right of the heatmap clearly distinguishes MLV and HIV sequences. TFBSs are under-represented in HIV sequences compared to MLV, while sequences from the HIVmIN vector share a 7-motif branch with MLV vector in the column dendrogram (detailed dendrogram in **Appendix 6.3**; complete list of over-represented Jaspar motifs in **Appendix 7**).

A PCA (**Figure 34**) confirmed the cluster analysis. The scatter plot of the first two components (accounting for 33.8% of the total variability) reveals three main groups, corresponding to the vector type. The first component (23.1% of total variability) discriminates between MLV and HIV sequences. The second component (10.7% of total variability) differentiates HIV from HIVmIN sequences but does not distinguish MLV from HIVmIN group. The corresponding loadings plot shows a peculiar set of 8 motifs associated to MLV sequences, mostly belonging to the ETS family (MA0056, MA0098, MA0081, MA0080, MA0053, MA0020, MA0038, MA0087). A second group of seven motifs, mostly belonging to the Zn-finger C_2H_2 family, is in common between HIVmIN and MLV sequences (MA0084, MA0063, MA0021, MA0012, MA0013, MA0049). Most of these motifs were identified also by the hierarchical cluster analysis (**Figure 33**).

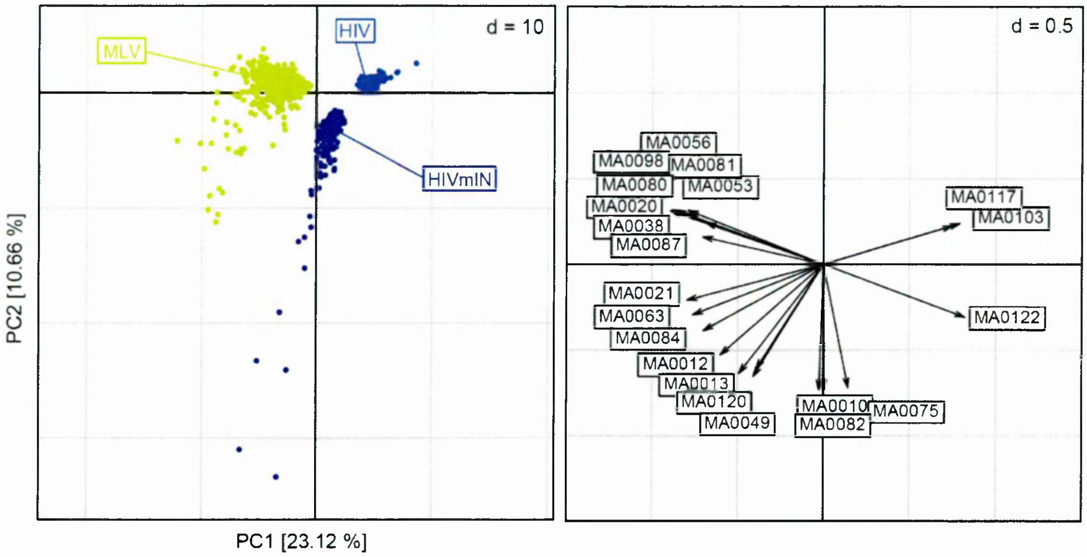


Figure 32. PCA of TFBS motifs enriched around retroviral insertions in human HeLa cells: effect of MLV integrase. Principal Components Analysis of likelihood ratio values from the Clover analysis of the 49 Jaspar motifs enriched ±1,000 bp around insertion sites of an MLV vector, an HIV vector, and an HIV vector packaged with an MLV integrase (HIVmIN) in HeLa cells. The scatter plot of the first two PCs (together accounting for 33.8% of the total variability) reveals three main independent groups, corresponding to each vector type. The first component discriminates MLV from HIV sequences; the second PC discriminates HIV from HIVmIN sequences, but not HIVmIN from MLV. The corresponding loadings plot shows a set of MLV-specific motifs (MA0056, MA0098, MA0081, MA000, MA0053, MA0020, MA0038, MA0087), and a second group of motifs in common between HIVmIN and MLV sequences (MA0084, MA0063, MA0021, MA0120, MA0013, MA0049).

7. Conclusions

Retroviral vectors, like their parental viruses, are characterized by strong biases and preferences in their integration into target cell genome, which differ significantly among different retroviral families. Gamma-retroviruses (RV) favor integration nearby TSSs and CpG islands, lentiviruses (LV) integrate preferentially within active transcription units, while alpha-retroviruses, such as ASLV, are relatively indifferent to genes or active regions in their integration site selection (section 3.2). Such differential preferences have a significant impact in predicting the risk of insertional gene activation by retroviral gene-transfer vectors. The recent adverse events following gene therapy for a blood monogenic disorder with MLV-transduced hematopoietic stem/progenitor cells (HSCs)[5-7] further accentuated the importance of understanding the molecular basis underlying retroviral integration targeting, with a particular attention to the relevant cell context. The probability of dominant activation of potentially cancer-causing genes (those involved in the control of stem-cell self-renewal, growth, and differentiation in the case of HSCs) could in fact differ significantly between RV and LV vectors, simply because of a different frequency by which they may target those genes. It has recently been suggested that LV vectors, due to their different integration preferences and LTR enhancer-free design, could be associated with a lower genotoxic risk compared to conventional RV vectors[199-201]. However, the current poor knowledge of the molecular mechanisms at the basis of target site selection represents a serious obstacle in the rational design of safer and more efficient gene transfer technology. Understanding in more detail the interactions between retroviral PICs and the human genome, the viral and cellular determinants of target site selection, and the role of functional vector components (enhancers, promoters, splicing and polyadenylation signals) in influencing integration as well as gene expression after integration, is crucial to assess the genotoxic characteristics of different vector families and designs.

7.1 Thesis conclusions

I have here reported a detailed analysis of large numbers of RV and LV integration sites in human CD34$^+$ HSCs tranduced in the same conditions used in clinical applications and analyzed short-term after infection, in the absence of selection. The general integration preferences of the two vector families were similar to those previously described for other mammalian hematopoietic and non-hematopoietic cells, and showed on average a 2-fold higher probability for RV vectors to target gene-dense regions, highly active genes, and promoter-proximal regions. More interestingly, RV, but not LV integration, occurred at high frequency (> 20%) at genomic locations (hot spots) significantly enriched in proto-oncogenes and genes involved in the control of cell proliferation and hematopoietic-specific functions.

A high frequency of hot spots, defined by statistical criteria previously applied to the definition of CISs[161], appears to be a hallmark of RV integrations in human CD34$^+$ HSCs. More than one-fifth of the RV integrations met the definition criteria, a frequency more than 7-fold higher than expected from the analysis of a randomly cloned collection of human DNA sequences, and almost 3-fold higher than that found in a collection of LV integrations of comparable size. The average extension of RV hot spots (*i.e.*, the maximum distance between all insertions within each hot spot) was well within the definition criteria, and significantly smaller than that of LV hot spots, spanning less than 10 kb in half of the cases and less than 2 kb in one-fourth of the cases. RV integration appears therefore to have high preference for restricted genomic locations, which may exhibit specific chromatin conformations or features that favor tethering of the preintegration complexes with higher probability. These features do not include gene density, proximity to promoters, or gene expression *per se*, since hot spot integrations show exactly the same preferences observed in the entire collection of RV integrations. The situation was completely different in the case of LV hot spots, which showed strikingly different characteristics with respect to the general LV integration preferences, being greatly enriched in gene-dense regions and expressed genes. These data suggest that LV integration may happen in a much wider portion of the HSC genome, and that hot spots are generated at low frequency by locations that are more favorable than others to PIC interaction, apparently those with a high density of expressed genes. Such explanation is consistent with the available evidence that LV PICs are tethered to the human genome by the widely distributed chromatin component LEDGF, and possibly by other chromatin remodeling or DNA-repair complexes (section 3.3).

Previous studies carried out in patients as well as in animal models have indicated that integrations in cancer-associated CISs and growth-controlling genes are enriched in the progeny of RV-transduced, repopulating HSCs (section 3.4). The major conclusion of these studies was that certain viral insertions lead to clonal selection of stem/ progenitor cells *in vivo*. However, the pretransplantation frequency of these insertion events was never accurately measured in the relevant cell population. Indeed, the results of this thesis indicate that a bias toward integration into or around certain categories of genes (*i.e.*, those involved in signal transduction, cell cycle, chromatin remodeling, and transcription), is already present in nontransplanted, unselected hematopoietic progenitors, and is augmented in integration hot spots. In particular, proto-oncogenes and cancer-related CISs are enriched at 3- to 5-fold the expected frequency in RV hot spots, indicating a specific preference for genomic locations containing these categories of genes. These include proto-oncogenes specifically expressed in hematopoietic progenitors and involved in hematopoietic cell neoplasia, such as LMO2 and EVI2-NF1, targeted at frequency of approximately 1:350, LYL1 and MYB (1:500), and others. Importantly, there was no difference in the number of integrations contributing to oncogene-containing hot spots between non-expanded (BM- and PB-derived) or moderately expanded (CB-derived) cell populations, arguing against the likelihood of clonal outgrowth generated in culture by insertional activation of growth-promoting genes.

A network-based pathway analysis indicated that a significant number of genes targeted by retroviral integration are functionally linked in transcription-, signal transduction-, apoptosis-, and tumorigenesis-related networks. Interestingly, genes involved in hematopoietic and system development and function were targeted at uniquely high frequency by RV integrations, and further enriched in RV hot spots, suggesting that the gene expression program of a cycling hematopoietic cell is, at least in part, instrumental in directing RV PICs to certain regions of the genome. Consistently, almost none of the genes targeted by CD34$^+$ hot spots were found in hot spots from HeLa cells, which most likely operate different regulatory networks. Kustikova et al reached similar conclusions.[15] in compiling their “insertional dominance database” (section 3.4) from the clonal progeny of serially transplanted HSCs in mice. The authors interpreted the observed over-representation of certain gene categories as the result of *in vivo* selection, rather than of intrinsic properties of the RV integration machinery. Indeed, 18% to 34% of the genes present in their IDDb (depending on the stringency of the comparative analysis) are also present among RV target gene list in this thesis, arguing against an exclusive role for *in vivo* selection in determining most of the frequency biases. A notable exception is the EVI1-MDS1 locus, which I found only once in non-transplanted cells, although it was retrieved at exceedingly high frequency *in vivo* from mice, non-human primates and humans (section 3.4). Insertional activation of such locus should therefore be considered a factor favoring clonal amplification and/or selection *in vivo* independently of the frequency by which it is targeted by RV integration before transplantation. It is worth noticing, however, that my data come from a population of hematopoietic progenitors in which the proportion of repopulating stem cells is admittedly low, leaving the possibility that stem cell-specific hot spots went undetected. Unfortunately, no integration analysis is currently possible in pretransplant, long-term repopulating stem cells, and it is therefore difficult to come to definitive conclusions as to what proportion of the biases detected in the stem cell progeny *in vivo* is due to vector preferences, and what proportion is due to *in vivo* selection.

Pursuing the idea that cell-specific transcriptional profiles are instrumental in directing RV PICs to favorable sites in the human genome, I proceeded further and investigated a possible interplay between retroviral integration and cell transcription. In strict collaboration with a bioinformatics group at IFOM-IEO campus, in Milan, I analyzed the abundance and arrangement of putative transcription factor binding sites (TFBSs) around RV and LV insertion sites. We were able to identify a previously disregarded feature of the regions targeted by RV PICs, *i.e.*, an elevated content of TFBSs. By analyzing the sequences flanking the insertion sites of RV and LV vectors in human HSCs, and of mutants carrying deletions or replacements of the LTR U3 enhancers, we showed that integration in TFBS-rich regions of the genome is peculiar to RV vectors with an entire RV LTR (either Mo-MLV- or SFFV-derived). Deletion of the U3 element strongly reduced the TFBS over-representation around the integration sites and, in turn, the relative frequency of TSS-proximal integrations. This indicated that U3 enhancer is an important determinant of RV target site selection. Statistical analyses pointed out that TFBS enrichment is only slightly dependent on the relative position or distance of the integration sites with respect to transcription units, with a modest increase in TFBS content around MLV and SFFV-MLV TSS-proximal integrations. This would suggest that selection of TFBS-rich regions may in fact underlie all known RV integration preferences, in particular those for TSSs, CpG islands and DNaseI hypersensitive sites (section 3.2), where TFBS-rich regulatory regions are highly represented.

On the other hand, TFBSs are significantly under-represented nearby LV integrations, independently of the presence of HIV U3 element in the LTR. Replacement of the HIV with an MLV U3 element in an LV vector removed this negative bias, but was not sufficient alone to introduce a positive one like that of RV vectors. No effect at all was seen, instead, when a single-copy MLV LTR was placed internally of a ΔU3-LV vector. Interestingly, performing the same analysis with a previously published collection of integration sites of MLV, HIV, and an HIV vector packaged with an MLV integrase (HIVmIN) in HeLa cells[121], we found that the MLV integrase re-directs the integration of an LV vector towards regions significantly enriched in TFBSs. Such observation, together with the effect of MLV U3 deletion, identifies MLV integrase and the LTR U3 region as the major viral determinants of the RV-specific selection of TFBS-rich target sites into the genome. Chromatin immunoprecipitation studies performed in hematopoietic cells transduced with an LV vector containing an MLV U3 enhancer showed that TFs belonging to the ETS family and YY1 are bound to PICs into the nucleus prior to integration. Indeed, unintegrated viral LTRs are transcriptionally active already 10 hours after infection, as demonstrated by cytofluorimetric analysis and RT-PCR on full-length viral mRNAs. Bound TFs are likely the cellular mediators of the LTR-associated component of the RV integration preferences. The resulting hypothesis is that cellular TFs binding MLV U3 enhancer cooperate with the integrase in directing PICs towards regulatory regions actively engaged by the transcriptional machinery. Such cooperation may be interpreted as an evolution of the mechanisms by which retrotransposons target their integration to specific genomic regions, tethered by host cell proteins (section 3.3.1). The specific domain of the retrotransposase direct tethering is lacking in the RV integrases, and may have been functionally replaced by the association with LTR-bound TFs. As a result, RV PICs are able to target a large collection of Pol II-specific regulatory elements throughout the genome, rather than few Pol III-specific elements. A mechanism coupling target site selection to gene regulation may have evolved to maximize the probability for gamma-retroviruses to be transcribed in the target cell genome, and possibly to induce expansion of infected cells by insertional deregulation of cell-specific growth regulators. HIV has evidently evolved a different strategy to target open chromatin regions while minimizing interference with the cell transcriptional machinery. Consistently, recent data emerging from large-scale studies associate HIV insertion sites with histone modifications specifically associated to transcribed chromatin rather than to enhancers, promoters and other regulatory regions[122].

Additional hints of a connection between cell-specific transcription programs and integration targeting came from the comparison between the TFBS motifs associated to RV insertions in HSCs and in the non-hematopoietic HeLa cell line.

We showed the existence of both cell-type specific, as well as common TFBS clusters between hematopoietic and epithelial cells. This suggests an indirect tethering model in which ubiquitous TFs bound within RV PICs interact with general components of the enhancer-binding complexes, such as co-regulators, chromatin remodeling or mediator complexes, rather than with specific TFs or TF families. Tethering of PICs to transcription factories, where promoters and regulatory regions are relocated by cell-specific mechanisms, may in turn be the cause of the RV-specific, high frequency of integration hot spots and preferred targeting of genes associated to cell-specific regulatory networks described above. Indeed, TFBS specifically associated with RV integration in HSCs include binding sites for hematopoietic regulators of cell proliferation, differentiation or quiescence, like LMO2, AML1/RUNX1, and FOXO3.

The different propensity of RV and LV vectors to target regulatory regions, and the frequency and characteristics of their integration hot spots herein described have an obvious impact on the design of safer gene transfer vectors for clinical applications. Although self-inactivating (ΔU3) design is predicted, also by the TFBS analysis, to improve the safety profile of MLV-based vectors, the MLV integrase remains an undesirable protagonist of RV vector tendency to target potentially dangerous regions of the genome. This thesis also shows the importance of the cell context in determining the frequency of integration into certain genomic regions, and predicts that targeting of dominantly acting proto-oncogenes may have a different likelihood in different cells. As an example, the LMO2 locus is targeted at very high frequency in HSCs, but not in T-cells, where it is not expressed, as was observed in our laboratory in the context of other integration studies. On the contrary, the use of HIV-derived vectors would minimize insertional gene activation by generally reducing integration in the proximity of regulatory enhancers and promoters. Analysis of TFBSs close to the integration sites provides therefore an additional readout to study the potential genotoxicity of vectors containing different promoters, enhancers and regulatory elements in a specific cell context.

7.2 Summary of contributions

This thesis gives significant contributions both to the fields of gene therapy and basic virology, identifying previously unrecognized features of the integration properties of gamma-retroviral and lentiviral vectors in the clinically relevant context of human hematopoietic stem/progenitor cells, and defining new parameters to predict the genotoxic risk associated to different vector designs.

a) This project was the first to retrieve and thoroughly analyze large numbers of RV and LV vector integrations from pretransplant, human CD34$^+$ HSCs. The short-term culture period guarantees that all the observed characteristics are not due to a clonal selection, but derive from retroviral specific preferences.

b) The already described general RV and LV integration preferences for active genes, gene-dense regions and, for the sole MLV, promoter proximal regions were here confirmed also in human HSCs.

c) A comparative analysis between RV and LV integration pattern revealed a 2-fold higher probability for RV vectors to target gene-dense regions, highly active genes, and promoter-proximal regions. Both RV and LV vectors tend to integrate near genes involved in the regulation of cell growth and proliferation, but only RV vectors have a specific bias for genes belonging to hematopoietic-specific pathways and/or involved in oncogenic transformation of hematopoietic tissues.

d) A large proportion (20%) of RV, but not LV, insertion sites was highly clustered to form integration hot spots. The integrations forming these hot spots recapitulate the general preferences of RV vectors in terms of gene density, gene expression and gene organization of targeted genomic regions. Instead, the list of genes surrounding RV hot spots resulted particularly enriched in cancer-associated

CISs, proto-oncogenes and genes involved in hematopoietic-specific functions, with two major implications:

d1) the bias towards certain gene categories observed in the clonal progeny of transduced HSCs *in vivo* is already detectable in non-transplanted hematopoietic progenitors, and is therefore imputable, at least in part, to intrinsic properties of the RV integration machinery, rather than exclusively to *in vivo* selection;

d2) the host cell transcriptional program might be instrumental in directing PICs to favorable sites in the genome; a comparison between CD34$^+$ RV hot spots and RV hot spots retrieved from a completely different cell type (*i.e.*, epithelial HeLa cells) confirmed this idea, since very few genes were found in common between the two target gene lists.

e) LV integrations originated just few hot spots, but these mapped to genomic loci extremely dense of active genes, independently of their function. In other words, LV hot spots simply mark those regions where the features generally attracting LV vectors (active genes, gene dense regions) are particularly enriched.

f) In addition to RV propensity for hot spots, this thesis reveals another previously unrecognized feature or RV integration, *i.e.*, an extremely high content of transcription factor binding sites (TFBSs) in genomic sequences adjacent to the insertions. Conversely, genomic regions flanking LV insertion sites are depleted of TFBSs, again highlighting different targeting strategies for the two viral families.

g) Using LTR-modified RV and LV vectors, I here demonstrate that RV enrichment in TFBS motifs depends on the presence of an RV entire U3 region. U3 deletion from both viral LTRs results in a strong reduction of the number of TFBSs, with some of them virtually "disappearing" from the integration surroundings.

Consistently, replacement of the HIV U3 enhancer with the MLV U3 element skews LV integrations towards TFBS-richer regions.

h) Chromatin immunoprecipitation experiments performed on unintegrated viral LTRs suggest that cellular TFs actually bind viral PICs prior to integration in a trascriptionally active conformation, which is not necessarily the same required for proviral expression.

i) Re-analyzing previously published integrations from an HIV vector packaged with an MLV integrase, I demonstrate that the RV integrase plays a substantial role, in cooperation with MLV U3 enhancer, in directing PICs to TFBS-rich regions.

j) A comparative analysis of RV and HIV TFBS patterns in CD34$^+$ and HeLa cells identified both cell-type specific and non-specific motifs, suggesting a targeting model in which viral PICs are tethered to chromatin by general components of enhancer-binding complexes, rather than specific TFs or TF families.

k) On the basis of the results summarized above, I proposed a model for RV integration targeting in which TFs bound within RV U3 region may cooperate with viral integrase to contact general components of the host cell transcriptional machinery, which, in turn, would tether PICs to active transcription factories, where integration finally occurs.

l) The results of this thesis have also some implications in the choice of transfer vectors for gene therapy applications. The weak propensity of LV vectors to target regulatory regions predicts a better safety profile for them with respect to the recently promoted ΔU3 RV vectors retaining MLV integrase, which was here demonstrated to have a dominant role in tethering PICs to regulatory regions. Moreover, the analyses of integration hot spots and of TFBSs described within this thesis may represent alternative readouts to study the potential genotoxicity of vectors containing different promoters, enhancers and regulatory elements in a specific cell context.

m) The content of the first part of this thesis (section 5.1) has been published in 2007 in the journal Blood[202], while the results of the second part (section 5.2) have been recently submitted to, and are at present under revision by, PLoS ONE journal.

7.3 Future research

This thesis has provided some new insights into the mechanism of target site selection by retroviral vectors, and in particular by gamma-retroviral vectors. At the same time it has raised several questions that are worth answering to get a deeper understanding on the molecular basis of integration targeting in the human genome.

A first issue regards integration hot spots. During the study, I have noticed that the frequency of RV hot spots grew progressively, following the increase of the sample size in an almost linear fashion. The situation was completely different in the case of LV hot spots, the frequency of which increased only slightly with the increase of sample size and appeared to *plateau*. This may suggest that, by analyzing a much higher number of sequences, all RV integrations could be confined to a defined subset of genomic regions, all having the appropriate features recognized by the PICs, while LV proviruses would be still spread all along active transcription units, with no particular clustering. Recently developed sequencing strategies, such as large-scale pyrosequencing, allowed achieving an approximately 100-fold increase in throughput over the classical Sanger technology[203]. Properly modified, such strategies have been successfully applied to sequence thousands of genomic integration sites at once[122,204]. Increasing my RV and LV integration numbers of 1 to 2 logs would be extremely useful to establish their trend of hot spot formation, and to collect sufficient events and add statistical robustness to the analyses described herein.

A second, extremely relevant issue is the experimental validation of the TF motifs that resulted enriched around RV and LV integrations by the TFBS analysis. This is anything but a trivial aspect, and requires a two-step effort. First of all, over-represented TFBSs must be associated to their corresponding transcription factors.

The Jaspar collection of experimentally validated matrices that we used to find enriched motifs includes in fact transcription factors from several multicellular eukaryotes, for which human orthologues must be identified. Even then, one has to keep in mind that several TFs can recognize the same binding site and, *vice versa*, two or more related motifs can be bound by a single TF. A good starting point could be the STAMP analysis we performed to merge Jaspar and UCSC Conserved results, which possibly returned the most promising factors. Once a list of putative TFs has been compiled, the second step is the demonstration that those factors actually bind genomic sequences flanking the integrations around which their binding sites were scored over-represented. A potent tool is represented by a large-scale chromatin immunoprecipitation procedure called ChIP on chip technology. In this technique, a classical ChIP is first performed, and the immunoprecipitated DNA is then amplified by LM-PCR, labeled with a fluorescent tag and hybridized to custom-designed microarrays. Specific "integrome" chips can be synthesized, spotting thousands of oligos to cover the integration surroundings, and used to reveal TF binding in proximity of the insertion sites. Single TFs can be validated one after the other in this way.

"Integrome" chips open another research chapter, *i.e.*, the study of the general chromatin *status* around retroviral insertion sites. At present, the issue has been addressed from an entirely bioinformatics point of view by the Bushman group[122], and has revealed some interesting connections between LV integration and histone post-translational modifications. The study was a clear indication that this is a worthwhile question to tackle. Dozens of histone modifications have been now described, often in conjunction with certain transcriptional control processes[205,206]; the identification of those histone acetylations or methylations specifically associated to insertion sites would logically link them to the related regulation processes and to cellular components participating in them.

Acronyms and abbreviations

ADA: adenosine deaminase

ASLV: avian sarcoma leucosis virus

BM: bone marrow

CA: capsid

CB: cord blood

CGD: chronic granulomatous disease

ChIP: chromatin immunoprecipitation

CIS: common insertion site

DAVID: Database for Annotation, Visualization and Integrated Discovery

EASE: Expression Analysis Systematic Explorer

GO: Gene Ontology

HIV: human immunodeficiency virus

HSCs: hematopoietic stem/progenitor cells

IDDb: insertion dominance database

IN: integrase

IPA: Ingenuity Pathways Analysis

IPKB: Ingenuity Pathways Knowledge Base

LAM-PCR: linear amplification-mediated PCR

LM-PCR: linker-mediated PCR

LTR: long terminal repeat

LV: lentivirus

MA: matrix

MOI: multiplicity of infection

Mo-MLV: Moloney murine leukemia virus

NRE: negative response element

PB: peripheral blood

PC: principal component

PCA: Principal Components Analysis

PCR: polymerase chain reaction

PIC: preintegration complex

PR: protease

RT: reverse transcriptase

RTC: reverse transcription complex

RV: gamma-retrovirus

SCID: severe combined immunodeficiency

SFFV: spleen focus forming virus

SIV: simian immunodeficiency virus

TF: transcription factor

TFBS: transcription factor binding site

TSS: transcription start site

UCR: upstream conserved region

VSV-G: vesicular stomatitis virus glycoprotein

References

1. 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.
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. 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.
4. 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.
5. 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.
6. Gansbacher B. Report of a second serious adverse event in a clinical trial of gene therapy for X-linked severe combined immune deficiency (X-SCID). Position of the European Society of Gene Therapy (ESGT). J Gene Med. 2003;5:261-262.
7. 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.
8. 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.
9. 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.
10. Laufs S, Gentner B, Nagy KZ, et al. Retroviral vector integration occurs in preferred genomic targets of human bone marrow-repopulating cells. Blood. 2003;101:2191-2198.
11. 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.
12. 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.

13. 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.

14. Suzuki T, Shen H, Akagi K, et al. New genes involved in cancer identified by retroviral tagging. Nat Genet. 2002;32:166-174.

15. Kustikova O, Fehse B, Modlich U, et al. Clonal dominance of hematopoietic stem cells triggered by retroviral gene marking. Science. 2005;308:1171-1174.

16. 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.

17. Coffin JM. Retroviruses: Cold Spring Harbor, New York: Cold Spring Harbor Laboratory Press; 1997.

18. Lee MS, Craigie R. Protection of retroviral DNA from autointegration: involvement of a cellular factor. Proc Natl Acad Sci U S A. 1994;91:9823-9827.

19. Suzuki Y, Craigie R. Regulatory mechanisms by which barrier-to-autointegration factor blocks autointegration and stimulates intermolecular integration of Moloney murine leukemia virus preintegration complexes. J Virol. 2002;76:12376-12380.

20. Mansharamani M, Graham DR, Monie D, et al. Barrier-to-autointegration factor BAF binds p55 Gag and matrix and is a host component of human immunodeficiency virus type 1 virions. J Virol. 2003;77:13084-13092.

21. Chen H, Engelman A. The barrier-to-autointegration protein is a host factor for HIV type 1 integration. Proc Natl Acad Sci U S A. 1998;95:15270-15274.

22. Lin CW, Engelman A. The barrier-to-autointegration factor is a component of functional human immunodeficiency virus type 1 preintegration complexes. J Virol. 2003;77:5030-5036.

23. Butler SL, Hansen MS, Bushman FD. A quantitative assay for HIV DNA integration in vivo. Nat Med. 2001;7:631-634.

24. Butler SL, Johnson EP, Bushman FD. Human immunodeficiency virus cDNA metabolism: notable stability of two-long terminal repeat circles. J Virol. 2002;76:3739-3747.

25. O'Doherty U, Swiggard WJ, Jeyakumar D, McGain D, Malim MH. A sensitive, quantitative assay for human immunodeficiency virus type 1 integration. J Virol. 2002;76:10942-10950.

26. Nermut MV, Fassati A. Structural analyses of purified human immunodeficiency virus type 1 intracellular reverse transcription complexes. J Virol. 2003;77:8196-8206.

27. Bukrinsky MI, Sharova N, McDonald TL, Pushkarskaya T, Tarpley WG, Stevenson M. Association of integrase, matrix, and reverse transcriptase antigens of human immunodeficiency virus type 1 with viral nucleic acids following acute infection. Proc Natl Acad Sci U S A. 1993;90:6125-6129.
28. Miller MD, Farnet CM, Bushman FD. Human immunodeficiency virus type 1 preintegration complexes: studies of organization and composition. J Virol. 1997;71:5382-5390.
29. Farnet CM, Haseltine WA. Determination of viral proteins present in the human immunodeficiency virus type 1 preintegration complex. J Virol. 1991;65:1910-1915.
30. Karageorgos L, Li P, Burrell C. Characterization of HIV replication complexes early after cell-to-cell infection. AIDS Res Hum Retroviruses. 1993;9:817-823.
31. Bowerman B, Brown PO, Bishop JM, Varmus HE. A nucleoprotein complex mediates the integration of retroviral DNA. Genes Dev. 1989;3:469-478.
32. Fassati A, Goff SP. Characterization of intracellular reverse transcription complexes of Moloney murine leukemia virus. J Virol. 1999;73:8919-8925.
33. Suzuki Y, Craigie R. The road to chromatin - nuclear entry of retroviruses. Nat Rev Microbiol. 2007;5:187-196.
34. Mulder LC, Chakrabarti LA, Muesing MA. Interaction of HIV-1 integrase with DNA repair protein hRad18. J Biol Chem. 2002;277:27489-27493.
35. Daniel R, Katz RA, Skalka AM. A role for DNA-PK in retroviral DNA integration. Science. 1999;284:644-647.
36. Ha HC, Juluri K, Zhou Y, Leung S, Hermankova M, Snyder SH. Poly(ADP-ribose) polymerase-1 is required for efficient HIV-1 integration. Proc Natl Acad Sci U S A. 2001;98:3364-3368.
37. Kalpana GV, Marmon S, Wang W, Crabtree GR, Goff SP. Binding and stimulation of HIV-1 integrase by a human homolog of yeast transcription factor SNF5. Science. 1994;266:2002-2006.
38. Violot S, Hong SS, Rakotobe D, et al. The human polycomb group EED protein interacts with the integrase of human immunodeficiency virus type 1. J Virol. 2003;77:12507-12522.
39. Li L, Yoder K, Hansen MS, Olvera J, Miller MD, Bushman FD. Retroviral cDNA integration: stimulation by HMG I family proteins. J Virol. 2000;74:10965-10974.
40. Farnet CM, Bushman FD. HIV-1 cDNA integration: requirement of HMG I(Y) protein for function of preintegration complexes in vitro. Cell. 1997;88:483-492.

41. Li L, Farnet CM, Anderson WF, Bushman FD. Modulation of activity of Moloney murine leukemia virus preintegration complexes by host factors in vitro. J Virol. 1998;72:2125-2131.
42. Cherepanov P, Maertens G, Proost P, et al. HIV-1 integrase forms stable tetramers and associates with LEDGF/p75 protein in human cells. J Biol Chem. 2003;278:372-381.
43. Llano M, Vanegas M, Fregoso O, et al. LEDGF/p75 determines cellular trafficking of diverse lentiviral but not murine oncoretroviral integrase proteins and is a component of functional lentiviral preintegration complexes. J Virol. 2004;78:9524-9537.
44. Ciuffi A, Llano M, Poeschla E, et al. A role for LEDGF/p75 in targeting HIV DNA integration. Nat Med. 2005;11:1287-1289.
45. Hombrouck A, De Rijck J, Hendrix J, et al. Virus evolution reveals an exclusive role for LEDGF/p75 in chromosomal tethering of HIV. PLoS Pathog. 2007;3:e47.
46. Shun MC, Raghavendra NK, Vandegraaff N, et al. LEDGF/p75 functions downstream from preintegration complex formation to effect gene-specific HIV-1 integration. Genes Dev. 2007;21:1767-1778.
47. Ramji DP, Foka P. CCAAT/enhancer-binding proteins: structure, function and regulation. Biochem J. 2002;365:561-575.
48. Wahlers A, Kustikova O, Zipfel PF, et al. Upstream conserved sequences of mouse leukemia viruses are important for high transgene expression in lymphoid and hematopoietic cells. Mol Ther. 2002;6:313-320.
49. Tsukiyama T, Ueda H, Hirose S, Niwa O. Embryonal long terminal repeat-binding protein is a murine homolog of FTZ-F1, a member of the steroid receptor superfamily. Mol Cell Biol. 1992;12:1286-1291.
50. Flanagan JR, Krieg AM, Max EE, Khan AS. Negative control region at the 5' end of murine leukemia virus long terminal repeats. Mol Cell Biol. 1989;9:739-746.
51. Rao A, Luo C, Hogan PG. Transcription factors of the NFAT family: regulation and function. Annu Rev Immunol. 1997;15:707-747.
52. DeFranco D, Yamamoto KR. Two different factors act separately or together to specify functionally distinct activities at a single transcriptional enhancer. Mol Cell Biol. 1986;6:993-1001.
53. Miksicek R, Heber A, Schmid W, et al. Glucocorticoid responsiveness of the transcriptional enhancer of Moloney murine sarcoma virus. Cell. 1986;46:283-290.
54. Granger SW, Fan H. In vivo footprinting of the enhancer sequences in the upstream long terminal repeat of Moloney murine leukemia virus: differential binding of nuclear factors in different cell types. J Virol. 1998;72:8961-8970.

55. Speck NA, Baltimore D. Six distinct nuclear factors interact with the 75-base-pair repeat of the Moloney murine leukemia virus enhancer. Mol Cell Biol. 1987;7:1101-1110.

56. Reisman D. Nuclear factor-1 (NF-1) binds to multiple sites within the transcriptional enhancer of Moloney murine leukemia virus. FEBS Lett. 1990;277:209-211.

57. Nye JA, Petersen JM, Gunther CV, Jonsen MD, Graves BJ. Interaction of murine ets-1 with GGA-binding sites establishes the ETS domain as a new DNA-binding motif. Genes Dev. 1992;6:975-990.

58. Manley NR, O'Connell M, Sun W, Speck NA, Hopkins N. Two factors that bind to highly conserved sequences in mammalian type C retroviral enhancers. J Virol. 1993;67:1967-1975.

59. Verger A, Duterque-Coquillaud M. When Ets transcription factors meet their partners. Bioessays. 2002;24:362-370.

60. Wang S, Wang Q, Crute BE, Melnikova IN, Keller SR, Speck NA. Cloning and characterization of subunits of the T-cell receptor and murine leukemia virus enhancer core-binding factor. Mol Cell Biol. 1993;13:3324-3339.

61. Sun W, Graves BJ, Speck NA. Transactivation of the Moloney murine leukemia virus and T-cell receptor beta-chain enhancers by cbf and ets requires intact binding sites for both proteins. J Virol. 1995;69:4941-4949.

62. Sun W, O'Connell M, Speck NA. Characterization of a protein that binds multiple sequences in mammalian type C retrovirus enhancers. J Virol. 1993;67:1976-1986.

63. Chao SH, Walker JR, Chanda SK, Gray NS, Caldwell JS. Identification of homeodomain proteins, PBX1 and PREP1, involved in the transcription of murine leukemia virus. Mol Cell Biol. 2003;23:831-841.

64. Kretzschmar M, Meisterernst M, Scheidereit C, Li G, Roeder RG. Transcriptional regulation of the HIV-1 promoter by NF-kappa B in vitro. Genes Dev. 1992;6:761-774.

65. Nabel G, Baltimore D. An inducible transcription factor activates expression of human immunodeficiency virus in T cells. Nature. 1987;326:711-713.

66. Fujita T, Nolan GP, Ghosh S, Baltimore D. Independent modes of transcriptional activation by the p50 and p65 subunits of NF-kappa B. Genes Dev. 1992;6:775-787.

67. Perkins ND, Edwards NL, Duckett CS, Agranoff AB, Schmid RM, Nabel GJ. A cooperative interaction between NF-kappa B and Sp1 is required for HIV-1 enhancer activation. Embo J. 1993;12:3551-3558.

68. Cooney AJ, Tsai SY, O'Malley BW, Tsai MJ. Chicken ovalbumin upstream promoter transcription factor binds to a negative regulatory

region in the human immunodeficiency virus type 1 long terminal repeat. J Virol. 1991;65:2853-2860.

69. Lu YC, Touzjian N, Stenzel M, Dorfman T, Sodroski JG, Haseltine WA. Identification of cis-acting repressive sequences within the negative regulatory element of human immunodeficiency virus type 1. J Virol. 1990;64:5226-5229.

70. Markovitz DM, Hannibal MC, Smith MJ, Cossman R, Nabel GJ. Activation of the human immunodeficiency virus type 1 enhancer is not dependent on NFAT-1. J Virol. 1992;66:3961-3965.

71. Michael NL, D'Arcy L, Ehrenberg PK, Redfield RR. Naturally occurring genotypes of the human immunodeficiency virus type 1 long terminal repeat display a wide range of basal and Tat-induced transcriptional activities. J Virol. 1994;68:3163-3174.

72. Sieweke MH, Tekotte H, Jarosch U, Graf T. Cooperative interaction of ets-1 with USF-1 required for HIV-1 enhancer activity in T cells. Embo J. 1998;17:1728-1739.

73. Romerio F, Gabriel MN, Margolis DM. Repression of human immunodeficiency virus type 1 through the novel cooperation of human factors YY1 and LSF. J Virol. 1997;71:9375-9382.

74. Yoon JB, Li G, Roeder RG. Characterization of a family of related cellular transcription factors which can modulate human immunodeficiency virus type 1 transcription in vitro. Mol Cell Biol. 1994;14:1776-1785.

75. Ou SH, Wu F, Harrich D, Garcia-Martinez LF, Gaynor RB. Cloning and characterization of a novel cellular protein, TDP-43, that binds to human immunodeficiency virus type 1 TAR DNA sequence motifs. J Virol. 1995;69:3584-3596.

76. Kato H, Horikoshi M, Roeder RG. Repression of HIV-1 transcription by a cellular protein. Science. 1991;251:1476-1479.

77. Dayton AI, Sodroski JG, Rosen CA, Goh WC, Haseltine WA. The trans-activator gene of the human T cell lymphotropic virus type III is required for replication. Cell. 1986;44:941-947.

78. Fisher AG, Feinberg MB, Josephs SF, et al. The trans-activator gene of HTLV-III is essential for virus replication. Nature. 1986;320:367-371.

79. Okamoto H, Sheline CT, Corden JL, Jones KA, Peterlin BM. Trans-activation by human immunodeficiency virus Tat protein requires the C-terminal domain of RNA polymerase II. Proc Natl Acad Sci U S A. 1996;93:11575-11579.

80. Chiu YL, Ho CK, Saha N, Schwer B, Shuman S, Rana TM. Tat stimulates cotranscriptional capping of HIV mRNA. Mol Cell. 2002;10:585-597.

81. Berro R, Kehn K, de la Fuente C, et al. Acetylated Tat regulates human immunodeficiency virus type 1 splicing through its interaction with the splicing regulator p32. J Virol. 2006;80:3189-3204.
82. 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;2:426-436.
83. 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;39:935-944.
84. Mann R, Mulligan RC, Baltimore D. Construction of a retrovirus packaging mutant and its use to produce helper-free defective retrovirus. Cell. 1983;33:153-159.
85. Watanabe S, Temin HM. Construction of a helper cell line for avian reticuloendotheliosis virus cloning vectors. Mol Cell Biol. 1983;3:2241-2249.
86. Yu SF, von Ruden T, Kantoff PW, et al. Self-inactivating retroviral vectors designed for transfer of whole genes into mammalian cells. Proc Natl Acad Sci U S A. 1986;83:3194-3198.
87. Miyoshi H, Blomer U, Takahashi M, Gage FH, Verma IM. Development of a self-inactivating lentivirus vector. J Virol. 1998;72:8150-8157.
88. Zufferey R, Dull T, Mandel RJ, et al. Self-inactivating lentivirus vector for safe and efficient in vivo gene delivery. J Virol. 1998;72:9873-9880.
89. Yang Y, Vanin EF, Whitt MA, et al. Inducible, high-level production of infectious murine leukemia retroviral vector particles pseudotyped with vesicular stomatitis virus G envelope protein. Hum Gene Ther. 1995;6:1203-1213.
90. Bartz SR, Vodicka MA. Production of high-titer human immunodeficiency virus type 1 pseudotyped with vesicular stomatitis virus glycoprotein. Methods. 1997;12:337-342.
91. Gaspar HB, Thrasher AJ. Gene therapy for severe combined immunodeficiencies. Expert Opin Biol Ther. 2005;5:1175-1182.
92. Lebensburger J, Persons DA. Progress toward safe and effective gene therapy for beta-thalassemia and sickle cell disease. Curr Opin Drug Discov Devel. 2008;11:225-232.
93. Capotondo A, Cesani M, Pepe S, et al. Safety of arylsulfatase A overexpression for gene therapy of metachromatic leukodystrophy. Hum Gene Ther. 2007;18:821-836.
94. McIntyre C, Derrick Roberts AL, Ranieri E, Clements PR, Byers S, Anson DS. Lentiviral-mediated gene therapy for murine mucopolysaccharidosis type IIIA. Mol Genet Metab. 2008;93:411-418.

95. Traas AM, Wang P, Ma X, et al. Correction of clinical manifestations of canine mucopolysaccharidosis I with neonatal retroviral vector gene therapy. Mol Ther. 2007;15:1423-1431.
96. Mavilio F, Pellegrini G, Ferrari S, et al. Correction of junctional epidermolysis bullosa by transplantation of genetically modified epidermal stem cells. Nat Med. 2006;12:1397-1402.
97. Brown BD, Cantore A, Annoni A, et al. A microRNA-regulated lentiviral vector mediates stable correction of hemophilia B mice. Blood. 2007;110:4144-4152.
98. Matsui H, Shibata M, Brown B, et al. Ex vivo gene therapy for hemophilia A that enhances safe delivery and sustained in vivo factor VIII expression from lentivirally engineered endothelial progenitors. Stem Cells. 2007;25:2660-2669.
99. Bank A, Dorazio R, Leboulch P. A phase I/II clinical trial of beta-globin gene therapy for beta-thalassemia. Ann N Y Acad Sci. 2005;1054:308-316.
100. Levine BL, Humeau LM, Boyer J, et al. Gene transfer in humans using a conditionally replicating lentiviral vector. Proc Natl Acad Sci U S A. 2006;103:17372-17377.
101. Manilla P, Rebello T, Afable C, et al. Regulatory considerations for novel gene therapy products: a review of the process leading to the first clinical lentiviral vector. Hum Gene Ther. 2005;16:17-25.
102. http://www.wiley.co.uk/genetherapy/clinical/ Gene Therapy Clinical Trials Worldwide, provided by the Journal of Gene Medicine. Accessed June 2008.
103. Pryciak PM, Varmus HE. Nucleosomes, DNA-binding proteins, and DNA sequence modulate retroviral integration target site selection. Cell. 1992;69:769-780.
104. Pruss D, Reeves R, Bushman FD, Wolffe AP. The influence of DNA and nucleosome structure on integration events directed by HIV integrase. J Biol Chem. 1994;269:25031-25041.
105. Dudley JP. Tag, you're hit: retroviral insertions identify genes involved in cancer. Trends Mol Med. 2003;9:43-45.
106. Li Z, Dullmann J, Schiedlmeier B, et al. Murine leukemia induced by retroviral gene marking. Science. 2002;296:497.
107. Warren AJ, Colledge WH, Carlton MB, Evans MJ, Smith AJ, Rabbitts TH. The oncogenic cysteine-rich LIM domain protein rbtn2 is essential for erythroid development. Cell. 1994;78:45-57.
108. Fisch P, Boehm T, Lavenir I, et al. T-cell acute lymphoblastic lymphoma induced in transgenic mice by the RBTN1 and RBTN2 LIM-domain genes. Oncogene. 1992;7:2389-2397.
109. Larson RC, Fisch P, Larson TA, et al. T cell tumours of disparate phenotype in mice transgenic for Rbtn-2. Oncogene. 1994;9:3675-3681.

110. 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;350:913-922.
111. Dave UP, Jenkins NA, Copeland NG. Gene therapy insertional mutagenesis insights. Science. 2004;303:333.
112. Woods NB, Bottero V, Schmidt M, von Kalle C, Verma IM. Gene therapy: therapeutic gene causing lymphoma. Nature. 2006;440:1123.
113. Pike-Overzet K, de Ridder D, Weerkamp F, et al. Ectopic retroviral expression of LMO2, but not IL2Rgamma, blocks human T-cell development from CD34+ cells: implications for leukemogenesis in gene therapy. Leukemia. 2007;21:754-763.
114. 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.
115. 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;103:11730-11735.
116. Mooslehner K, Karls U, Harbers K. Retroviral integration sites in transgenic Mov mice frequently map in the vicinity of transcribed DNA regions. J Virol. 1990;64:3056-3058.
117. Scherdin U, Rhodes K, Breindl M. Transcriptionally active genome regions are preferred targets for retrovirus integration. J Virol. 1990;64:907-912.
118. Weidhaas JB, Angelichio EL, Fenner S, Coffin JM. Relationship between retroviral DNA integration and gene expression. J Virol. 2000;74:8382-8389.
119. 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;72:4005-4014.
120. Bester AC, Schwartz M, Schmidt M, et al. Fragile sites are preferential targets for integrations of MLV vectors in gene therapy. Gene Ther. 2006;13:1057-1059.
121. Lewinski MK, Yamashita M, Emerman M, et al. Retroviral DNA integration: viral and cellular determinants of target-site selection. PLoS Pathog. 2006;2:e60.
122. Wang GP, Ciuffi A, Leipzig J, Berry CC, Bushman FD. HIV integration site selection: analysis by massively parallel pyrosequencing reveals association with epigenetic modifications. Genome Res. 2007;17:1186-1194.
123. Bushman F, Lewinski M, Ciuffi A, et al. Genome-wide analysis of retroviral DNA integration. Nat Rev Microbiol. 2005;3:848-858.

124. Wu X, Li Y, Crise B, Burgess SM, Munroe DJ. Weak palindromic consensus sequences are a common feature found at the integration target sites of many retroviruses. J Virol. 2005;79:5211-5214.
125. Berry C, Hannenhalli S, Leipzig J, Bushman FD. Selection of target sites for mobile DNA integration in the human genome. PLoS Comput Biol. 2006;2:e157.
126. Lander ES, Linton LM, Birren B, et al. Initial sequencing and analysis of the human genome. Nature. 2001;409:860-921.
127. 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.
128. Barr SD, Ciuffi A, Leipzig J, Shinn P, Ecker JR, Bushman FD. HIV integration site selection: targeting in macrophages and the effects of different routes of viral entry. Mol Ther. 2006;14:218-225.
129. Beard BC, Dickerson D, Beebe K, et al. Comparison of HIV-derived lentiviral and MLV-based gammaretroviral vector integration sites in primate repopulating cells. Mol Ther. 2007;15:1356-1365.
130. Ciuffi A, Mitchell RS, Hoffmann C, et al. Integration site selection by HIV-based vectors in dividing and growth-arrested IMR-90 lung fibroblasts. Mol Ther. 2006;13:366-373.
131. Narezkina A, Taganov KD, Litwin S, et al. Genome-wide analyses of avian sarcoma virus integration sites. J Virol. 2004;78:11656-11663.
132. Kirchner J, Connolly CM, Sandmeyer SB. Requirement of RNA polymerase III transcription factors for in vitro position-specific integration of a retroviruslike element. Science. 1995;267:1488-1491.
133. Mou Z, Kenny AE, Curcio MJ. Hos2 and Set3 promote integration of Ty1 retrotransposons at tRNA genes in Saccharomyces cerevisiae. Genetics. 2006;172:2157-2167.
134. Zhu Y, Zou S, Wright DA, Voytas DF. Tagging chromatin with retrotransposons: target specificity of the Saccharomyces Ty5 retrotransposon changes with the chromosomal localization of Sir3p and Sir4p. Genes Dev. 1999;13:2738-2749.
135. Zhu Y, Dai J, Fuerst PG, Voytas DF. Controlling integration specificity of a yeast retrotransposon. Proc Natl Acad Sci U S A. 2003;100:5891-5895.
136. Bushman FD. Tethering human immunodeficiency virus 1 integrase to a DNA site directs integration to nearby sequences. Proc Natl Acad Sci U S A. 1994;91:9233-9237.
137. Goulaouic H, Chow SA. Directed integration of viral DNA mediated by fusion proteins consisting of human immunodeficiency virus type 1 integrase and Escherichia coli LexA protein. J Virol. 1996;70:37-46.

138. Tan W, Zhu K, Segal DJ, Barbas CF, 3rd, Chow SA. Fusion proteins consisting of human immunodeficiency virus type 1 integrase and the designed polydactyl zinc finger protein E2C direct integration of viral DNA into specific sites. J Virol. 2004;78:1301-1313.
139. Bushman FD, Miller MD. Tethering human immunodeficiency virus type 1 preintegration complexes to target DNA promotes integration at nearby sites. J Virol. 1997;71:458-464.
140. Katz RA, Merkel G, Skalka AM. Targeting of retroviral integrase by fusion to a heterologous DNA binding domain: in vitro activities and incorporation of a fusion protein into viral particles. Virology. 1996;217:178-190.
141. Tan W, Dong Z, Wilkinson TA, Barbas CF, 3rd, Chow SA. Human immunodeficiency virus type 1 incorporated with fusion proteins consisting of integrase and the designed polydactyl zinc finger protein E2C can bias integration of viral DNA into a predetermined chromosomal region in human cells. J Virol. 2006;80:1939-1948.
142. Engelman A, Cherepanov P. The lentiviral integrase binding protein LEDGF/p75 and HIV-1 replication. PLoS Pathog. 2008;4:e1000046.
143. Emiliani S, Mousnier A, Busschots K, et al. Integrase mutants defective for interaction with LEDGF/p75 are impaired in chromosome tethering and HIV-1 replication. J Biol Chem. 2005;280:25517-25523.
144. Turlure F, Devroe E, Silver PA, Engelman A. Human cell proteins and human immunodeficiency virus DNA integration. Front Biosci. 2004;9:3187-3208.
145. Qiu C, Sawada K, Zhang X, Cheng X. The PWWP domain of mammalian DNA methyltransferase Dnmt3b defines a new family of DNA-binding folds. Nat Struct Biol. 2002;9:217-224.
146. Stec I, Nagl SB, van Ommen GJ, den Dunnen JT. The PWWP domain: a potential protein-protein interaction domain in nuclear proteins influencing differentiation? FEBS Lett. 2000;473:1-5.
147. Izumoto Y, Kuroda T, Harada H, Kishimoto T, Nakamura H. Hepatoma-derived growth factor belongs to a gene family in mice showing significant homology in the amino terminus. Biochem Biophys Res Commun. 1997;238:26-32.
148. Ikegame K, Yamamoto M, Kishima Y, et al. A new member of a hepatoma-derived growth factor gene family can translocate to the nucleus. Biochem Biophys Res Commun. 1999;266:81-87.
149. Cherepanov P, Devroe E, Silver PA, Engelman A. Identification of an evolutionarily conserved domain in human lens epithelium-derived growth factor/transcriptional co-activator p75 (LEDGF/p75) that binds HIV-1 integrase. J Biol Chem. 2004;279:48883-48892.

150. Llano M, Vanegas M, Hutchins N, Thompson D, Delgado S, Poeschla EM. Identification and characterization of the chromatin-binding domains of the HIV-1 integrase interactor LEDGF/p75. J Mol Biol. 2006;360:760-773.
151. Turlure F, Maertens G, Rahman S, Cherepanov P, Engelman A. A tripartite DNA-binding element, comprised of the nuclear localization signal and two AT-hook motifs, mediates the association of LEDGF/p75 with chromatin in vivo. Nucleic Acids Res. 2006;34:1653-1675.
152. Fatma N, Singh DP, Shinohara T, Chylack LT, Jr. Transcriptional regulation of the antioxidant protein 2 gene, a thiol-specific antioxidant, by lens epithelium-derived growth factor to protect cells from oxidative stress. J Biol Chem. 2001;276:48899-48907.
153. Singh DP, Fatma N, Kimura A, Chylack LT, Jr., Shinohara T. LEDGF binds to heat shock and stress-related element to activate the expression of stress-related genes. Biochem Biophys Res Commun. 2001;283:943-955.
154. Maertens G, Cherepanov P, Pluymers W, et al. LEDGF/p75 is essential for nuclear and chromosomal targeting of HIV-1 integrase in human cells. J Biol Chem. 2003;278:33528-33539.
155. Llano M, Delgado S, Vanegas M, Poeschla EM. Lens epithelium-derived growth factor/p75 prevents proteasomal degradation of HIV-1 integrase. J Biol Chem. 2004;279:55570-55577.
156. Marshall HM, Ronen K, Berry C, et al. Role of PSIP1/LEDGF/p75 in lentiviral infectivity and integration targeting. PLoS ONE. 2007;2:e1340.
157. Sutherland HG, Newton K, Brownstein DG, et al. Disruption of Ledgf/Psip1 results in perinatal mortality and homeotic skeletal transformations. Mol Cell Biol. 2006;26:7201-7210.
158. Kohn DB, Sadelain M, Dunbar C, et al. American Society of Gene Therapy (ASGT) ad hoc subcommittee on retroviral-mediated gene transfer to hematopoietic stem cells. Mol Ther. 2003;8:180-187.
159. Kiem HP, Sellers S, Thomasson B, et al. Long-term clinical and molecular follow-up of large animals receiving retrovirally transduced stem and progenitor cells: no progression to clonal hematopoiesis or leukemia. Mol Ther. 2004;9:389-395.
160. Bonini C, Grez M, Traversari C, et al. Safety of retroviral gene marking with a truncated NGF receptor. Nat Med. 2003;9:367-369.
161. Wu X, Luke BT, Burgess SM. Redefining the common insertion site. Virology. 2006;344:292-295.
162. Akagi K, Suzuki T, Stephens RM, Jenkins NA, Copeland NG. RTCGD: retroviral tagged cancer gene database. Nucleic Acids Res. 2004;32:D523-527.

163. 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.
164. Nishikata I, Sasaki H, Iga M, et al. A novel EVI1 gene family, MEL1, lacking a PR domain (MEL1S) is expressed mainly in t(1;3)(p36;q21)-positive AML and blocks G-CSF-induced myeloid differentiation. Blood. 2003;102:3323-3332.
165. Minakuchi M, Kakazu N, Gorrin-Rivas MJ, et al. Identification and characterization of SEB, a novel protein that binds to the acute undifferentiated leukemia-associated protein SET. Eur J Biochem. 2001;268:1340-1351.
166. 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.
167. 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.
168. Kustikova OS, Geiger H, Li Z, et al. Retroviral vector insertion sites associated with dominant hematopoietic clones mark "stemness" pathways. Blood. 2006.
169. Lewinski MK, Bisgrove D, Shinn P, et al. Genome-wide analysis of chromosomal features repressing human immunodeficiency virus transcription. J Virol. 2005;79:6610-6619.
170. Persaud D, Zhou Y, Siliciano JM, Siliciano RF. Latency in human immunodeficiency virus type 1 infection: no easy answers. J Virol. 2003;77:1659-1665.
171. Testa A, Lotti F, Cairns L, et al. Deletion of a negatively acting sequence in a chimeric GATA-1 enhancer-long terminal repeat greatly increases retrovirally mediated erythroid expression. J Biol Chem. 2004;279:10523-10531.
172. Schambach A, Wodrich H, Hildinger M, Bohne J, Krausslich HG, Baum C. Context dependence of different modules for posttranscriptional enhancement of gene expression from retroviral vectors. Mol Ther. 2000;2:435-445.
173. Follenzi A, Sabatino G, Lombardo A, Boccaccio C, Naldini L. Efficient gene delivery and targeted expression to hepatocytes in vivo by improved lentiviral vectors. Hum Gene Ther. 2002;13:243-260.
174. Lotti F, Menguzzato E, Rossi C, et al. Transcriptional targeting of lentiviral vectors by long terminal repeat enhancer replacement. J Virol. 2002;76:3996-4007.

175. Dull T, Zufferey R, Kelly M, et al. A third-generation lentivirus vector with a conditional packaging system. J Virol. 1998;72:8463-8471.
176. 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.
177. Schmidt M, Zickler P, Hoffmann G, et al. Polyclonal long-term repopulating stem cell clones in a primate model. Blood. 2002;100:2737-2743.
178. Kent WJ. BLAT--the BLAST-like alignment tool. Genome Res. 2002;12:656-664.
179. 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.
180. Dennis G, Jr., Sherman BT, Hosack DA, et al. DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol. 2003;4:P3.
181. http://rtcgd.ncifcrf.gov/ Mouse Retrovirus Tagged Cancer Gene Database. Accessed January 2007.
182. http://embryology.med.unsw.edu.au Embryology DNA-Tumor Suppressor and Oncogene Database, provided by the University of New South Wales (Sydney, Australia). Accessed January 2007.
183. http://www.tumor-gene.org Tumor Gene Database. Accessed January 2007.
184. 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.
185. Sandelin A, Alkema W, Engstrom P, Wasserman WW, Lenhard B. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. Nucleic Acids Res. 2004;32:D91-94.
186. Shimodaira H. An approximately unbiased test of phylogenetic tree selection. Syst Biol. 2002;51:492-508.
187. Smith LI. A tutorial on Principal Components Analysis. www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf. 2002.
188. Mahony S, Benos PV. STAMP: a web tool for exploring DNA-binding motif similarities. Nucleic Acids Res. 2007;35:W253-258.
189. Testa A, Donati G, Yan P, et al. Chromatin immunoprecipitation (ChIP) on chip experiments uncover a widespread distribution of NF-Y binding CCAAT sites outside of core promoters. J Biol Chem. 2005;280:13606-13615.
190. Harris MA, Clark J, Ireland A, et al. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res. 2004;32:D258-261.

191. Wieser R. The oncogene and developmental regulator EVI1: expression, biochemical properties, and biological functions. Gene. 2007;396:346-357.
192. Ito Y. RUNX genes in development and cancer: regulation of viral gene expression and the discovery of RUNX family genes. Adv Cancer Res. 2008;99:33-76.
193. Paschka P. Core binding factor acute myeloid leukemia. Semin Oncol. 2008;35:410-417.
194. Haylock DN, Nilsson SK. Osteopontin: a bridge between bone and blood. Br J Haematol. 2006;134:467-474.
195. Gallant S, Gilkeson G. ETS transcription factors and regulation of immunity. Arch Immunol Ther Exp (Warsz). 2006;54:149-163.
196. O'Neil J, Look AT. Mechanisms of transcription factor deregulation in lymphoid cell transformation. Oncogene. 2007;26:6838-6849.
197. Douglass TG, Driggers L, Zhang JG, et al. Macrophage colony stimulating factor: Not just for macrophages anymore! A gateway into complex biologies. Int Immunopharmacol. 2008;8:1354-1376.
198. Boehrer S, Nowak D, Hoelzer D, Mitrou PS, Chow KU. The molecular biology of TRAIL-mediated signaling and its potential therapeutic exploitation in hematopoietic malignancies. Curr Med Chem. 2006;13:2091-2100.
199. Nienhuis AW. Assays to evaluate the genotoxicity of retroviral vectors. Mol Ther. 2006;14:459-460.
200. Bushman FD. Retroviral integration and human gene therapy. J Clin Invest. 2007;117:2083-2086.
201. Porteus MH, Connelly JP, Pruett SM. A look to future directions in gene therapy research for monogenic diseases. PLoS Genet. 2006;2:e133.
202. Cattoglio C, Facchini G, Sartori D, et al. Hot spots of retroviral integration in human CD34+ hematopoietic cells. Blood. 2007;110:1770-1778.
203. Margulies M, Egholm M, Altman WE, et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature. 2005;437:376-380.
204. Wang GP, Garrigue A, Ciuffi A, et al. DNA bar coding and pyrosequencing to analyze adverse events in therapeutic gene transfer. Nucleic Acids Res. 2008;36:e49.
205. Wang Z, Zang C, Rosenfeld JA, et al. Combinatorial patterns of histone acetylations and methylations in the human genome. Nat Genet. 2008;40:897-903.
206. Barski A, Cuddapah S, Cui K, et al. High-resolution profiling of histone methylations in the human genome. Cell. 2007;129:823-837.

Acknowledgements

Many people participated in this project, giving practical and/or intellectual support.

Special thanks go to Fulvio Mavilio, Alessandra Recchia, Serenella Sartori, Giulia Facchini, and Anna Testa, all contributing more than substantially to the project design and development.

Thanks to Giuliana Ferrari's laboratory (Annarita Miccio, Antonella Antonelli, Claudia Rossi; TIGET, Milan, Italy) for their help with the lentiviral vectors, and to the Statistics Core of the Vita-Salute San Raffaele University (Alessandro Ambrosi, Clelia Di Serio; CUSSB, Milan, Italy) for statistical support.

Barbara Felice, Davide Cittaro and Lucilla Luzi (IFOM-IEO Campus, Milan, Italy) were essential to the Bioinformatics analysis of transcription factor binding sites; Francesco Ferrari (University of Padova, Italy) and Francesco Cattoglio (Politecnico di Milano, Italy) provided continuous informatics assistance.

I finally thank Prof. Andrew M. L. Lever for scientific and academic supervision.

Appendix 1.

Genes targeted by retroviral integrations. List of all Known Genes (UCSC definition) hit by RV and LV proviruses, inside or at a distance of ≤30 kb upstream or downstream, in human CD34$^+$ HSCs. Target genes of the control library of LM-PCR-amplified human CD34$^+$ HSC DNA are also listed. Gene symbols are in alphabetical order.

DATA SET	GENE SYMBOL
RV all	
Intragenic insertions	13CDNA73, AAK1, AB002324, ABCB9, ABCC1, ABCC4, ABLIM1, ABLIM1, ACACA, ACAT1, ACOT11, ADAM12, ADD3, AF035037, AF116699, AF119871, AF271775, AF336876, AFF1, AFF3, AHCYL1, AHI1, AK001094, AK054856, AK097804, AK122582, AK125044, AK125299, AK128348, AK129685, AK7, AKAP10, ALK, ALKBH3, ALOX5AP, ALS2, ANKHD1, ANKRD15, ANKRD28, ARHGAP6, ARHGDIB, ARHGEF12, ARHGEF17, ARHGEF3, ARHGEF7, ASPH, ATE1, ATP2B4, ATP5J, ATP6V0A2, ATP6V1E1, ATP8B4, ATPIIA, ATXN1, AY138860, AY240960, B3GALNT2, BBX, BC036863, BC040277, BC053534, BC062756, BC063432, BCL2, BCL2L1, BRE, BTBD11, BTN1A1, BX641037, C10orf107, C10orf26, C10orf59, C11orf49, C11orf59, C12orf31, C14orf150, C14orf153, C14orf178, C14orf43, C14orf46, C18orf1, C19orf22, C1orf164, C1orf186, C22orf9, C6orf111, C6orf198, C6orf68, C6orf89, C6orf93, C7orf10, C8orf36, C9orf3, C9orf77, CA5B, CAGE1, CALCRL, CAPG, CAPG, CAPZA1, CARS, CASP8, CBX5, CCDC26, CCM2, CD200, CD34, CD47, CD53, CD6, CDH26, CDK6, CDKL1, CEP135, CFLAR, CGI-09, CHCHD4, CHST11, CHSY1, CIP29, CITED2, CLASP1, COL24A1, CP110, CR749496, CR749644, CRADD, CRIM1, CRLF3, CRMP1, CS, CSK, CSNK1G1, CTCF, CTNNBL1, CTNND2, CTSD, CUGBP2, CXorf12, CXXC5, DACH1, DAPK2, DCBLD2, DDX59, DIP2C, DIRC2, DKFZp761D221, DLGAP1, DNAJB6, DNM3, DNMT3B, DOCK8, DSCR3, DUSP14, DYM, DYNLL1, E2F3, EAF2, EGF, ELOVL5, EMR1, ENTPD1, ERG, ETNK1, ETV6, EVI1, EVI2A, EVI2B, EVL, FAM107B, FAM20B, FARS2, FBN1, FBXL10, FBXL11, FBXL18, FBXO15, FGFR1OP2, FHIT, FHL1, FLJ10597, FLJ11336, FLJ14213, FLJ14624, FLJ20097, FLJ23577, FLJ35155, FLNB, FOLR2, FOXK1, FOXK2, FOXP1, FRAT1, FSIP1, GAB2, GARNL4, GART, GLG1, GLRA3, GMPS, GNAQ, GOLGA3, GOLPH4, GPR125, GPR128, GPR155, GPR97, GRB2, H2AFV, HCCA2, HECW2, HHEX, HIVEP3, HLA-B, HMBOX1, HMBS, HMGA2, HNRPF, HOM-TES-103, HORMAD2, HPS3, HSD17B1, IFNAR2, IL1RL1, INPP4B, IQGAP2, ITGAE, ITGB2, ITPR1, ITSN2, JMJD1C, KAZALD1, KIAA0063, KIAA0247, KIAA0350, KIAA0427, KIAA1115, KIAA1815, KIAA1840, KIF5A, KIT, KLF13, KLF7, KRT10, KSR1, LAMA3, LCP1, LDLRAD3, LDLRAD3, LGMN, LIMS1, LMO2, LNX2, LOC196264, LOC196394, LOC339789, LOC387921, LOC389634, LOC440712, LOC441108, LOC647353, LOC652848, LRP1B, LRP5, LRP8, LRRFIP1, LRRFIP2, LTBP2, M13231, MACF1, MAD1L1, MAK10, MAML2, MAML3, MAPKAPK3, MAPRE2, MARK3, MBD4, MBP,

	MDFIC, MDS006, MED18, MGAT4A, MGC12981, "MGC15523, ", MGC17624, MGC26816, MGLL, MIPEP, MKL1, MLL3, MLLT10, MORN3, MRC2, MRPL48, MRPS18B, MRPS28, MRVI1, MSI2, MTIF3, MTMR1, MTX1, MYB, MYCBP2, MYO5A, MYO5C, NANOG, NBPF1, NEK6, NEK7, NF1, NICN1, NLGN4X, NPAL2, NPR3, NR5A2, NRG1, NSE2, NUDCD1(CML66), OBFC1, OGFOD2, OPA3, OPHN1, ORC4L, OSBPL3, P2RY8, PAQR8, PCID1, PDE1A, PDE4B, PDE4D, PDGFC, PEN2, PEX1, PHF21A, PHGDHL1, PHTF1, PI4KII, PICALM, PIK3AP1, PIP5K1B, PIWIL3, PKD2, PKIG, PLCB1, PLCB4, PLCG2, PLCL2, PLCXD1, PPP1R12A, PPP3CA, PRDM10, PRDM2, PRG-3, PRKACB, PRKCB1, PRKCE, PRKCQ, PRMT8, PROM1, PSCD4, PSMA3, PTCH2, PTK2, PTPRC, QRICH1, RAB31, RAB8A, RABGAP1, RAD51L1, RAD54B, RALGPS2, RAPGEF1, raptor, RARG, RASGRP3, RASSF5, RAVER2, RBM15, RDBP, RFX2, RGS18, RHOG, RHOH, RNF130, RNF175, RNF4, RPL30, RPS4X, RPS6KA1, RRBP1, RREB1, RTN4, RTN4IP1, RUNX1, RXRA, RXRB, SAMD4, SCAMP2, SCAP1, SCFD2, SCHIP1, SEC23A, SEH1L, SENP8, SEPT2, SETBP1, SFRS5, SH2D1A, SH3BP1, SH3KBP1, SH3MD1, SH3PXD2A, SHB, SIPA1L3, SLC12A7, SLC16A10, SLC24A3, SLC25A17, SLC44A1, SLC4A4, SLC7A5, SLCO3A1, SMAD3, SMARCD2, SMG6, SMG7, SMYD3, SND1, SNX10, SOC, SPATS2, ST3GAL4, ST3GAL6, ST8SIA6, STAG1, STS-1, STX8, SUCNR1, SV2C, SYN3, SYNJ2, TA-NFKBH, TAOK3, TAPBPL, TBL1X, TCBA1, TCF7L2, TESK2, TGDS, TGFBR1, TGM5, THRB, TK2, TLN1, TLR4, TM7SF3, TMEM103, TMEM144, TMEM49, TMEM77, TMPRSS13, TNFRSF9, TNIK, TNRC6B, TPCN1, TRCB1, TRERF1, TRIM24, TRIM26, TRIO, TRIP4, TRPM3, TRPS1, TSC22D2, TTC27, TUBA8, TUBB1, TXNDC5, U96396, UBE1L, USP6NL, UTRN, VAC14, VANGL1, VAV3, VTI1A, VWF, WDR26, WDR40A, WDR51A, WRNIP1, XPO6, XRCC2, ZBTB16, ZBTB40, ZNF323, ZNF335, ZNF406, ZNF429, ZNF438, ZNF659, ZNF675, ZNFN1A2, ZNRF1
Upstream insertions (≤ 30 kb)	7A5, AB037861, AB051446, ABCC5, ACP2, ACTB, ADRB2, ADRBK1, AF090940, AF161366, AF318337, AF336887, AF466365, AK124259, AK021772, AK026318, AK123819, AK124332, AK124574, AK125490, AK126124, AK127315, AK127522, AK128368, AK128605, AK130688, AK131506, AKR1A1, AL832992, AL833872, ALDH3B2, AMICA1, AMPD2, AMT, ANXA1, APEX1, APOLD1, ARHGAP19, ARHGDIB, ARID3A, ARL11, ARL6IP4, ARMCX3, ASB13, ASB2, ASF1A, ASMTL, ASXL1, ATF4, ATP6V0A1, AY459291, BAG5, BC006122, BC027954, BC031632, BC035592, BC038104, BC043157, BC051789, BC058822, BC060886, BC093721, BCL2L13, BRP44, BTBD14B, BTEB1, BTN2A2, BTN3A2, BX538238, C10orf79, C10orf83, C11orf56, C12orf11, C12orf46, C13orf23, C14orf48, C16orf50, C17orf41, C17orf64, C20orf135, C20orf74, C21orf51, C21orf56, C21orf70, C3orf48, C4orf14, C6orf134, C6orf61, C8orf47, C9orf123, C9orf85, C9orf88, CAP1, CARD15, CAT, CBFA2T3, CBX5, CCDC48, CCKBR, CCL18, CCL23, CCM2, CD244, CD34, CD38, CD59, CD97, CDC27, CEP135, CEPT1, CLASP2, CNGA4, COASY, COL11A2, COPG, COPS7A, COPZ1, COQ3, CR936780, CREB3, CRYBB1, CSF3R, CTSZ, CWF19L1, CXCL6, CXorf9, CYB5R3, CYP11A1, CYP1B1, CYTL1, DBNDD1, DCTN2, DIP, DKFZp547D2210, DKFZP564O0523, DKFZP566M1046, DKFZp686A15192, DLD, DPH5, E2F5, EDEM1, EGFL7, ELF1, ENG, EPS8L3, ETV6, EVI2B, F25965, FAM19A2, FDPS, FIS, FLI1, FLJ12584, FLJ23322, FLJ31951, FNBP1, FRS2, FTCD, FTSJ3, FXN, GABPA, GAS8, GDF15, GGA2, GGCX, GLRX, GNA13, GNAT2, GPAM, GPR107, GPR113, GPR174, GPR18, GPR180, GPR21, GPR65, GSG2, GSN, GSTO1, GTPBP1, HCFC1, HCMOGT-1, HCST, HDLBP, HIST1H2BF, HIST1H4E, HIVEP1, HK2, HLA-C, HMG20, HNRPA1, HNRPF, HOOK2, HSD17B1, HSD17B8, HSH2D, HSPA1A, HSPA1B, HSPA4, HTATSF1, HTR3D, IFNGR1, IGFBP7, IL10RB, IL18R1, IL27RA, IMP4, INPPL1, IQCB1, IQWD1, ITGAL, ITGB7, IVD, JAK3, JUNB, JUND, KCNE1, KIAA0101, KIAA0125, KIAA0415, KIAA0746, KIAA1267, KIAA1458, KIAA1683, KISS1R, KLF6, KRT25D, LAIR1, LARP1, LASS4, LDB1, LITAF, LMAN1L, LMO2, LOC139886, LOC155100, LOC284948, LOC340156, LOC348840, LOC51255, LOC90288, LOC92154, LRRC33, LRRC37B, LY64, LYL1, MADD, MAFK, MAP2K1IP1, MAP2K5, MAP3K14, MAPKAPK3, MARCH7, MBNL1, MCM3, MCM8, MDC1, MEF2D, MGC11332, MGC23985, MGC26597, MGC30208, MGC40574, MGC9850, MGST1, MIR16, MIST, MLL3, MLX, MPP7, MRPS15, MS4A7, MYCBP2, MYL4, MYL6, MYL6B, MYO1G, MYO9A,

	N4BP1, N4BP2, NAGLU, NAP1L4, NCOR1, NFAM1, NFE2, NOSIP, NOTCH2, NPC2, NRM, NSMAF, NUMA1, NYD-SP21, OMG, OMG, P2RX5, PAQR8, PARP1, PCAF, PCBP1, PDE6H, PEX11B, PF4V1, PGM2L1, PHOX2A, PIGL, PIK3CB, PKIG, PKLR, PKN3, PNMA1, POLR1D, POLR2B, PORIMIN, PPIL1, PPP1R10, PPP4C, PPRC1, PRC1, PRKACG, PRSS23, PTPN18, QRSL1, RAB6A, RBL2, RBM8A, RCBTB1, RENBP, RFX1, RHBDL6, RHOA, RHOBTB3, RHOH, RICTOR, RING1, RLN3, RNASEH2A, RPS26, RPS6KA5, RUSC1, SAC, SCO1, SEC15L1, SEC31L1, SELI, SEPHS2, SERINC3, SET, SFRS9, SIGLEC6, SIP1, SLC35B4, SLC39A7, SLFN5, SLIC1, SMCR7L, SNAP23, SNW1, SNX27, SON, SPON2, SPTA1, SSR1, ST7L, STCH, STK32B, SUDS3, SURB7, TA-LRRP, TBN, TBX6, TCF12, TCP10L2, THAP9, THBS3, THRAP5, TM9SF2, TMEM43, TMEM99, TMSB10, TNIP1, TOB1, TPD52L2, TRIAD3, TTC19, TTYH1, TUBB, UHRF2, USP25, USP32, VAMP5, VPS13A, VPS33B, WDR10, WDR19, WRNIP1, XBP1, XYLT1, YPEL3, ZCCHC10, ZDHHC6, ZFX, ZNF175, ZNF207, ZNF306, ZNF496, ZNF594, ZNF608, ZNF740
Downstream insertions (≤ 30 kb)	7A5, AB075850, ABCF1, ACSL5, ACTR10, ADAMTSL3, ADORA2B, ADRB1, AF090940, AF116668, AF190162, AF194537, AF220263, AFG3L1, AFTPH, AGPAT2, AK025445, AK091504, AK096194, AK098012, AK1, AK124606, AK124993, AK125756, AK127109, AK127414, AK127711, AK127982, AKAP8, AKR1B10, ALDOA, ALS2CR12, ANKRD15, ANXA6, ARL4A, ARL8B, ARMCX2, ASH1L, ATP5E, ATP6V0A1, AVPI1, AVPR1B, AY151139, AY358240, AY358255, AY358798, BCAS1, BCL2, BIRC6, BLOC1S2, BRS3, BTN2A1, BTN3A3, C10orf86, C11orf42, C11orf51, C12orf38, C14orf129, C14orf172, C14orf9 , C15orf23, C17orf55, C17orf75, C19orf34, C1orf104, C20orf121, C20orf174, C20orf45, C20orf67, C21orf67, C21orf69, C21orf91, C21orf96, C3orf54, C6orf198, CAPSL, CASP10, CCDC42, CCDC73, CCR4, CCR6, CD34, CD47, CHD4, CHGB, CISH, CITED1, CLMN, CNAP1, COQ10A, COQ5, CP, CR627381, CR749643, CRYBA4, CTBP1, CTCFL, CUL5, CYGB, DDB2, DDX42, DEPDC6, DHRS9, DISP2, DKFZP434L0117, DNAJB12, DNAJB14, DPAGT1, DSCR1, EBI2, EIF2B3, EIF3S7, EML2, ENSA, EPGN, ETS2, EVA1, EVC, EVI2A, FAM62A, FAM82A, FAS, FBXO3, FCHSD2, FEZ2, FKSG44, FLJ13725, FLJ22938, FLJ33814, FLJ45244, FLJ90024, FLOT1, FRAG1, FRAT2, FYB, GAPDH, GBA2, GGA1, GLB1, GNRHR2, GOLPH3L, GPR15, GPR31, GPR4, GPR56, GTPBP6, H2AF, HBB, HBD, HBE1, HBG1, HBG2, HBLD1, HCN3, HEMK1, HIF1AN, HIST1H1E, HIST1H2AD, HIST1H2BD, HIST1H2BE, HIST1H3D, HIST1H4D, HLA-E, HMG1L1, HNRPA1, HSD17B1, HSD17B12, HSPBP1, IER3, IL23R, IL7R, IQCH, IQGAP3, IQSEC1, IRF2BP2, IRF8, ITLN1, ITPR1, JAM2, KCNE2, KIAA1600, KIAA1764, KIF3B, L12685, LOC113179, LOC162427, LOC388969, LOC442535, LOC91614, LRP2, LRRC25, LRRC51, LSM4, LSS, LY86, LYZL2, MATN2, MBD6, MGAT3, MGC33407, MGC34761, MGC34805, MGC4677, MGC4677 , MLF2, MTHFD2L, MTM1, MYOZ1, N4BP2, NAGLU, NET1, NFIX, NOL1, NOTCH2NL, NR1H3, NR5A2 , NT5M, NTN1, ODZ1, OR6K2, OR6K3, OR7C2, OSBPL6, OSGEP, P2RY6, PAP2D, PARP6, PGEA1, PGPEP1, PHACTR2, PIP5K1B, PITPNM2, PKD2L1, PMAIP1, POLG2, POU2F3, PRDX1, PRDX2, PRKRA, PRRG2, PSMC3IP, PSMC5, PSMD1, PTAFR, QARS, RBED1, RFWD3, RHBDL3, RNF40, RPS4Y1, RSBN1, RUTBC3, SCAMP4, SEC22L1, SEMA5A, SEMA7A, SEPT3, SERPINB1, SERTAD2, SFXN3, SH3BGRL3, SH3GLB1, SLC10A1, SLC14A2, SLC1A6, SLC22A5, SLC2A14, SLC9A2, SMARCC2, SPAG4L, SPP1, ST13, ST6GALNAC6, STRF7, SULT1E1, SUZ12, SYNPO2L, SYPL1, TADA2L, TAGAP , TCTA, THADA, TMBIM4, TMC5, TMEM105, TMEM136, TMEM142A, TMEM4, TNFRSF7, TOMM22, TPM4, TPT1, TRIB1, TRIM15, TRMT1, TUFT1, TYROBP, U2AF1L3, UNQ5783, USP52, USP6, VAMP1, VAMP8, VMD2L1, VPS11, WFIKKN2, XPNPEP2, ZC3H10, ZDHHC9, ZFP36L2, ZNFN1A4, ZSWIM4

LV all	
Intragenic insertions	ABCB1, ABCC1, ABI1, ABO, ACOT8, ACTN4, ADAM9, ADAMTS6, AF116714, AF118072, AK091523, AK092155, AK092861 , AK123224 , AK128476, AKAP6, ALG9, AMFR, AMICA1, ANKRD13D, ANKS1A, ANTXR2, ANXA4, AP2A1, AP2B1, AP3B1, APBB1IP, ARHGAP15, ARHGAP5, ARHGEF12, ARID4A, ARID5B, ARIH1, ARIH2, ARL6IP2, ASCC3, ASRGL1, ATAD2, ATP11C, ATP2A2, ATP8B4, ATRNL1, ATXN10, ATXN2, AZI1, BAT5, BAZ1B, BBX, BC031691 , BC063703, BC071171 , BCL7B, BICD1, BIRC3, BMP2K, BOLA2, BRE, BRF1, BRIP1, BRWD3, BTBD14B, BTBD5, BTBD7, C10orf6, C10orf84, C11orf23, C11orf49, C12orf26, C13orf7, C15orf29, C16orf45, C1orf27, C2orf25, C3orf26, C3orf62, C3orf63, C3orf9, C4orf13, C5orf31, C6orf10, C6orf106, C6orf125, C6orf142, C6orf167, C7orf10, CAD, CALCR, CALCRL, CANX, CARD8, CASP4, CBFA2T2, CBFA2T3, CBFB, CC2D1A, CCDC101, CCDC57, CD109, CD163L1, CD2AP, CD47, CDC2, CDC25C, CDC73, CDH7, CDK5RAP2, CDK6, CDKAL1, CENTB2, CEP170, CEP350, CHCHD3, CHEK1, CHKA, CHMP2B, CLCN6, CNAP1, CNOT2, CNTN1, CNTNAP2, COBLL1, COG6, COL4A3BP, COMMD10, CRADD, CSE1L, CSF1R, CXorf17, CYB5R4, DACH1, DAG1, DCAL1, DCBLD2, DEC1, DEK, DGKB, DIAPH2, DIAPH3, DKFZP586P0123, DLG1, DNAJB6, DOC1, DPP10, DPYD, DR1, DST, DTNB, E2F8, EFTUD2, EHBP1, EHHADH, EHMT1, EHMT2, EIF4ENIF1, EPB41, EPC2, EPHA3, ERG, ERO1L, ESR1, ETS1, EXDL2, EXOC4, FBN2, FBXL11, FBXL12, FBXL17, FBXL20, FBXL4, FBXW11, FCHO2, FCHSD2, FEZ2, FGD4, FLJ12716, FLJ14803, FLJ20152, FLJ20298, FLJ22028, FLJ25421, FLJ32363, FLJ32951, FLJ39370, FLJ90396, FMNL3, FOLR2, FOXP1, FXR2, FYTTD1, GAB3, GALNS, GANAB, GARNL3, GCNT2, GIT2, GK, GLCE, GLUL, GMCL1, GNAQ, GNG7, GNRHR, GOLPH3L, GPHN, GPIAP1, GRID2, GTDC1, GTF3C5, HDHD1A, HERPUD1, HERPUD2, HIVEP1, HLA-DPA1, HORMAD2, HOXA10, HSF2, HSF5, HTATIP, IBRDC1, IFT80, IGHA1 , IGHA1 , IHPK1, IL1RAPL2, INPP5E, INPP5F, INVS, IPO11, IPO7, ITGAV, ITGB1, ITM2B, ITPR2, JAKMIP2, JMJD1C, KBTBD3, KBTBD8, KCTD3, KGFLP1, KIAA0391, KIAA0528, KIAA0746, KIAA0776, KIAA0831, KIAA1128, KIAA1160, KIAA1219, KIAA1604, KIAA1715, KIAA1826, KIF11, KIF14, KLHDC4, KLHL2, KNTC2, LMBRD1, LMBRD2, LOC169355 , LOC285636, LOC317671, LOC647353, LOC652848, LPGAT1, LRBA, LRRC35, LRRC7, LUC7L, LUZP5, LYPLAL1, LYST, MADD, MAGED1, MAMDC1, MAN2A1, MAP3K3, MAP3K7IP2, MAP4K5, MAPK1, MBP, MCART6, MDS1, MECP2, MEF2A, METTL4, MGC24039, MGC46496, MGC72104, MIER2, MKLN1, MLL3, MLLT10, MLR2, MOACT5, MOV10L1, MRPL22, MRPS28, MSI2, MTAC2D1, MTBP, MTG1, MTMR3, MTPN, MTR, MUM1, MVD, MYH10, MYO10, N4BP2, NAG, NAPE-PLD, NBR1, NCAM2, NCOA7, NCOR1, NDUFA5, NEGR1, NF1, NFATC3, NFYC, NHLRC2, NHN1, NIPBL, NIPSNAP3A, NLGN1, NLK, NNT, NRXN1, NSD1, NUDCD1, NUDCD3, NUDT10, NUMB, NUP188, NUP88, NUP98, OAS2, OCRL, ODF2L, OPA1, ORC4L, OSBL1A, PACS1, PAK1, PARK2, PCDH7, PCTK2, PCTP, PDCD4, PDE11A, PDE4D, PDE7A, PDGFC, PGS1, PHF12, PHF16, PHF20L1, PHF3, PHIP, PIAS1, PIB5PA, PICALM, PIK3CD, PKN1, PKNOX1, PLCB1, PLEKHA3, PLEKHF2, POGZ, POLA1, POLE, POLR2A, PPME1, PPP3CA, PPTC7, PRDM5, PRKAA1, PRKAG1, PRKAR2B, PRKRA, PRNPIP, PRPF6, PRPF8, PSCD1, PSMD13, PSPC1, PTK2B, PTPRA, PTPRD, PTPRK, PUS7L, QKI, RAB11FIP3, RAB3GAP1, RAB6IP1, RAD12, RAD18, RAD51L1, RALA, RALGPS2, RASA2, RBPMS, RC3H1, RCN2, RDBP, RFWD2, RFX3, RHOT1, RNASEL, RNF130, RNF24, RNMT, RNPC3, RNPS1, ROBO1, ROCK1, RPRC1, RRM2B, RSNL2, RUNDC1, RUNX1, RUNX2, SAE1, SAPS3, SCAMP2, SEC11L1, SEC14L1, SEL1L, SEMA3A, SENP6, SESN1, SETBP1, SETD2, SF3B2, SFRS15, SFXN1, SGK3, SH2D4B, SH3GL1, SH3RF1, SHARPIN, SIL1, SIN3A, SLC2A5, SLC30A7, SLC3A2, SLCO4C1, SLCO5A1, SLFN11, SLK, SMAD2, SMAP1L, SMARCA1, SMARCC1, SMCX, SMG1, SMYD3, SNAPC3, SOCS7, SOS2, SP3, SPATA16, SPATA5, SPATA9, SPCS3, SPPL3, SRCAP, SRPK2, SRR, SSH2, ST7, ST7L, ST8SIA4, STAG1, STAG2, STAT5B,

	STK17A, STK3, STOM, STS-1, STXBP4, STXBP5, SUFU, SULT1E1, SUPT3H, SUSD1, SUV39H2, SWAP70, SYNE1, TACC1, TBC1D23, TBC1D5, TCBA1, TCF4, TEC, TEDDM1, TFR2, THADA, THRAP1, TMEFF1, TMEM2, TMEM49, TMEM55A, TNIK, TNKS, TNRC6C, TOP2B, TOX, TPCN1, TPP2, TRAF2, TRIM38, TRIM44, TRIO, TRPM7, TRPS1, TSR1, TXNDC10, UBA2, UBE2E2, UBE2G1, UBR2, UCHL3, UHRF2, USP12, USP25, USP33, USP34, USP53, UTP14A, UTP20, UTRN, VPS13B, VPS13D, VPS26A, VPS41, VPS54, VRK3, WDR17, WDR35, WDR7, WDTC1, XRCC1, XRCC4, YEATS2, YTDHF1, ZBTB20, ZBTB8OS, ZC3H5, ZC3H7A, ZCCHC7, ZCCHC8, ZFAND3, ZFP64, ZNF276, ZNF45, ZNF559, ZNF583, ZNF6, ZNF644, ZNF766, ZNFN1A2, ZZEF1
Upstream insertions (≤ 30 kb)	AARSD1, AASDHPPT, AB082528, AF118080, AF130057, AIF1, AK074565, AK090761, AK092305, AK094323, AK095276, AK096194 , AK126169, AK126463, AK126488, AK126785, AK127468, AK128353, AK129559, AK129966, AK130228, AK130247, AK130865, AMY2B, ANAPC11, ANKFY1, APRT, ARHGDIA, ARL6IP, ATP5I, AURKA, AY203928, AY484516 , B3GAT3, BAT2, BAT4, BC029660, BC030956 , BC038573, BC090057, BTBD12, C11orf68, C15orf27, C16orf7, C16orf70, C17orf38, C19orf57, C1orf120, C20orf32, C3orf1, C3orf18, C3orf54, C6orf21, C6orf47, C9orf163, C9orf38, CCDC103, CCHCR1, CEL, CELL, CETP, CHAF1A, CHD3, CLEC2B, COMMD6, CSRP3, CST3, CTDSP2, CUEDC2, CYBA, DCI, DDN, DDX56, DFNB59, DIP, DKFZP586P0123, DMTF1, ECGF1, EPC1, FAM53C, FBXL13, FBXO4, FHL5, FKBP7, FLJ21657, FLJ35834, FLJ36840, FLJ38482, FLJ45530, FUZ, GAL3ST3, GALK1, GALNS, GAPDH, GBA2, GEMIN5, GM632, GPR179, H3F3B, HCG27, HDAC9, HIST1H2AB, HIST1H2AK, HIST1H2AL, HIST1H3B, HIST1H3C, HIST1H4B, HIST1H4K, HLA-DPB1, HOXA5, HOXA6, HOXA7, HOXA9, IER2, IFI35, IL10RA, INPP4B, INPP5E, INPPL1, IQCD, IRAK4, ITFG3, JUND, KIAA0423, KIAA0963, KIAA1683, KIAA1875, KLHDC7B, KLHL13, KNTC2, LOC124446, LOC283331, LOC440836, LRCH3, LSMD1, LTB, LY6G5C, LY6G6D, LYL1, MAF1, MAML1, MAN1A1, MAPKAPK3, ME1, MFSD7, MLC1, MPI, MRPL16, MRPL51, MTHFR, NAT10, NCR3, NDST4, NDUFS7, NIPSNAP3B, NPB, NT5C3L, NUDT6, NUP50, OR8G1, OSBPL8, P2RY1, P8, PACS1, PACS2, PCGF3, PDE4C, PH-4, PIGL, PIN1, POU4F1, POU5F1, PPME1, PSD, PTOV1, PYGM, RASL10B, RBBP4, RC3H1, REEP3, RGS1, RILP, RNF185, RPAIN, RPL27, RSHL2, RUTBC1, RWDD4A, RY1, SAMD10, SCO2, SDCCAG3, SELM, SEMA4G, SERPINB1, SF1, SFI1, SHPRH, SIRT3, SLC25A17, SLC44A4, SLC5A6, SMG6, SNRPB, SOX15, SSH3, STAT5A, STK11, TAOK2, TM2D2, TMEM15, TMEM180, TMEM39A, TMEM64, TMSL3, TNNC2, TPR, TRAPPC3, TRIM35, TRMT1, TSGA10IP, TSKS, TTC30A, UBE2J1, UBL5, UCKL1, WDR13, WDR74, XCR1, ZBTB4, ZFP2, ZNF114, ZNF221, ZNF320, ZNF473, ZNF480, ZNF490, ZNF576, ZNF582, ZNF9, ZSWIM3
Downstream insertions (≤ 30 kb)	A2M, ABCA1, ABCA3, ACLY, ACTL6B, ADCY1, ADRBK1, AF130061, AF132200, AK074565, AK097463, AK098728, AK125829, AK128497, AK128554, AKT1S1, ANXA4, AOF1, ARFGEF2, ARMC2, ARRDC2, BAT3, BAT5, BC009783 , BC033532, BTG1, C10orf95, C11orf48, C12orf33, C14orf100, C17orf56, C1QBP, C2orf28, C2orf4, CCDC36, CCDC44, CCDC66, CD68, CD69, CDT1, CISH, CLEC2A, CPLX3, CPT1B, CREB3, CRIM1, CSNK2B, CSTF1, CYB5D1, CYC1, DCLRE1C, DDX59, DNASE1L2, DPH5, DRAP1, DUSP6, EIF3S12, EIF4A1, ERG, EXOSC4, FANCA, FBXL15, FGF2, FKBP10, FKBP7, FKSG44, FLJ11021, FLJ22688, FOLR1, GIF, GIPC1, GNGT1, GPAA1, HEMK1, HIF1AN, HIGD1B, HIST1H1B, HIST1H1C, HIST1H2BB, HIST1H2BN, HIST1H3A, HIST1H3I, HIST1H4A, HIST1H4L, HLA-E, HOM-TES-103, HORMAD1, HOXA11, HOXA13, HSPC176, HTRA4, IHPK3, IL12RB1, IL17C, INTS5, IRGQ, ITPR3, KCNN1, KIAA0258, KLHL10, KLHL11, LOC144233, LOC196264, LOC196541, LOC389118, LOC401252, LSM4, LST1, LY6G5B, LY6G6C, LY6G6E, MAP4K2, MED25, MEN1, MGC15523, MGC26885, MGC35402, MGC39372, MLL2, MPDU1, MYL5, NCAPH2, NEU1, NHN1, NOTCH1, NPPA, NUDT21, OCLM, OR8G2, OTUD6B, PAIP1, PAOX, PCYT2, PDE6B, PDSS1, PHIP, PHOX2A, PHYHD1, PIK3R5, PLA2G3, PMPCA, PNKP, PODNL1, POF1B, POU1F1, PPP1R2P9, PPP3CC, PSIP1,

	PTGER1, PTPN9, Q6ZSU9, RABEP1, RAD9B, RARB, RBM3, RECQL, RFESD, RGS16, RGSL1, RIC8A, RNASEH2C, ROM1, RPS6KA5, SEC23IP, SEPP1, SERTAD4, SEZ6, SH2B3, SIRT7, SLC12A3, SLC30A3, SMTN, SNAI3, SNTB1, SNX21, SPAG6, SQFE253, SRR, ST13, STGC3, STT3A, STX10, STX3, SVH, TAGAP, TBL2, TBPL2, TCF19, TCHP, TCN1, THEG, THOC4, TMED4, TMEM88, TNRC6C, TOMM7, TSR1, TWISTNB, UAP1, UBE2C, UGP2, USH2A, USP4, VAT1, ZBTB12, ZNF326, ZNF343, ZNF354B, ZNF667, ZNF718, ZNFN1A2
Controls Intragenic sequences	ABCA1, ACTL6B, ADAMTS3, AIG1, AK127078, AK5, ALS2CR19, ANK3, APBB2, ARHGAP15, ARL2BP, ARL6IP2, ARL8B, ARNTL2, ASTN2, ATE1, AUH, BACH2, BBS7, BCAS3, BCL2, BRIP1, BRWD1, BTBD9, BZW2, C10orf30, C12orf50, C14orf37, C1orf101, C1orf139, C1orf26, C4orf13, C4orf16, C4orf8, C9orf39, CACHD1, CAMK1D, CBFA2T2, CCBE1, CCDC26, CCDC46, CCT4, CD2AP, CDH13, CDH18, CDH2, CDH8, CDKN2C, ChGn, CHN2, CIT, CLOCK, CNTN5, CNTNAP4, COG2, CPE, CPNE4, CREB5, CRSP2, CSMD1, CSS3, CTNNAL1, CYP4Z1, DAB1, DDEF1, DDX17, DEPDC2, DIAPH2, DKFZp667M2411, DLC1, DLG2, DNAJC1, DNAJC18, DNER, DP58, DPP10, DSCAML1, DSG1, DYNLT1, EIF4G3, ELK3, EMID2, EPB41L4B, EPHA3, EPS15, ERBB4, ERGIC1, ESR1, EVA1, EXOC4, FAM19A, FAM19A1, FARP2, FARS2, FBXL17, FER, FLJ13576, FLJ22104, FLJ22624, FLJ30294, FLJ36166, FLJ39739, FSIP1, FYN, GAB1, GCC2, GFRA1, GLS, GPD2, GRIN2B, GSK3B, HCTSL-s, HECW2, HERC4, HHIP, HNRPR, IGHA1, IGSF4D, IL10RB, IL18, IL1RAPL1, JAZF1, JMJD2C, KCNQ5, KCTD2, KCTD3, KHDRBS2, KIAA0133, KIAA0555, KIAA0564, KIAA0828, KIAA1344, KIAA1432, KIAA1463, KIAA1900, KIAA1958, KIF13B, L10374, LGMN, LIN7A, LNX1, LOC138046, LOC153561, LOC400986, LOC51057, LOC51334, LPHN3, LY75, LYAR, MACF1, MAGI2, MAMDC1, MAML3, MAP2, MAPKAP1, MCCC2, MCTS1, MLR2, MOBKL1A, MPP4, MRPL44, MTA3, MTRF1L, MYH10, MYO3B, NAALADL2, NAV2, NBEAL1, NCBP2, NEK1, NELL1, NF2, NFAT5, NLGN1, NOMO1, NOTCH2NL, NRG1, NRG3, OSBPL6, OSBPL8, OXR1, PACRG, PAK3, PALLD, PARP8, PASK, PB1, PCDH15, PCNXL2, PFTK1, PHACTR1, PIK4CA, PLCE1, PLEKHA5, POU6F2, PPA2, PPP1R10, PPP2R2B, PPP3CA, PRKCB1, PTK7, PTPN23, R3HDM1, RAP80, RAPGEF1, RASA1, RASAL2, RBBP5, RBL2, RBMS3, RELN, RERE, RGL1, RGS6, RHAG, RIC8B, RIMS2, RIT1, RORA, RSPO4, RUVBL1, SCN1A, SCPEP1, SEC23A, SEMA3D, SEPT10, SGCZ, SH3BGRL2, SIPA1L2, SLC1A2, SLC25A17, SLC2A13, SLC40A1, SLC44A5, SLC5A8, SMA3, SMAD2, SNAP91, SNTB2, SNTG1, SOX5, SOX6, SP6, SPAG16, SPAST, SRD5A2L, SRPK1, SSBP2, STAU2, STK32B, STN2, STXBP5, STXBP6, SUPT3H, SURB7, SYNPR, TACC1, TBL1XR1, TBX15, TBX4, TCERG1L, TCF4, TCF7L1, THADA, THRB, TMEM16D, TMEM16F, TNIK, TNNI3K, TNPO1, TPP2, TRIM9, UBE4B, UGT2B4, USP34, VAPB, VGCNL1, VTI1A, WAPAL, WDR7, WWOX, XRCC5, YAP1, ZADH1, ZCCHC16, ZDHHC21, ZFYVE28, ZMYM6, ZNF273, ZNF277, ZNF714, ZSWIM5
Upstream sequences (≤ 30 kb)	ABCC8, ADAMTS17, AK025522, AK093210, AK125488, AK126028, AK126853, AK127866, AKR1C1, AL832683, ANKMY2, ARNT2, ASB4, ATP5H, BC044620, BC050563, C12orf29, C12orf58, C13orf23, C18orf14, C18orf17, C2orf26, C6orf134, CAMKK1, CAMKV, CCNA2, CCNI, CD1D, CD300LF, CDH11, CEP170, CTDSP2, CXorf38, DAB2, DNAJB13, DOCK2, EIF3S9, ENAH, FAF1, FBXO5, FLJ42289, GAS2, GNRHR, H2AFY, HSA277841, HTLF, IFRG28, IKBKAP, ITGB6, KCTD14, KERA, KIAA1026, KRTAP21-1, KRTAP21-2, KRTHB1, KRTHB6, LAMB1, LOC154907, LOC221442, LOC388730, LOC442247, LRRN5, LUZP4, LZTR2, MDM2, MEF2A, MGAT4C, MGC13159, MICAL2, MITF, MRPS18B, MRPS28, MTMR2, NDUFV2, NFIA, NIT2, OR6C1, OR6C3, OSBPL6, PACAP, PEMT, PHC2, PHF6, PIK3CA, PMPCB, PNOC, POTE15, PPP1R7, PRKAB2, PTD004, PYGB, Q6ZUH1, RAB37, RANBP2, RBMS3, RSL1D1, SDCCAG8, SERPIND1, SFMBT2, SLC30A6,

	SLC9A3R1, SMCP, SNX8, SPRY2, SYTL3, TAF5L, THRB, TPD52, TRAIP, TSPAN13, USP16, WDFY1, ZNF155, ZNF16, ZNF2, ZNF230, ZNF294, ZNF410, ZNF509, ZNF514, ZSWIM5
Downstream sequences (≤ 30 kb)	ABCF1, ABRA, AGT, AK095410, AK096196, AK124028, ALS2, ARID5B, ARRDC3, BLNK, BNC2, C11orf69, C15orf41, C1GALT1C1, C1orf171, C21orf6, C9orf6, CD59, CDC2, CHD2, CHKB, CNTN2, COCH, CPT1B, ELP3, ENTPD6, EPM2A, EXOSC9, FAM50B, FLJ20160, FLJ20364, FLJ32786, FLJ90013, FMO5, GCKR, GLDN, GNAI1, GRM6, GSPT1, H1FNT, HERV-FRD, ICT1, IFNAR2, KIAA0907, KIAA1008, KIAA1423, KLC4, KLHDC7B, KRT8, KRTHB3, L10374, LAMA3, LAMB4, LCE1A, LIMS1, LOC387921, LOC441931, LOC92691, MGC70870, MKRN3, MON1A, MPPED2, MRPL22, MST1R, NFYA, NPAL1, NRXN1, NUP155, PIGZ, RAP1GDS1 , RGMA, RGS17, RNASE11, RSPRY1, RSU1, SEC61A1, SENP5, SERPINE2, SLC26A8, SLC35B4, SNCA, SP100, SPZ1, STK32A, STRN3, SYT11, TBC1D23, TFAP2A, TG, TIFA, TM4SF11, TSPAN5, TWISTNB, TXK, UGT8, UNQ9217, USH1C, WDR71, ZNF221, ZNF454, ZNF518, ZNF641

Appendix 2.

Ingenuity network analysis of genes targeted by RV and LV integrations, and control sequences. Networks with the best Ingenuity score and made exclusively of genes targeted by RV or LV integrations are shown. A single network was obtained with the control gene list.

	Network ID	Genes	Score	Focus genes	Top functions
RV	1	AHCYL1, ALK, ALOX5AP, ATXN1, BCL2, BCL2L1, BRE, CASP8, CASP10, CFLAR, CHGB, CRADD, DDB2, DYNLL1, FAS, ITPR1, MAD1L1, MADD, MKL1, MLX, MYB, NOSIP, PARP1, PMAIP1, PRKCQ, RASGRP3, RTN4, RTN4IP1, SLC1A6, SNW1, TNFRSF7, TPT1, TUBA8, TUBB, TUBB1	42	35	Apoptosis ($p<10^{-7}$)
	2	ACTB, ADRB1, ANXA1, CBFA2T3, CSF3R, DACH1, EMR1, ERG, ETV6, FLI1, FRS2, FYB, GAB2, GRB2, GSN, HBB (includes EG:3043), HNRPA1, HSPA1B, INPPL1, IQGAP2, IRF8, KLF7, MIST, NCOR1, PLCG2, RAPGEF1, RHOG, SCAP1, SHB, SMARCC2, SMARCD2, TBL1X, TNIK, VAV3, VWF	42	35	Cell adhesion ($p<10^{-10}$) Transcription ($p<10^{-4}$) Cell Differentiation ($p<10^{-4}$)
	3	ANXA6, ARHGDIB, ARHGEF3, ARHGEF12, ARHGEF17, BRS3, CD53, GABPA, GNA13, GNAQ, HBD, HBE1, HBG1, HBG2, HMBS, HSPA1A, ITGAE, ITGB2, ITGB7, LAMA3, MAFK, NFE2, OMG, PIP5K1B, PPP1R12A, PRKCB1, PTK2, RGS18, RHOA, RHOH, SFRS5, TLN1, TRIO, UTRN, ZNF335	42	35	Stress fiber formation ($p<10^{-8}$) Transformation ($p<10^{-3}$)
	4	ATP2B4, CCL18, CDC27, CTBP1, CXXC5, E2F5, ENG, EVI1, FOLR2, FOXP1, FRAT1, HHEX, JUND, KLF6, KLF13, MAP3K14, NR1H3, PCAF, PPP3CA, RARG, RFX1, RFX2, RPL30, RREB1, RUNX1, RXRB, SMAD3, SND1, SPP1, SURB7, TADA2L, TCF7L2, TGFBR1, TNIP1, TOB1	42	35	Transcription ($p<10^{-13}$) Cell Proliferation ($p<10^{-4}$) Tumorigenesis ($p<10^{-3}$)

	5	ATF4, CBX5, CD34, CD38, CSK, CXCL6, CYP1B1, ETS2, FDPS, GLG1, HK2, IER3, IL1RL1, JMJD1C, JUNB, MBP, NR5A2, NRG1, PIK3CB, PKLR, PRKACB, PRKACG, PSMC5, PSMC3IP, PSMD1, PTPN18, PTPRC, RXRA, SLC10A1, SMYD3, THRB, TLR4, TRIM24, TRIP4, ZFP36L2	42	35	Transcription ($p<10^{-4}$) Cell proliferation ($p<10^{-3}$) Cell differentiation ($p<10^{-3}$)
LV	1	ABI1, AKAP6, BMP2K, BRF1, CAD, CBFB, CSF1R, DUSP6, FOXP1, GAB3, GIPC1, GIT2, GNRHR, GTF3C5, JUND, MADD, MAGED1, MAPK1, MAPKAPK3, MBP, PAK1, PDE4C, PDE4D, PTPRA, PTPRD, PYGM, RGS16, RPS6KA5, RUNX1, SH2B3, SH3GL1, SNAPC3, SOS2, THOC4, TRPM7	46	35	Cell growth ($p<10^{-3}$) Signaling pathway ($p<10^{-3}$)
	2	ABCB1, ACTL6B, ARID4A, BAZ1B, CDC25C, CDK6, CHAF1A, CHD3, CHEK1, CSNK2B, DMTF1, ETS1, FANCA, FGF2, GLUL, HOXA7, HSF2, KIF11, MAML1, MECP2, MEN1, MLL2, NOTCH1, PDGFC, PHOX2A, PIN1, POU1F1, PPP3CA, RBBP4, RUNX2, SIN3A, SMARCA1, SMARCC1, SP3, ZNFN1A2	46	35	Transcription ($p<10^{-8}$) Cell proliferation ($p<10^{-7}$) Apoptosis ($p<10^{-4}$)
	3	ADRBK1, ARHGEF12, ATP2A2, CALCR, CBFA2T2, CBFA2T3, DACH1, ESR1, EXOC4, GNAQ, GNGT1, IER2, ITPR3, KBTBD8, LCOR, LTB, MAP3K3, MAP3K7IP2, MEF2A, MOV10L1, MTPN, NCOA7, NCOR1, NFYC, NPPA, P2RY1, PDE7A, PIAS1, PKN1, PLCB1, POLR2A, RALA, SUFU, TEC, THRAP1	46	35	Calcium signaling ($p<10^{-5}$) Transcription ($p<10^{-3}$)
	4	A2M, ABCA1, ADAM9, AP2A1, AP2B1, ARHGAP5, BIRC3, CANX, CD47, CISH, DAG1, DLG1, DTNB, EPB41, FOLR2, HOXA9, HOXA11, ITGAV, ITGB1, MAP4K2, MAP4K5, NLGN1, NRXN1, NUMB, PICALM, PTK2B, RARB, SLC3A2, SNTB1, ST8SIA4, STAT5A, STAT5B, TNIK, TRAF2, UTRN	46	35	Cell adhesion ($p<10^{-5}$)
Control	1	BCL2, BLNK, CCNA2, CDC2, CRSP2, DAB1, DAB2, DDX17, EPS15, ERBB4, ESR1, GAB1, GNRHR, GSK3B, LCOR, MDM2, MST1R, MTA3, NF2, NRG1, NRG3, PIK3CA, RANBP2, RAPGEF1, RBL2, RELN, SERPINE2, SLC9A3R1, SMAD2, SNAP91, SRPK1, STON2, SURB7, WWOX, YAP1	56	35	Cancer ($p<10^{-8}$) Apoptosis ($p<10^{-6}$) Transcription ($p<10^{-5}$)

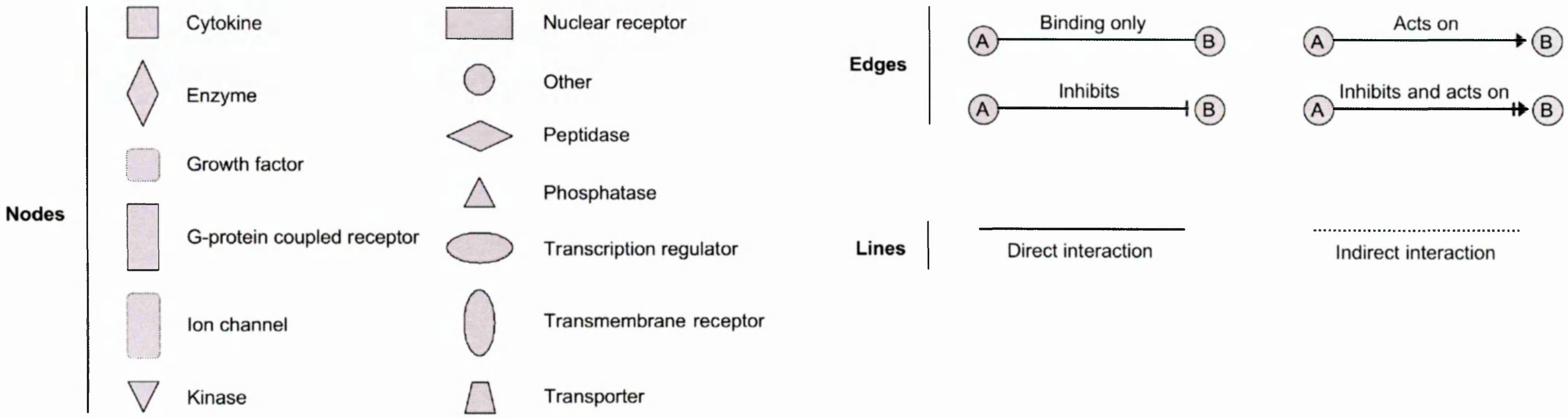
Legend of Ingenuity network analysis symbols. The legend provides a key of the main features of Ingenuity network explorer, including node shapes, edge labels and types. The color gray identifies the Focus Genes from a data set (*i.e.*, the genes targeted by RV and/or LV integrations).

"Acts on" and "Inhibits" edges may also include a binding event.

Direct interactions require that two nodes (*i.e.*, genes or gene products) make direct physical contact with each other, with no intermediate steps.

Direct interactions can also include chemical modifications (*e.g.*, phosphorylation), provided that there is evidence that the two factors involved interact directly rather than through an intermediate.

Indirect interactions do not require that there is physical contact between the two nodes.



Appendix 3.

Complete statistics of Figure 20. Comparison of the frequency of consecutive insertion sites having a certain distance one from each other, computed for 8 distance bins (1-10 bp; 10-100 bp; 100-1,000 bp; 1,000-10,000 bp; 10,000-100,000 bp; 100,000-1,000,000 bp; 10,000,000-100,000,000 bp); all possible combinations of sample diversities within each bin were assessed by a 2-sample test for equality of proportions with continuity correction.

The sign '-' is used when the statistical test was not reliable, *i.e.*, numbers were too low for the *chi*-squared approximation to be considered valid.

	Controls	**RV**	**LV**	**HeLa**	**Controls v.s. RV**	**Controls v.s. LV**	**RV v.s. LV**
bin 1	0	0	0	0	-	-	-
bin 2	3	2	0	4	-	-	-
bin 3	1	21	5	12	0.0005	-	0.0129
bin 4	2	50	15	24	1.07e-08	5.35e-03	4.15e-04
bin 5	28	93	54	55	4.29e-06	1.18e-02	0.03783
bin 6	168	287	229	220	0.001	0.007	0.6767
bin 7	521	492	473	479	1.85e-14	2.97e-05	7.57e-04
bin 8	51	58	53	52	0.5466	0.954	0.655

Modulo function. The distance between two consecutive insertion sites is a mono-dimensional value, to be plotted along a single axis. However, when hundreds of values have to be plotted together, it becomes hard to visualize them graphically. Hence I decided to apply a function to my data so that they would be arbitrarily scattered along a second, *y*, axis. I used a slightly modified *modulo* operation, which finds the remainder of division of one number by another. Given two numbers, *a* (the dividend) and *n* (the divisor), a modulo *n* (abbreviated as *a mod n*) is the remainder, on division of *a* by *n*. For instance, the expression "7 *mod* 3" would evaluate to 1, while "9 *mod* 3" would evaluate to 0. Let *x* be the distance between two consecutive insertion sites, in base pairs. To associate a y value to each x value, I applied the following function:

$y\ (x) = (x\ mod\ 100)\ /\ 100$

In this way x values were scattered along the y axis on 99 virtual rows, assuming values ranging from 0.00 to 1.00, practically corresponding to the last 2 digits of the bp distance. For example:

$y\ (13{,}367) = (13{,}367\ mod\ 100)\ /\ 100 = 0.67$

Appendix 4.

Integration hot spots. Complete list of hot spots of RV integrations in CD34$^+$ and HeLa cells, of LV integrations in CD34$^+$ cells, and of "false-positive" hot spots (controls) generated by applying the definition criteria to the control list of sequences. Range indicates the maximum distance between hits contained in each hot spot.

	Chromosome	Position	Range (bp)	N° hits	Gene symbol	GeneID	Origin*
CD34$^+$ RV hot spots (n=97)	1p34.1	44550505 44556772 44558256 44559928	9423	4	FLJ10597	55182	CB-RV (4)
	14q24.3	73278789 73289432 73291968 73292671	13882	4	C14orf43 PNMA1	91748 9240	CB-RV (2) CB-ΔRV (2)
	17q23.2	55201823 55216872 55217426 55219527	17704	4	TMEM49	81671	CB-RV (4)
	6p25.2	2697508 2714551 2764400 2764413	66905	4	LOC34015 SERPINB1 WRNIP1	340156 1992 56897	BM-ADA (2) CB-ΔRV (1) PB-ND (1)

17q21.2	37863134	103951	4	ATP6V0A1	535	PB-ND (2)
	37936799			COASY	80347	BM-ADA (1)
	37960561			HSD17B1	3292	CB-RV (1)
	37967085			LOC16242	162427	
				NAGLU	4669	
				MLX	6945	
				PSMC3IP	29893	
10q21.2	63178757	2767	3	C10orf107	219621	BM-ADA (2)
	63179101					CB-ΔRV (1)
	63181524					
10q24.33	105508395	4797	3	SH3MD1	9644	BM-ADA (1)
	105512424					CB-RV (1)
	105513192					CB-ΔRV (1)
17q11.2	26661186	7827	3	EVI2A	2123	BM-X-SCID (1)
	26662552			EVI2B	2124	CB-RV (2)
	26669013			NF1	4763	
				OMG	4974	
1q31.2	188861554	8221	3	RGS18	64407	BM-ADA (1)
	188862119					BM-X-SCID (1)
	188869775					CB-ΔRV (1)
17q21.32	42631527	8327	3	CDC27	996	CB-RV (1)
	42636597			MYL4	4635	PB-ND (2)
	42639854					
1q32.1	202792256	13006	3	AVPR1B	553	BM-ADA (1)
	202801803			C1orf186	440712	CB-RV (1)
	202805262					CB-ΔRV (1)
18p11.21	13552526	18214	3	AF090940		CB-RV (1)
	13566992			C18orf1	753	CB-ΔRV (2)
	13570740					
1q32.1	196855882	28635	3	NR5A2	2494	CB-ΔRV (3)
	196856552					
	196884517					
2q13	111847549	36605	3	AK123819		CB-RV (1)
	111881069			MGC4677	112597	CB-ΔRV (2)

2q13	111847549	36605	3	AK123819		CB-RV (1)
	111881069			MGC4677	112597	CB-ΔRV (2)
	111884154					
1q32.2	204437422	37264	3	CD34	947	CB-ΔRV (1)
	204471786					CB-RV (1)
	204474686					PB-ND (1)
10p14	11211805	37510	3	CUGBP2	10659	BM-ADA (1)
	11226458					CB-RV (1)
	11249315					CB-ΔRV (1)
12q24.22	115534910	38624	3	intergenic		BM-ADA (1)
	115535352					CB-RV (1)
	115573534					CB-ΔRV (1)
20q13.2	51757705	41605	3	intergenic		PB-ND (1)
	51789775					CB-RV (1)
	51799310					CB-ΔRV (1)
11p13	33860999	48661	3	AF116668		BM-ADA (1)
	33889289			LMO2	4005	CB-RV (1)
	33909660					CB-ΔRV (1)
9p24.3	683102	55917	3	ANKRD15	23189	CB-RV (3)
	696469					
	739019					
5q32	148164841	79	2	ADRB2	154	BM-ADA (2)
	148164920					
12q24.31	120564071	104	2	MORN3	84876	BM-X-SCID (1)
	120564175			TMEM142A	283385	CB-RV (1)
20p12.3	8082923	136	2	PLCB1	23236	CB-ΔRV (2)
	8083059					
19p13.13	13075943	137	2	BTBD14B	112939	BM-X-SCID (1)
	13076080			LYL1	4066	CB-RV (1)
				NFIX	4784	
				TRMT1	55621	
2p16.1	60504592	274	2	intergenic		BM-X-SCID (1)
	60504866					CB-RV (1)
6q21	109731942	293	2	intergenic		BM-X-SCID (1)

21q11.2	14970504 14970863	359	2	intergenic		BM-ADA (1) CB-ΔRV (1)
11p15.4	6213567 6214039	472	2	AK026318 BC040277 C11orf42 C11orf56 CCKBR CNGA4 DKFZP566M	 160298 84067 887 1262 84067	CB-RV (1) PB-ND (1)
Xp22.11	23924605 23925096	491	2	ZFX	7543	CB-RV (1) CB-ΔRV (1)
15q26.3	99596986 99597497	511	2	CHSY1	22856	CB-RV (2)
20p11.23	20664971 20665584	613	2	C20orf74	57186	BM-X-SCID (1) CB-RV (1)
17q21.31	41627675 41628452	777	2	KIAA1267	284058	BM-ADA (1) CB-RV (1)
2p11.2	85544636 85545415	779	2	CAPG LOC284948 RBED1	822 284948 84173	PB-ND (2)
8q24.12	121146776 121147669	893	2	DEPDC6	64798	PB-ND (1) CB-RV (1)
2p21	43360006 43360981	975	2	AK025445 MGC40574 THADA ZFP36L2	 285048 63892 678	CB-RV (2)
13q32.3	98948370 98949514	1144	2	TM9SF2	9375	CB-ΔRV (1) PB-ND (1)
3p13	71522689 71523950	1261	2	FOXP1	27086	PB-ND (1) CB-RV (1)
Xp22.12	19643071 19644418	1347	2	SH3KBP1	30011	CB-RV (2)

5p13.1	39110796	1373	2	FYB	2533	CB-RV (2)
	39112169			RICTOR	253260	
22q13.2	41164222	1498	2	NFAM1	150372	CB-RV (1)
	41165720					CB-ΔRV (1)
11p13	36388453	1715	2	FLJ14213	79899	BM-X-SCID (1)
	36390168					CB-RV (1)
9p13.2	38029205	1811	2	SHB	6461	BM-ADA (2)
	38031016					
10q25.2	111823809	1920	2	ADD3	120	BM-ADA (1)
	111825729					CB-ΔRV (1)
10p12.31	22580508	1977	2	intergenic		CB-RV (2)
	22582485					
6p24.3	7090865	1991	2	RREB1	6239	CB-RV (1)
	7092856					CB-ΔRV (1)
4p16.1	10381173	2068	2	MIST	116449	CB-ΔRV (1)
	10383241					PB-ND (1)
15q24.2	73207330	2251	2	intergenic		CB-RV (1)
	73209581					CB-ΔRV (1)
18q21.33	58892975	2747	2	intergenic		CB-RV (1)
	58895722					CB-ΔRV (1)
4q13.3	75559394	2810	2	EPGN	255324	CB-RV (1)
	75562204			MTHFD2L	441024	CB-ΔRV (1)
17q25.3	76986470	2915	2	intergenic		CB-RV (1)
	76989385					PB-ND (1)
20q11.21	30825230	3355	2	DNMT3B	1789	CB-RV (2)
	30828585					
1p13.3	108174011	3673	2	VAV3	10451	CB-RV (2)
	108177684					
4p16.2	5153853	4225	2	CYTL1	54360	CB-ΔRV (1)
	5158078			STK32B	55351	PB-ND (1)
10q22.1	73745534	5129	2	DNAJB12	54788	CB-ΔRV (2)
	73750663					
11p13	36068419	5677	2	LDLRAD3	143458	BM-X-SCID (1)
	36074096					CB-RV (1)

1p13.3	111128305 111134079	5774	2	CD53	963	BM-X-SCID (1) CB-ΔRV (1)
3p14.3	56784711 56791063	6352	2	ARHGEF3	50650	BM-X-SCID (1) CB-RV (1)
12p13.2	11825451 11832811	7360	2	ETV6	2120	CB-RV (2)
12p12.3	15000262 15008965	8703	2	ARHGDIB C12orf46 PDE6H	397 121506 5149	CB-RV (1) CB-ΔRV (1)
18q23	72897211 72905976	8765	2	MBP	4155	CB-RV (1) CB-ΔRV (1)
13q13.1	31580431 31589216	8785	2	13CDNA73	10129	CB-ΔRV (2)
22q12.1	27521637 27530857	9220	2	FLJ33814 XBP1	150275 7494	BM-ADA (1) CB-RV (1)
6q23.3	135547695 135557117	9422	2	MYB	4602	CB-RV (2)
3q13.12	109291185 109301095	9910	2	CD47	961	CB-RV (2)
22q13.31	45309539 45320437	10898	2	DIP	23151	CB-RV (1) PB-ND (1)
10q11.21	43223980 43235621	11641	2	HNRPF	3185	PB-ND (2)
4p14	40009255 40021254	11999	2	N4BP2 RHOH	55728 399	CB-RV (2)
20q13.2	51972400 51985519	13119	2	BCAS1 CR749643	8537	CB-ΔRV (2)
11q24.3	128038248 128052395	14147	2	FLI1	2313	BM-ADA (1) CB-RV (1)
20p12.2	9074545 9090002	15457	2	PLCB4	5332	CB-ΔRV (1) CB-RV (1)
4q12	56655069 56670532	15463	2	CEP135	9662	BM-X-SCID (1) CB-RV (1)

8p23.1	8226211	15832	2	AK122582	157285	CB-RV (1)
	8242043			AL833872		CB-ΔRV (1)
10p12.33	17495013	16203	2	AK127982		BM-ADA (1)
	17511216			ST8SIA6	338596	BM-X-SCID (1)
3p24.3	17987279	16944	2	intergenic		BM-ADA (1)
	18004223					CB-RV (1)
2p23.2	28072768	18488	2	BRE	9577	BM-ADA (1)
	28091256					CB-ΔRV (1)
16p13.11	16101813	18559	2	ABCC1	4363	BM-ADA (1)
	16120372					CB-RV (1)
2q14.3	122640188	19023	2	intergenic		BM-ADA (1)
	122659211					PB-ND (1)
20q13.12	42585420	19100	2	C20orf121	79183	CB-RV (2)
	42604520			PKIG	11142	
				SERINC3	10955	
3p21.31	50611909	20434	2	CISH	1154	BM-ADA (1)
	50632343			HEMK1	51409	CB-ΔRV (1)
				MAPKAPK3	7867	
14q24.1	69151131	21166	2	KIAA0247	9766	CB-RV (2)
	69172297					
16p13.13	11116122	21183	2	KIAA0350	23274	BM-ADA (1)
	11137305					BM-X-SCID (1)
13q32.1	94700653	22036	2	ABCC4	10257	BM-X-SCID (1)
	94722689					CB-ΔRV (1)
13q12.3	29844569	22741	2	intergenic		PB-ND (1)
	29867310					CB-ΔRV (1)
11q23.2	113488532	22851	2	ZBTB16	7704	BM-ADA (1)
	113511383					CB-RV (1)
3p14.3	57912744	24023	2	intergenic		CB-RV (2)
	57936767					
17q11.2	22898705	25919	2	KSR1	8844	BM-X-SCID (1)
	22924624					CB-RV (1)
11q24.1	122047138	27515	2	STS-1	84959	CB-ΔRV (1)
	122074653					PB-ND (1)

	12q24.23	117231701	27674	2	SUDS3	64426	BM-ADA (1)
		117259375			TAOK3	51347	CB-RV (1)
	20q13.32	55461409	28106	2	CTCFL	140690	BM-ADA (2)
		55489515			HMG1L1	10357	
	16q23.2	80390251	29371	2	PLCG2	5336	CB-RV (2)
		80419622					
	22q13.1	39153777	29588	2	AB051446	85373	BM-X-SCID (1)
		39183365			MKL1	57591	CB-ΔRV (1)
					RUTBC3	64783	
	3p26.1	4756100	30251	2	ITPR1	3708	BM-ADA (1)
		4786351					CB-RV (1)
	1p31.2	66515186	33770	2	L12685		CB-RV (1)
		66548956			PDE4B	5142	CB-ΔRV (1)
	7q34	137604232	35078	2	TRIM24	8805	CB-RV (1)
		137639310					PB-ND (1)
	16q24.1	84537775	35257	2	IRF8	3394	CB-RV (1)
		84573032					CB-ΔRV (1)
	19q13.13	43168201	35630	2	BC036863		BM-ADA (1)
		43203831			SIPA1L3	23094	CB-RV (1)
	7p13	44807804	35985	2	AL832992		BM-X-SCID (1)
		44843789			CCM2	83605	CB-RV (1)
					MYO1G	64005	
HeLa RV hot spots	19q13.2	47448798	29710	4	AK124207		§
(n=33)		47449000			CIC	23152	
		47478494			ERF	2077	
		47478508			GSK3A	2931	
					MGC70924	284338	
					PAFAH1B3	5050	
					ZNF526	116115	
	20q13.33	61326173	8638	3	BIRC7	79444	§
		61334598			C20 orf58	128414	

	61334811			YTHDF1	54915	
1q42.3	231399911	16782	3	intergenic		§
	231411975					
	231416693					
8p11.3	19438109	45960	3	ChGn	55790	§
	19482820					
	19484069					
1p21.3	94757430	67	2	intergenic		§
	74757497					
6p21.33	31851433	96	2	C6orf27	80737	§
	31851529			LSM2	57819	
				MSH5	4439	
				RDBP	7936	
				VARS	7407	
9q22.33	98090654	97	2	TBC1D2	55357	§
	98090751					
16p12.1	24565525	184	2	FLJ45256	400511	§
	24565709					
15q26.3	97258251	342	2	IGF1R	3480	§
	97258593					
9p13.3	33224502	353	2	BAG1	573	§
	33224855			SPINK4	27290	
5p15.1	16970165	365	2	MYO10	4651	§
	16970500					
12q22	92573626	365	2	CRADD	8738	§
	92573991					
14q24.3	74604776	369	2	ACYP1	97	§
	74605145			C14orf140	79696	
				MLH3	27030	
				NEK9	91754	
22q12.1	26585833	416	2	PITPNB	23760	§
	26586249					
17q11.2	26182221	582	2	C17orf41	79915	§
	26182803			CRLF3	51379	
12p13.2	10404056	754	2	KLRK1	22914	§

12p13.2	10404056	754	2	KLRK1	22914	§
	10404810					
Xq12	66566773	763	2	AR	367	§
	66567536					
10q23.33	96978922	1171	2	C10orf129	142827	§
	96980093			PDLIM1	9124	
11q13.3	69651813	1186	2	TMEM16A	55107	§
	69652999					
19q13.2	47079672	1239	2	ARHGEF1	9138	§
	47080911			CD79A	973	
				RPS19	6223	
7p21.3	12531894	1354	2	ARL4A	10124	§
	12533248					
9p24.1	4668861	1491	2	BC014133		§
	4670352			C9orf68	55064	
				CDC37L1	55664	
				PPAPDC2	403313	
1q41	214943437	1499	2	TGFB2	7042	§
	214944936					
11p13	32065944	1543	2	RCN1	5954	§
	32067487					
19p13.3	3555330	1794	2	C19orf29OS	404665	§
	3557124			GIPC3	126326	
				HMG20B	10362	
				PIP5K1C	23396	
				TBXA2R	6915	
18q11.2	17506483	2456	2	ABHD3	171586	§
	17508939					
9q21.13	72993839	2556	2	ANXA1	301	§
	72996395					
2p15	61898929	2623	2	intergenic	84140	§
	61901552				7514	
8q23.2	110688292	2712	2	FLJ20366	55638	§
	110691004					
15q26.2	95989920	3273	2	intergenic		§

7q21.11	79748528	3743	2	intergenic		§
	79752271					
7p11.2	54915760	4179	2	EGFR	1956	§
	54919939					
14q22.1	51613502	4306	2	C14orf166	51637	§
	51617808			NID2	22795	
17q11.2	22880336	4678	2	KSR1	8844	§
	22885014					
4p13	41550873	4756	2	DKFZP686A01247	22998	§
	41555629					
18q23	73946264	7191	2	intergenic	2587	§
	73953455				27164	
Xp22.11	23863173	8223	2	EIF2S3	1968	§
	23871396					
10q21.2	63180583	8886	2	C10orf107	219621	§
	63189469					
1p36.32	3443085	10461	2	ARHGEF1	9138	§
	3453546			MEGF6	1953	
7q22.3	106402101	11768	2	COG5	10466	§
	106413869			HBP1	26959	
				PRKAR2B	5577	
9p21.3	20598051	12134	2	MLLT3	4300	§
	20610185					
22q13.2	41478223	12764	2	A4GALT	53947	§
	41490987			ARFGAP3	26286	
10q23.32	93339808	17911	2	HECTD2	143279	§
	93357719			PPP1R3C	5507	
1q23.3	161429505	19165	2	PBX1	5087	§
	161448670					
12q23.3	103758195	24589	2	SLC41A2	84102	§
	103782784					
15q21.1	46413805	25413	2	DUT	1854	§
	46439218					
5p13.1	39128214	25846	2	FYB	2533	§
	39154060			RICTOR	253260	

	1q42.3	232373371	26113	2	LYST	1130	§
		232399484			NID1	4811	
	17q11.2	24483384	26189	2	MYO18A	399687	§
		24509573					
	9p24.2	4134826	28598	2	GLIS3	169792	§
		4163424					
	7q31.33	122981636	32737	2	HYAL4	23553	§
		123014373			WASL	8976	
	7p15.3	22635018	33996	2	DRCTNNB1A	84668	§
		22669014			TOMM7	54543	
CD34$^+$ LV hot spots	9q34.3	136607347	31043	3	AK074565		CB-LV (1)
(n=33)		136635022			Ak130247		CB-ΔLV (2)
		136638390			C9orf163	158055	
					INPP5E	56623	
					NOTCH1	4851	
					PMPCA	23023	
					SDCCAG3	10807	
	11q13.1	65587227	40000	3	AK090761		CB-LV (1)
		65587336			AK126488		CB-ΔLV (2)
		65627227			GAL3ST3	89792	
					PACS1	55690	
					SF3B2	10992	
	11q13.2	66706159	38546	3	AF118080		CB-LV (3)
		66739476			AK096194		
		66744705			FBXL11	22992	
	17q23.2	55188652	43091	3	TMEM49	81671	CB-LV (1)
		55229297					CB-ΔLV (2)
	12p13.31	6488535	942	2	CNAP1	9918	CB-LV (1)
		6489477			GAPDH	2597	CB-ΔLV (1)
					HOM-TES-10	25900	
					MRPL51	51258	

5q15	95018564	2300	2	RFESD	317671	CB-LV (1)
	95020864			SPATA9	83890	CB-ΔLV (1)
7p14.1	38694054	3290	2	VPS41	27072	CB-LV (2)
	38697344					
11p15.5	229114	4030	2	PSMD13	5719	CB-LV (2)
	233144			RIC8A	60626	
				SIRT3	23410	
9p24.1	6465526	4761	2	C9orf38		CB-ΔLV (2)
	6470287			UHRF2	115426	
13q31.1	78086565	4936	2	C13orf7	79596	CB-ΔLV (2)
	78091501			POU4F1	5457	
13q31.3	88838073	6952	2	intergenic		CB-ΔLV (2)
	88845025					
11q13.2	68034754	7330	2	C11orf23	53839	CB-LV (1)
	68042084			SAPS3	55291	CB-ΔLV (1)
11q13.1	64950133	7486	2	FKSG44	83786	CB-LV (1)
	64957619					CB-ΔLV (1)
6q14.1	79704830	7725	2	PHIP	55023	CB-LV (1)
	79712555					CB-ΔLV (1)
11q12.3	61874307	9359	2	ASRGL1	80150	CB-LV (1)
	61883666					CB-ΔLV (1)
3p21.31	49515892	9792	2	DAG1	1605	CB-ΔLV (2)
	49525684					
3p21.31	48962110	11013	2	ARIH2	10425	CB-LV (1)
	48973123			PH-4	54681	CB-ΔLV (1)
19q13.32	53431732	11507	2	CARD8	22900	CB-LV (1)
	53443239			ZNF114	163071	CB-ΔLV (1)
4q22.3	94671055	12281	2	GRID2	2895	CB-ΔLV (2)
	94683336					
20p13	3909280	12510	2	RNF24	11237	CB-LV (1)
	3921790					CB-ΔLV (1)
6q21	109414949	14905	2	ARMC2	84071	CB-LV (1)
	109429854			SESN1	27244	CB-ΔLV (1)

	6p21.31	34730265 34745513	15248	2	C6orf106	64771	CB-ΔLV (2)
	19p13.11	17986953 18002338	15385	2	ARRDC2 IL12RB1 KCNN1	27106 3594 3780	CB-LV (2)
	2p21	43643725 43665831	22106	2	THADA	63892	CB-LV (1) CB-ΔLV (1)
	6p21.1	42659594 42682583	22989	2	UBR2	23304	CB-ΔLV (2)
	13q33.1	103304775 103328165	23390	2	intergenic		CB-ΔLV (2)
	20p12.3	8323564 8347696	24132	2	PLCB1	23236	CB-LV (2)
	17p13.3	2158295 2184113	25818	2	RUTBC1 SMG6 SRR TSR1	9905 23293 63826 55720	CB-LV (2)
	2p13.3	69979229 70006225	26996	2	GMCL1 RY1	64395 11017	CB-ΔLV (2)
	2q14.3	124267757 124295138	27381	2	intergenic		CB-LV (1) CB-ΔLV(1)
	12p12.1	22536363 22565652	29289	2	KIAA0528	9847	CB-LV (2)
	17q23.2	57198972 57230593	31621	2	BRIP1	83990	CB-LV (1) CB-ΔLV (1)
	Xq28	152828754 152863883	35129	2	MECP2	4204	CB-ΔLV (2)
Controls (n=11)	4p16.3	2369595 2369710	115	2	AK126028 ZFYVE28	 57732	
	3q13.33	121289159 121291243	2084	2	GSK3B	2932	
	6q25.1	152351498	4561	2	ESR1	2099	

6q25.1	152351498 152356059	4561	2	ESR1	2099
2q31.2	178861724 178872384	10660	2	OSBPL6	114880
4q21.21	80722235 80738227	15992	2	intergenic	
4p16.2	5389283 5413524	24241	2	STK32B	55351
2q32.3	194366032 194390363	24331	2	intergenic	
12p12.3	19177339 19207223	29884	2	PLEKHA5	54477
4q35.2	188799358 188831151	31793	2	intergenic	
4q32.3	170027856 170059943	32087	2	PALLD	23022
2q22.3	147690603	34919	2	intergenic	

* Source of hot spot integrations (the number in parentheses indicates the number of hits for each category):

CB-RV: cord blood-derived CD34$^+$ cells transduced with wild-type LTR RV

CB-ΔRV: cord blood-derived CD34$^+$ cells transduced with U3 deleted-LTR RV

BM-ADA: bone marrow-derived CD34$^+$ cells from ADA-SCID patients transduced with wt LTR RV

BM-X-SCID: bone marrow-derived CD34$^+$ cells from X-SCID patients transduced with wt LTR RV

PB-ND: peripheral blood-derived CD34$^+$ cells from normal donor transduced with wt LTR RV

CB-LV: cord blood-derived CD34$^+$ cells transduced with wild type-LTR LV

CB-ΔLV: cord blood-derived CD34$^+$ cells transduced with sin-18-LTR LV

§: HeLa cells transduced with wild-type LTR RV by Wu et al[12].

Appendix 5.

Complete statistics of Figures 21 and 27. Statistical comparison of the frequency of TFBS distributions (motif count/sequence) in **Fig. 21** and **27**. All combinations of sample diversities were in turn assessed by exact Wilcoxon rank sum test (one-sided, alternative hypothesis: "greater"), independently in CD34$^+$ HSCs and in HeLa cells. In CD34$^+$ cells, the control group distribution is significantly shifted to the right (greater) with respect to HIV, ΔU3-HIV[CMV] and ΔU3-HIV[MLV] distributions ($p < 2.2e\text{-}16$), but not to MLV, ΔU3-MLV, SFFV-MLV and MLV-HIV data sets ($p = 1$). The other way around, MLV distribution is significantly shifted to the right (greater) of all the other distributions, excepted for SFFV-MLV group. Similarly, in HeLa cells MLV distribution is significantly shifted to the right (greater) with respect to both HIV and HIVmIN distributions, which, in turn, are statistically different one from each other ($p < 2.2e\text{-}16$). Asterisks specify the values highlighted in **Figure 23**.

CD34$^+$ HSCs	Controls	MLV	ΔU3-MLV	SFFV-MLV	HIV	ΔU3-HIV[CMV]	ΔU3-HIV[MLV]	MVL-HIV
Controls	-	1	1	1	< 2.2e-16	< 2.2e-16	< 2.2e-16	1
MLV	< 2.2e-16 *	-	< 2.2e-16 *	0.900	< 2.2e-16	< 2.2e-16	< 2.2e-16	< 2.2e-16
ΔU3-MLV	< 2.2e-16	1	-	1	< 2.2e-16	< 2.2e-16	< 2.2e-16	< 4.4e-16
SFFV-MLV	< 2.2e-16	0.100	< 2.2e-16	-	< 2.2e-16	< 2.2e-16	< 2.2e-16	< 2.2e-16
HIV	1	1	1	1	-	1	1.6e-08	1
ΔU3-HIV[CMV]	1	1	1	1	6.6e-13	-	< 2.2e-16	1
ΔU3-HIV[MLV]	1	1	1	1	1	1	-	1
MLV-HIV	1.4e-08	1	1	1	< 2.2e-16	< 2.2e-16	< 2.2e-16 *	-

HeLa cells	**MVL**	**HIV**	**HIVmIN**
MLV	-	< 2.2e-16	< 2.2e-16
HIV	1	-	1
HIVmIN	1	< 2.2e-16	-

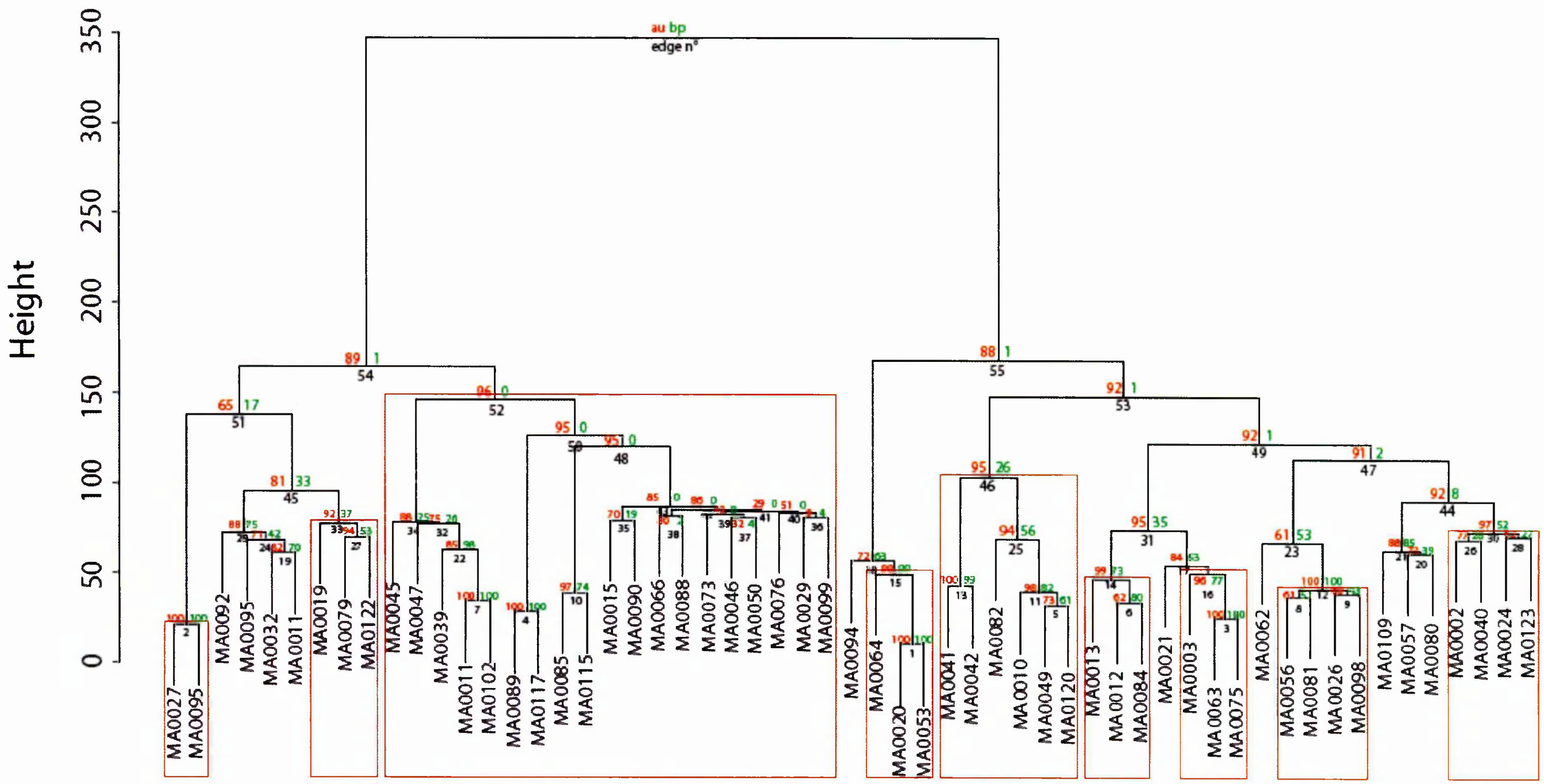
Appendix 6.

Bootstrapped matrix column dendrograms of the hierarchical cluster analyses shown in **Figures 22**, **28** and **31**. Column dendrograms have been sampled with 10,000 bootstrap replicates and approximately unbiased probabilities are reported on each node (AU, red).

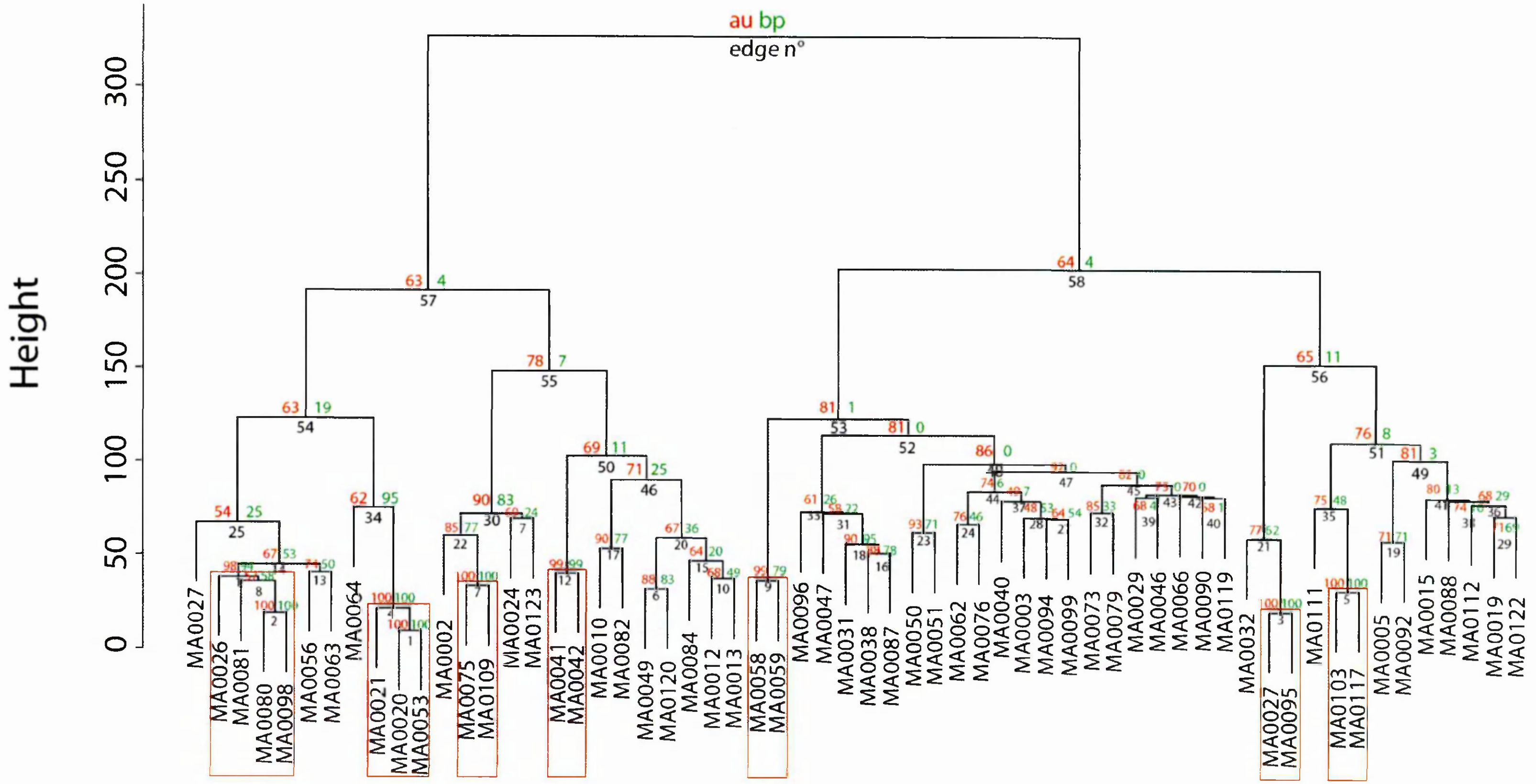
Red rectangles on the tree identify nodes with an AU value > 0.95, hence considered significant, stable nodes and corresponding to matrices highly connected to each vector TFBS profile.

Distance: euclidean; cluster method: ward.

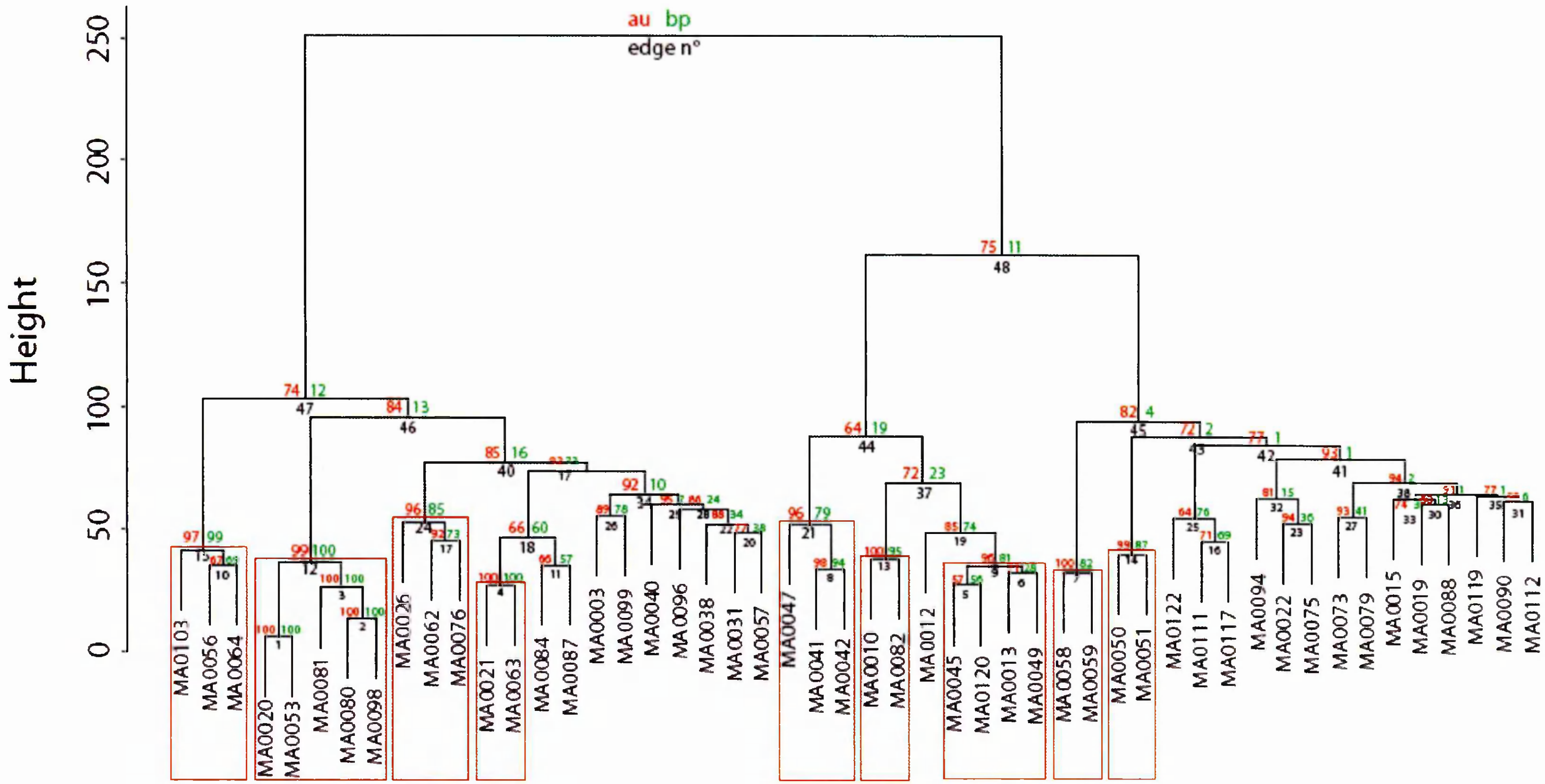
6.1 Bootstrapped matrix column dendrogram of the hierarchical cluster analysis shown in **Figure 24** (RV and LV vectors in CD34^{+} HSCs).



6.2 Bootstrapped matrix column dendrogram of the hierarchical cluster analysis shown in **Figure 30** (MLV and HIV vectors in CD34^{+} HSCs and HeLa cells.



_6.3_ Bootstrapped matrix column dendrogram of the hierarchical cluster analysis shown in **Figure 33** (HIV, MLV and HIVmIN vectors in HeLa cells).



Appendix 7.

Colver results. Distribution of Jaspar matrices (average number of matrices/sequence and 1st to 99th percentile range, in parentheses) in sequences flanking (± 1,000 bp) integration sites of different RV and LV vectors in human CD34$^+$ HSCs and in HeLa cells (see **Figure 11** for vector identification).

CD34$^+$ HSCs										
MATRIX	**TF**	**TF family**	**Control**	**MLV**	**ΔU3-MLV**	**SFFV-MLV**	**HIV**	**ΔU3-HIV[CMV]**	**ΔU3-HIV[MLV]**	**MLV-HIV**
MA0001	AGL3	MADS	0	0	0	0	0	0	0	0
MA0002	RUNX1	RUNT	0	1.38 (0-4)	0	0	0	0	0	0
MA0003	TFAP2A	AP2	0	3.28 (0-11)	0	3.27 (0-10)	0	0	0	0
MA0004	Arnt	bHLH	0	0	0	0	0	0	0	0
MA0005	Agamous	MADS	0	0	0	0	0	1.64 (0-6)	0	1.55 (0-5)
MA0006	Arnt-Ahr	bHLH	0	0	0	0	0	0	0	0
MA0007	Ar	Nuclear receptor	0	0	0	0	0	0	0	0
MA0008	Athb-1	HOMEO-ZIP	0	0	0	0	0	0	0	0
MA0009	T	T-BOX	0	0	0	0	0	0	0	0
MA0010	Broad-complex_1	Zn-finger, C2H2	0	4.53 (0-28.72)	0	0	0	0	0	0
MA0011	Broad-complex_2	Zn-finger, C2H2	0	0	0	1.12 (0-5.96)	0	0	0	0
MA0012	Broad-complex_3	Zn-finger, C2H2	0	1.84 (0-11.72)	0	2.11 (0-10.18)	0	0	0	0
MA0013	Broad-complex_4	Zn-finger, C2H2	0	2.77 (0-20)	0	3.00 (0-24.12)	0	0	0	0
MA0014	Pax5	PAIRED	0	0	0	0	0	0	0	0
MA0015	CF2-II	Zn-finger, C2H2	3.36 (0-46)	0	0	0	0	0	0	0
MA0016	CFI-USP	Nuclear receptor	0	0	0	0	0	0	0	0

MA0017	NR2F1	Nuclear receptor	0	0	0	0	0	0	0	0
MA0018	CREB1	bZIP	0	0	0	0	0	0	0	0
MA0019	Chop-cEBP	bZIP	0	0	0	0	2.63 (0-9.98)	2.70 (0-10.56)	3.19 (0-10)	0
MA0020	Dof2	Zn-finger, DOF	3.38 (0-9)	3.86 (0-10)	4.08 (0-11.01)	3.68 (0-11.06)	0	0	0	3.41 (0-9.01)
MA0021	Dof3	Zn-finger, DOF	0	3.50 (0-10)	0	3.77 (0-11)	0	0	0	2.91 (0-7.01)
MA0022	Dorsal_1	REL	0	0	0	0	0	0	0	0
MA0023	Dorsal_2	REL	0	0	0	0	0	0	0	0
MA0024	E2F1	Unknown	0	0.61 (0-3)	0	0	0	0	0	0
MA0025	NFIL3	bZIP	0	0	0	0	0	0	0	0
MA0026	E74A	ETS	0	2.34 (0-7)	2.37 (0-6.01)	2.30 (0-8)	0	0	0	0
MA0027	En1	HOMEO	0	0	0	0	0.64 (0-3)	0	0	0
MA0028	ELK1	ETS	0	0	0	0	0	0	0	0
MA0029	Evi1	Zn-finger, C2H2	0	0.83 (0-4)	0	0	0	0	0	0
MA0030	FOXF2	Forkhead	0	0	0	0	0	0	0	0
MA0031	FOXD1	Forkhead	0	0	0	0	0	0	0	0
MA0032	FOXC1	Forkhead	0	0	0	0	0.14 (0-2)	0.13 (0-2)	0.10 (0-1.01)	0.12 (0-2)
MA0032	FOXC1	Forkhead	0	0	0	0	0	0	0	0
MA0033	FOXL1	Forkhead	0	0	0	0	0	0	0	0
MA0034	GAMYB	TRP-CLUSTER	0	0	0	0	0	0	0	0
MA0035	Gata1	Zn-finger, GATA	0	0	0	0	0	0	0	0
MA0036	GATA2	Zn-finger, GATA	0	0	0	0	0	0	0	0
MA0037	GATA3	Zn-finger, GATA	0	0	0	0	0	0	0	0
MA0039	Klf4	Zn-finger, C2H2	0	0	4.98 (0-13.02)	4.97 (0-13.12)	0	0	0	0
MA0040	Foxq1	Forkhead	0	1.63 (0-6)	0	0	0	0	0	0
MA0041	Foxd3	Forkhead	0	4.18 (0-22.72)	0	4.70 (0-26.06)	0	0	0	0
MA0042	FOXI1	Forkhead	0	2.67 (0-11)	0	2.85 (0-10.06)	0	0	0	0
MA0043	HLF	bZIP	0	0	0	0	0	0	0	0
MA0044	HMG-1	HMG	0	0	0	0	0	0	0	0
MA0045	HMG-IY	HMG	0	0	0	6.85 (0-31.06)	0	0	0	0

MA0046	TCF1	HOMEO	0	1.05 (0-5)	0	1.23 (0-5)	0	0	0	0
MA0047	Foxa2	Forkhead	0	0	0	2.45 (0-10.06)	0	0	0	0
MA0048	NHLH1	bHLH	0	0	0	0	0	0	0	0
MA0049	Hunchback	Zn-finger, C2H2	0	6.31 (0-32.72)	0	7.13 (0-37.36)	0	0	0	0
MA0050	IRF1	TRP-CLUSTER	0	1.76 (0-6)	1.63 (0-5)	1.82 (0-7)	0	0	0	0
MA0051	IRF2	TRP-CLUSTER	0	0	0	0	0	0	0	0
MA0052	MEF2A	MADS	0	0	0	0	0	0	0	0
MA0053	MNB1A	Zn-finger, DOF	2.31 (0-8)	2.70 (0-9)	2.68 (0-9)	2.56 (0-8.06)	0	0	0	2.51 (0-9.01)
MA0054	MYB.ph3	TRP-CLUSTER	0	0	0	0	0	0	0	0
MA0055	Myf	bHLH	0	0	0	0	0	0	0	0
MA0056	ZNF42_1-4	Zn-finger, C2H2	0	3.68 (0-11)	4.06 (0-10)	3.90 (0-11.06)	0	0	0	0
MA0057	ZNF42_5-13	Zn-finger, C2H2	2.54 (0-8)	2.55 (0-8)	0	0	0	0	0	0
MA0058	MAX	bHLH-ZIP	0	0	0	0	0	0	0	0
MA0059	MYC-MAX	bHLH-ZIP	0	0	0	0	0	0	0	0
MA0060	NF-Y	CAAT-BOX	0	0	0	0	0	0	0	0
MA0061	NF-kappaB	REL	0	0	0	0	0	0	0	0
MA0062	GABPA	ETS	0	1.55 (0-5)	1.67 (0-5)	0	0	0	0	0
MA0063	Nkx2-5	HOMEO	0	1.07 (0-5)	0	1.12 (0-5.06)	0	0	0	0
MA0064	PBF	Zn-finger, DOF	1.97 (0-7)	2.44 (0-9)	2.42 (0-8.01)	2.36 (0-8)	1.97 (0-7)	0	0	2.15 (0-9.01)
MA0065	PPARG-RXRA	Nuclear receptor	0	0	0	0	0	0	0	0
MA0066	PPARG	Nuclear receptor	0	0	0	0	0	0	0	0.44 (0-3)
MA0067	Pax2	PAIRED	0	0	0	0	0	0	0	0
MA0068	Pax4	PAIRED-HOMEO	0	0	0	0	0	0	0	0
MA0069	Pax6	PAIRED	0	0	0	0	0	0	0	0
MA0070	Pbx	HOMEO	0	0	0	0	0	0	0	0
MA0071	RORA	Nuclear receptor	0	0	0	0	0	0	0	0
MA0072	RORA1	Nuclear receptor	0	0	0	0	0	0	0	0
MA0073	RREB1	Zn-finger, C2H2	0	0	0	3.90 (0-21.42)	0	0	0	0
MA0074	RXR-VDR	Nuclear receptor	0	0	0	0	0	0	0	0

MA0075	Prrx2	HOMEO	0	2.67 (0-12)	0	3.16 (0-12.12)	0	0	0	0
MA0076	ELK4	ETS	0	0	0.565 (0-3)	0	0	0	0	0
MA0077	SOX9	HMG	0	0	0	0	0	0	0	0
MA0078	Sox17	HMG	0	0	0	0	0	0	0	0
MA0079	SP1	Zn-finger, C2H2	0	3.19 (0-9.72)	0	3.39 (0-9)	0	3.45 (0-10)	0	3.52 (0-10)
MA0080	SPI1	ETS	3.18 (0-7)	3.70 (0-10)	3.71 (0-10)	3.44 (0-10)	0	0	0	0
MA0081	SPIB	ETS	0	3.78 (0-9)	3.94 (0-10)	3.56 (0-9)	0	0	0	0
MA0082	SQUA	MADS	0	2.70 (0-10)	0	3.03 (0-11.12)	0	0	0	0
MA0083	SRF	MADS	0	0	0	0	0	0	0	0
MA0084	SRY	HMG	0	2.15 (0-7.72)	0	2.35 (0-7.06)	0	0	0	0
MA0085	SU_h	IPT/TIG domain	0	0	0	1.35 (0-4.06)	0	0	0	0
MA0086	Snail	Zn-finger, C2H2	0	0	0	0	0	0	0	0
MA0087	Sox5	HMG	0	0	0	0	0	0	0	0
MA0088	Staf	Zn-finger, C2H2	0	0	0	0	2.52 (0-7)	2.50 (0-8)	0	2.59 (0-8)
MA0089	TCF11-MafG	bZIP	1.66 (0-6)	0	0	0	0	0	0	0
MA0090	TEAD	TEA	1.69 (0-5.56)	0	0	0	0	0	0	0
MA0091	TAL1-TCF3	bHLH	0	0	0	0	0	0	0	0
MA0092	HAND1-TCF3	bHLH	0	0	0	0	0	0	0	2.57 (0-7)
MA0093	USF1	bHLH-ZIP	0	0	0	0	0	0	0	0
MA0094	Ubx	HOMEO	7.00	0.10 (0-4)	0	0.09 (0-2.18)	0	0	0	5.00
MA0095	YY1	Zn-finger, C2H2	0	0	0	0	2.15 (0-6)	0	0	0
MA0096	bZIP910	bZIP	0	0	0	0	0	0	0	0
MA0097	bZIP911	bZIP	0	0	0	0	0	0	0	0
MA0098	c-ETS	ETS	0	1.09 (0-5.72)	0.99 (0-5)	0	0	0	0	0
MA0099	Fos	bZIP	2.21 (0-7)	2.27 (0-7)	2.26 (0-7)	0	0	0	2.51 (0-7)	2.41 (0-8)
MA0100	Myb	TRP-CLUSTER	0	0	0	0	0	0	0	0
MA0101	REL	REL	0	0	0	0	0	0	0	0
MA0102	cEBP	bZIP	0	0	0	1.58 (0-6)	0	0	0	0
MA0103	deltaEF1	Zn-finger, C2H2	0	0	0	0	0	0	0	0

MA0104	Mycn	bHLH-ZIP	0	0	0	0	0	0	0	0
MA0105	NFKB1	REL	0	0	0	0	0	0	0	0
MA0106	TP53	P53	0	0	0	0	0	0	0	0
MA0107	RELA	REL	0	0	0	0	0	0	0	0
MA0108	TBP	TATA-box	0	0	0	0	0	0	0	0
MA0109	RUSH1-alfa	Zn-finger, GATA	0.52 (0-3)	0.63 (0-3)	0	0	0	0.57 (0-3)	0	0
MA0110	ATHB5	HOMEO-ZIP	0	0	0	0	0	0	0	0
MA0111	Spz1	bHLH-ZIP	0	0	0	0	0	1.99 (0-6)	2.10 (0-5.01)	2.06 (0-5.01)
MA0112	ESR1	NUCLEAR	0	0	0	0	0	0	0	0
MA0113	NR3C1	NUCLEAR	0	0	0	0	0	0	0	0
MA0114	HNF4	NUCLEAR	0	0	0	0	0	0	0	0
MA0115	NR1H2-RXR	Nuclear receptor	0	0	0	0.08 (0-1)	0	0	0	0
MA0116	Roaz	Zn-finger, C2H2	0	0	0	0	0	0	0	0
MA0117	MafB	bZIP, MAF	1.64 (0-5)	0	0	0	0	0	0	0
MA0118	Macho-1	Zn-finger, C2H2	0	0	0	0	0	0	0	0
MA0119	Hox11-CTF1	HOMEO/CAAT	0	0	0	0	0	0	0	0
MA0120	ID1	Zn-finger, C2H2	0	3.07 (0-20.72)	0	3.05 (0-20.18)	0	0	0	0
MA0121	ARR10	TRP-CLUSTER	0	0	0	0	0	0	0	0
MA0122	Bapx1	HOMEO	2.64 (0-7)	2.69 (0-7)	0	0	2.77 (0-7)	2.89 (0-8)	2.76 (0-7)	2.80 (0-7)
MA0123	ABI4	AP2	0	2.08 (0-10)	0	0	0	0	0	0

HeLa cells

MATRIX	TF	TF family	MLV	HIV	HIVmIN
MA0001	AGL3	MADS	0	0	0
MA0002	RUNX1	RUNT	0	0	0
MA0003	TFAP2A	AP2	3.22 (0-11)	0	2.89 (0-9)
MA0004	Arnt	bHLH	0	0	0
MA0005	Agamous	MADS	0	0	0
MA0006	Arnt-Ahr	bHLH	0	0	0

MA0007	Ar	Nuclear receptor	0	0	0
MA0008	Athb-1	HOMEO-ZIP	0	0	0
MA0009	T	T-BOX	0	0	0
MA0010	Broad-complex_1	Zn-finger, C2H2	0	0	4.76 (0-29.76)
MA0011	Broad-complex_2	Zn-finger, C2H2	0	0	0
MA0012	Broad-complex_3	Zn-finger, C2H2	1.93 (0-13.37)	0	2.05 (0-13.52)
MA0013	Broad-complex_4	Zn-finger, C2H2	2.73 (0-25)	0	2.97 (0-23.52)
MA0014	Pax5	PAIRED	0	0	0
MA0015	CF2-II	Zn-finger, C2H2	0	0	0
MA0016	CFI-USP	Nuclear receptor	0	0	0
MA0017	NR2F1	Nuclear receptor	0	0	0
MA0018	CREB1	bZIP	0	0	0
MA0019	Chop-cEBP	bZIP	0	3.09 (0-10)	0
MA0020	Dof2	Zn-finger, DOF	3.89 (0-10)	0	0
MA0021	Dof3	Zn-finger, DOF	3.39 (0-9)	0	3.44 (0-9)
MA0022	Dorsal_1	REL	0	0	1.88 (0-6)
MA0023	Dorsal_2	REL	0	0	0
MA0024	E2F1	Unknown	0	0	0
MA0025	E74A	ETS	0	0	0
MA0026	E74A	ETS	2.18 (0-7)	0	1.98 (0-7)
MA0027	En1	HOMEO	0	0	0
MA0028	ELK1	ETS	0	0	0
MA0029	Evi1	Zn-finger, C2H2	0	0	0
MA0030	FOXF2	Forkhead	0	0	0
MA0031	FOXD1	Forkhead	1.39 (0-5)	0	0
MA0032	FOXC1	Forkhead	0	0	0
MA0032	FOXC1	Forkhead	0	0	0
MA0033	FOXL1	Forkhead	0	0	0
MA0034	GAMYB	TRP-CLUSTER	0	0	0
MA0035	Gata1	Zn-finger, GATA	0	0	0
MA0036	GATA2	Zn-finger, GATA	0	0	0

MA0037	GATA3	Zn-finger, GATA	0	0	0
MA0038	Gfi	Zn-finger, C2H2	2.29 (0-6)	0	0
MA0040	Foxq1	Forkhead	1.71 (0-5.37)	0	0
MA0041	Foxd3	Forkhead	4.04 (0-25.74)	0	4.49 (0-22)
MA0042	FOXI1	Forkhead	2.52 (0-9)	0	2.48 (0-11)
MA0043	HLF	bZIP	0	0	0
MA0044	HMG-1	HMG	0	0	0
MA0045	HMG-IY	HMG	0	0	6.99 (0-36.52)
MA0046	TCF1	HOMEO	0	0	0
MA0047	Foxa2	Forkhead	2.31 (0-8)	0	2.22 (0-7)
MA0048	NHLH1	bHLH	0	0	0
MA0049	Hunchback	Zn-finger, C2H2	5.89 (0-36)	0	6.64 (0-36.76)
MA0050	IRF1	TRP-CLUSTER	1.58 (0-6)	0	0
MA0051	IRF2	TRP-CLUSTER	0.82 (0-4)	0	0
MA0052	MEF2A	MADS	0	0	0
MA0053	MNB1A	Zn-finger, DOF	2.76 (0-9.37)	0	0
MA0054	MYB.ph3	TRP-CLUSTER	0	0	0
MA0055	Myf	bHLH	0	0	0
MA0056	ZNF42_1-4	Zn-finger, C2H2	3.68 (0-11)	0	0
MA0057	ZNF42_5-13	Zn-finger, C2H2	2.61 (0-8.37)	0	0
MA0058	MAX	bHLH-ZIP	1.06 (0-5)	0	0
MA0059	MYC-MAX	bHLH-ZIP	0.99 (0-5)	0	0
MA0060	NF-Y	CAAT-BOX	0	0	0
MA0061	NF-kappaB	REL	0	0	0
MA0062	GABPA	ETS	1.48 (0-5)	0	0
MA0063	Nkx2-5	HOMEO	1.04 (0-5)	0	0.97 (0-5)
MA0064	PBF	Zn-finger, DOF	2.47 (0-9.37)	2.00 (0-6.69)	0
MA0065	PPARG-RXRA	Nuclear receptor	0	0	0
MA0066	PPARG	Nuclear receptor	0	0	0
MA0067	Pax2	PAIRED	0	0	0
MA0068	Pax4	PAIRED-HOMEO	0	0	0

MA0069	Pax6	PAIRED	0	0	0
MA0070	Pbx	HOMEO	0	0	0
MA0071	RORA	Nuclear receptor	0	0	0
MA0072	RORA1	Nuclear receptor	0	0	0
MA0073	RREB1	Zn-finger, C2H2	3.21 (0-20)	0	0
MA0074	RXR-VDR	Nuclear receptor	0	0	0
MA0075	Prrx2	HOMEO	0	0	2.81 (0-11.76)
MA0076	ELK4	ETS	0.56 (0-3)	0	0
MA0077	SOX9	HMG	0	0	0
MA0078	Sox17	HMG	0	0	0
MA0079	SP1	Zn-finger, C2H2	3.28 (0-10)	3.67 (0-10.69)	3.29 (0-9)
MA0080	SPI1	ETS	3.56 (0-10)	0	0
MA0081	SPIB	ETS	3.50 (0-9)	0	0
MA0082	SQUA	MADS	0	0	2.99 (0-12.76)
MA0083	SRF	MADS	0	0	0
MA0084	SRY	HMG	2.19 (0-8)	0	2.21 (0-8)
MA0085	SU_h	IPT/TIG domain	0	0	0
MA0086	Snail	Zn-finger, C2H2	0	0	0
MA0087	Sox5	HMG	1.63 (0-6)	0	0
MA0088	Staf	Zn-finger, C2H2	0	2.67 (0-7)	0
MA0089	TCF11-MafG	bZIP	0	0	0
MA0090	TEAD	TEA	1.61 (0-5)	0	0
MA0091	TAL1-TCF3	bHLH	0	0	0
MA0092	HAND1-TCF3	bHLH	0	0	0
MA0093	USF1	bHLH-ZIP	0	0	0
MA0094	Ubx	HOMEO	0.11 (0-5)	0.01	0.08 (0-2.76)
MA0095	YY1	Zn-finger, C2H2	0	0	0
MA0096	bZIP910	bZIP	0.22 (0-2)	0	0
MA0097	bZIP911	bZIP	0	0	0
MA0098	c-ETS	ETS	1.11 (0-5.37)	0	0
MA0099	Fos	bZIP	2.84 (0-8)	0	2.47 (0-7.76)

MA0100	Myb	TRP-CLUSTER	0	0	0
MA0101	REL	REL	0	0	0
MA0102	cEBP	bZIP	0	0	0
MA0103	deltaEF1	Zn-finger, C2H2	0	3.85 (0-10)	0
MA0104	Mycn	bHLH-ZIP	0	0	0
MA0105	NFKB1	REL	0	0	0
MA0106	TP53	P53	0	0	0
MA0107	RELA	REL	0	0	0
MA0108	TBP	TATA-box	0	0	0
MA0109	RUSH1-alfa	Zn-finger, GATA	0	0	0
MA0110	ATHB5	HOMEO-ZIP	0	0	0
MA0111	Spz1	bHLH-ZIP	0	2.01 (0-5)	0
MA0112	ESR1	NUCLEAR	0	2.29 (0-8.38)	0
MA0113	NR3C1	NUCLEAR	0	0	0
MA0114	HNF4	NUCLEAR	0	0	0
MA0115	NR1H2-RXR	Nuclear receptor	0	0	0
MA0116	Roaz	Zn-finger, C2H2	0	0	0
MA0117	MafB	bZIP, MAF	0	1.62 (0-5)	0
MA0118	Macho-1	Zn-finger, C2H2	0	0	0
MA0119	Hox11-CTF1	HOMEO/CAAT	0.91 (0-4)	0	0
MA0120	ID1	Zn-finger, C2H2	3.01 (0-24.37)	0	3.39 (0-25.76)
MA0121	ARR10	TRP-CLUSTER	0	0	0
MA0122	Bapx1	HOMEO	0	0 (0-7)	2.64 (0-7)
MA0123	ABI4	AP2	0	0	0

Appendix 8.

Results of conserved TFBS analysis in human HSCs and in HeLa cells.

8.1 Statistical significance of evolutionarily conserved TFBSs found ±1,000 bp around retroviral insertions in CD34$^+$ hematopoietic cells and in HeLa cells. A Fisher's exact test (one-sided, alternative hypothesis: "greater") was applied to compare the percentage of sequences containing at least one conserved TFBS in the experimental data sets, including control sequences, with percentages found in corresponding fitted and random backgrounds (100,000 *in silico* generated sequences).

Cell type	**Data set**	**Data set *vs.* Background**	**Data set *vs.* Random**
CD34$^+$ cells	Controls	5.8e-05	9.7e-06
	MLV	6.4e-20	1.5e-33
	ΔU3-MLV	0.178	0.026
	SFFV-MLV	2.1e-04	2.1e-04
	HIV	0.511	0.023
	ΔU3-HIV[CMV]	0.098	2.3e-04
	ΔU3-HIV[MLV]	0.475	0.068
	MLV-HIV	1.4e-03	9.3e-08
HeLa cells			
	MLV	4.1e-18	3.3e-30
	HIV	0.053	3.1e-06
	HIVmIN	1.4e-06	6.8e-10

8.2 Total counts of TRANSFAC conserved matrices in sequences flanking (± 1,000 bp) the integration sites of different RV and LV vectors in human CD34$^+$ HSCs and HeLa cells (see **Figure 11** for vector identification).

* $p < 0.05$, Fisher's exact test over matched background.

CD34$^+$ cells

Matrix	Accession number	Binding factors	Controls	MLV	ΔU3-MLV	SFFV-MLV	HIV	ΔU3-HIV[CMV]	ΔU3-HIV[MLV]	MLV-HIV
M00002	V$E47_01	E47	0	0	0	0	0	0	0	0
M00005	V$AP4_01	AP-4	0	0	0	0	0	0	0	0
M00006	V$MEF2_01	MEF-2A	*11	22	0	4	6	*12	*6	2
M00007	V$ELK1_01	Elk-1	0	0	0	*2	0	0	0	0
M00017	V$ATF_01	ATF	0	2	0	2	0	0	0	0
M00024	V$E2F_01	E2F	0	*8	0	0	0	0	0	0
M00025	V$ELK1_02	Elk-1	0	4	0	0	0	0	0	0
M00026	V$RSRFC4_01	RSRFC4	7	6	0	*6	0	6	0	2
M00033	V$P300_01	p300	0	0	0	0	0	0	0	0
M00034	V$P53_01	p53	0	2	*4	0	0	0	0	0
M00037	V$NFE2_01	NF-E2	2	12	0	2	0	0	0	2
M00039	V$CREB_01	CREB, deltaCREB	0	2	0	2	0	0	0	0
M00040	V$CREBP1_01	ATF-2	1	6	2	*6	0	0	0	0
M00041	V$CREBP1CJUN_01	ATF-2, c-Jun	0	4	2	0	0	0	2	0
M00045	V$E4BP4_01	E4BP4	6	12	2	*10	*10	*14	2	4
M00050	V$E2F_02	E2F, E2F-1, E2F-2, E2F-3a, E2F-4, E2F-5	0	4	2	0	0	0	0	0
M00051	V$NFKAPPAB50_01	NF-kappaB1	0	0	0	0	2	0	0	0
M00052	V$NFKAPPAB65_01	RelA	2	4	0	0	0	*6	0	2
M00053	V$CREL_01	c-Rel	0	6	2	0	0	*8	0	0

M00054	V$NFKAPPAB_01	NF-kappaB, NF-kappaB1, RelA	0	0	0	0	0	4	0	0
M00056	V$MYOGNF1_01	NF-1	0	4	2	0	0	0	0	0
M00059	V$YY1_01	YY1	3	2	0	0	0	2	0	0
M00062	V$IRF1_01	IRF-1	4	*20	0	4	4	2	2	6
M00065	V$TAL1BETAE47_01	E47, Tal-1beta	2	*8	0	0	0	0	0	0
M00066	V$TAL1ALPHAE47_01	E47, Tal-1	2	8	0	0	0	2	0	4
M00069	V$YY1_02	YY1	2	*12	0	0	0	0	0	2
M00070	V$TAL1BETAITF2_01	ITF-2, Tal-1beta	1	*12	2	0	0	2	0	4
M00071	V$E47_02	E47	0	0	0	0	0	0	0	0
M00076	V$GATA2_01	GATA-2	0	0	0	0	0	0	0	0
M00077	V$GATA3_01	GATA-3	2	0	0	0	0	0	0	0
M00084	V$MZF1_02	MZF-1	0	2	0	0	0	0	0	0
M00085	V$ZID_01	ZID	3	4	0	2	0	0	0	0
M00095	V$CDP_01	CUTL1	4	12	2	2	4	4	2	6
M00096	V$PBX1_01	Pbx1a	2	10	0	*4	0	4	0	*6
M00097	V$PAX6_01	Pax-6	5	8	0	*6	2	2	0	0
M00098	V$PAX2_01	Pax-2	0	4	0	0	0	0	0	0
M00102	V$CDP_02	CUTL1	*23	16	4	6	12	8	6	14
M00104	V$CDPCR1_01	CUTL1	1	2	0	0	0	0	0	0
M00105	V$CDPCR3_01	CUTL1	7	10	0	2	2	4	0	0
M00106	V$CDPCR3HD_01	CUTL1	4	0	0	0	0	2	0	2
M00109	V$CEBPB_01	C/EBPbeta	2	4	0	0	2	4	2	0
M00113	V$CREB_02	CREB, deltaCREB	0	2	0	0	0	0	0	0
M00114	V$TAXCREB_01	CREB, deltaCREB	0	2	0	0	0	0	0	0
M00115	V$TAXCREB_02	CREB, deltaCREB	2	4	0	0	0	0	0	0
M00116	V$CEBPA_01	C/EBPalpha	3	4	2	0	2	2	2	2
M00117	V$CEBPB_02	C/EBPbeta	1	10	0	0	0	0	0	0
M00118	V$MYCMAX_01	c-Myc, Max1	0	2	2	0	0	0	2	0
M00119	V$MAX_01	Max1	0	0	*2	0	0	0	0	0
M00121	V$USF_01	USF1	0	2	*4	2	0	0	0	0

M00122	V$USF_02	USF1	0	0	0	0	0	0	0	0
M00123	V$MYCMAX_02	c-Myc, Max1	0	0	0	0	0	0	0	0
M00124	V$PBX1_02	Pbx1a	3	10	0	4	6	4	2	8
M00126	V$GATA1_02	GATA-1	*7	0	2	0	0	2	0	2
M00127	V$GATA1_03	GATA-1	0	6	2	0	0	0	0	2
M00128	V$GATA1_04	GATA-1	1	4	0	2	2	0	0	0
M00130	V$FOXD3_01	FOXD3	*6	6	0	2	0	0	0	4
M00132	V$HNF1_01	HNF-1A	*9	12	2	2	2	2	4	0
M00133	V$TST1_01	POU3F1	3	*12	0	0	2	4	0	0
M00134	V$HNF4_01	HNF-4alpha2	0	2	0	0	0	0	0	0
M00135	V$OCT1_01	POU2F1	6	*20	*6	2	2	4	2	6
M00136	V$OCT1_02	POU2F1	9	12	0	*8	*12	10	4	8
M00137	V$OCT1_03	POU2F1	5	2	0	2	2	4	0	0
M00138	V$OCT1_04	POU2F1	*5	8	*8	0	0	2	0	*8
M00143	V$PAX5_01	Pax-5	0	0	0	0	0	0	0	0
M00144	V$PAX5_02	Pax-5	0	2	0	0	0	2	0	2
M00145	V$BRN2_01	POU3F2	*12	16	0	0	6	8	0	8
M00146	V$HSF1_01	HSF1 (long)	4	4	0	2	2	2	2	2
M00147	V$HSF2_01	HSF2	2	8	0	2	0	0	2	0
M00152	V$SRF_01	SRF	1	2	0	0	2	4	0	4
M00155	V$ARP1_01	ARP-1	0	8	2	2	2	0	0	0
M00156	V$RORA1_01	RORalpha1	*4	0	0	0	2	2	0	0
M00157	V$RORA2_01	RORalpha2	*11	12	2	0	*12	*12	0	2
M00158	V$COUP_01	COUP-TF1, HNF-4alpha2	0	0	0	2	0	0	0	0
M00159	V$CEBP_01	C/EBPalpha	0	0	0	0	0	0	0	0
M00160	V$SRY_02	SRY	2	4	0	2	0	0	2	2
M00161	V$OCT1_05	POU2F1	7	16	6	0	4	8	0	*14
M00162	V$OCT1_06	POU2F1	*4	8	0	0	2	0	0	4
M00172	V$AP1FJ_Q2	AP-1, c-Fos, c-Jun	0	0	0	0	0	0	0	0
M00173	V$AP1_Q2	AP-1	0	2	2	0	0	0	0	0

M00174	V$AP1_Q6	AP-1	0	0	0	0	0	0	0	0
M00177	V$CREB_Q2	CREB	0	8	0	2	0	0	0	0
M00178	V$CREB_Q4	CREB	0	*8	2	0	0	0	0	0
M00179	V$CREBP1_Q2	ATF-2	0	8	2	2	0	0	0	0
M00183	V$MYB_Q6	c-Myb	1	8	0	0	*4	0	0	0
M00185	V$NFY_Q6	CP1A, CP1C, NF-Y, NF-YA	0	0	0	0	0	0	0	0
M00186	V$SRF_Q6	SRF	1	2	0	0	*6	0	0	0
M00187	V$USF_Q6	USF1	0	0	0	2	0	0	0	0
M00188	V$AP1_Q4	AP-1	2	8	0	0	0	0	0	0
M00189	V$AP2_Q6	AP-2alphaA, AP-2gamma	0	0	0	0	0	0	0	0
M00190	V$CEBP_Q2	C/EBPalpha	0	6	0	0	0	0	0	2
M00191	V$ER_Q6	ER-alpha	0	4	0	0	0	0	0	0
M00192	V$GR_Q6	GR-alpha, GR-beta	0	4	0	0	2	0	0	0
M00193	V$NF1_Q6	NF-1	0	0	0	0	0	0	0	0
M00194	V$NFKB_Q6	NF-kappaB, NF-kappaB1	0	2	0	0	0	0	0	0
M00195	V$OCT1_Q6	POU2F1	6	4	2	2	4	2	2	0
M00201	V$CEBP_C	C/EBPalpha	2	4	0	0	2	2	0	4
M00203	V$GATA_C	GATA-1, GATA-2, GATA-3	*3	4	2	0	0	0	2	2
M00205	V$GRE_C	GR-alpha	2	10	0	2	2	0	0	0
M00206	V$HNF1_C	HNF-1A	6	12	0	0	2	2	0	6
M00208	V$NFKB_C	NF-kappaB, NF-kappaB1, NF-kappaB2	0	2	0	0	0	0	2	0
M00209	V$NFY_C	CP1A, NF-Y, NF-YA	0	*12	0	2	0	0	0	2
M00210	V$OCT_C	Oct-B1, oct-B2, oct-B3, POU2F1, POU2F2, POU2F2(Oct-2.1), POU2F2B, POU2F2C	10	8	6	2	8	8	2	8
M00215	V$SRF_C	SRF	0	0	2	0	0	2	0	2
M00216	V$TATA_C	TBP, TFIID	3	10	4	0	4	0	0	2

M00220	V$SREBP1_01	SREBP-1a, SREBP-1b, SREBP-1c	0	*6	0	*4	0	2	0	0
M00221	V$SREBP1_02	SREBP-1a, SREBP-1b, SREBP-1c	2	0	0	2	0	0	0	0
M00222	V$HAND1E47_01	E47	0	2	2	0	0	0	0	4
M00223	V$STAT_01	STAT1alpha, STAT1beta, STAT2, STAT3, STAT4, STAT6	1	2	2	2	2	2	0	0
M00224	V$STAT1_01	STAT1alpha, STAT1beta	0	2	2	2	0	2	0	0
M00225	V$STAT3_01	STAT3	0	*8	*4	0	0	2	0	2
M00231	V$MEF2_02	MEF-2A	3	*8	0	*4	0	4	0	0
M00232	V$MEF2_03	MEF-2A	6	4	0	0	0	4	0	0
M00233	V$MEF2_04	MEF-2A	*7	12	4	0	2	2	4	4
M00235	V$AHRARNT_01	AhR, Arnt	0	*8	2	0	0	0	0	0
M00236	V$ARNT_01	Arnt	0	2	*2	2	0	0	0	0
M00237	V$AHRARNT_02	AhR, Arnt	0	0	0	0	0	0	0	0
M00243	V$EGR1_01	Egr-1	0	2	0	0	0	2	0	0
M00245	V$EGR3_01	Egr-3	0	4	0	0	0	2	0	0
M00246	V$EGR2_01	Egr-2	0	4	0	2	0	2	0	0
M00248	V$OCT1_07	POU2F1	12	18	4	6	6	*16	4	*14
M00249	V$CHOP_01	C/EBPalpha, CHOP-10	0	4	0	0	2	2	0	2
M00251	V$XBP1_01	XBP-1	0	4	2	0	0	0	0	0
M00252	V$TATA_01	TBP	1	0	0	0	0	0	0	0
M00256	V$NRSF_01	NRSF form 1, NRSF form 2	1	0	2	0	0	2	0	0
M00257	V$RREB1_01	RREB-1	2	6	0	2	0	0	0	2
M00258	V$ISRE_01	ISGF-3	1	20	4	2	2	0	2	2
M00260	V$HLF_01	Hlf	3	10	0	0	2	8	0	2
M00272	V$P53_02	p53	0	0	0	0	0	0	0	0
M00277	V$LMO2COM_01	Lmo2	0	0	0	0	0	0	0	0
M00278	V$LMO2COM_02	Lmo2	5	*18	0	4	2	0	0	0

M00279	V$MIF1_01	MIF-1	0	*8	0	2	2	0	0	0
M00280	V$RFX1_01	RFX1	*4	*8	0	2	2	2	2	0
M00281	V$RFX1_02	RFX1	0	*14	2	0	2	6	0	4
M00284	V$TCF11MAFG_01	LCR-F1	3	*12	*6	2	0	2	2	4
M00285	V$TCF11_01	LCR-F1	1	0	0	0	0	0	0	0
M00287	V$NFY_01	NF-Y	0	0	0	0	0	0	0	0
M00289	V$HFH3_01	FOXI1	2	*8	0	2	2	4	0	0
M00290	V$FREAC2_01	FOXF2	2	*20	0	2	2	4	2	6
M00291	V$FREAC3_01	FOXC1	*7	*20	0	0	4	4	2	4
M00292	V$FREAC4_01	FOXD1	3	*20	0	0	4	0	0	4
M00293	V$FREAC7_01	FOXL1	6	*14	0	4	6	6	2	6
M00302	V$NFAT_Q6	NF-AT1, NF-AT2, NF-AT3, NF-AT4	0	0	0	0	0	0	0	0
M00346	V$GATA1_05	GATA-1	4	18	2	0	4	4	0	6
M00410	V$SOX9_B1	Sox9	0	*12	0	2	0	0	0	4
M00412	V$AREB6_01	AREB6	0	2	0	0	2	0	0	0
M00413	V$AREB6_02	AREB6	0	0	0	0	0	0	0	0
M00414	V$AREB6_03	AREB6	0	4	*4	0	0	0	2	2
M00416	V$CART1_01	Cart-1	*14	10	6	2	10	14	2	10
M00418	V$TGIF_01	TGIF	1	4	0	0	0	0	0	2
M00419	V$MEIS1_01	Meis-1	0	2	0	0	0	0	0	0
M00420	V$MEIS1AHOXA9_01	HOXA9B, Meis-1	9	12	2	0	0	2	4	10
M00421	V$MEIS1BHOXA9_02	HOXA9B, Meis-1	7	4	2	0	0	4	2	8
M00422	V$FOXJ2_01	FOXJ2 (long isoform)	2	2	0	4	6	0	0	6
M00423	V$FOXJ2_02	FOXJ2 (long isoform)	10	14	6	0	14	10	4	16
M00424	V$NKX61_01	Nkx6-1	*11	14	0	0	2	4	0	*14
M00437	V$CHX10_01	Chx10	*10	10	4	4	2	10	0	*10
M00449	V$ZIC2_01	ZIC2	2	*8	0	0	0	0	0	0
M00451	V$NKX3A_01	Nkx3-1	3	4	2	2	2	2	2	6
M00453	V$IRF7_01	IRF-7A	3	*10	0	2	2	6	0	4
M00454	V$MRF2_01	MRF-2	4	*16	0	2	8	8	4	2

M00456	V$FAC1_01	FAC1	*6	*12	0	*6	0	4	0	4
M00457	V$STAT5A_01	STAT5A	2	*14	0	0	*6	2	0	0
M00459	V$STAT5B_01	STAT5B	2	*14	0	2	6	4	0	0
M00460	V$STAT5A_02	STAT5A	0	*10	2	0	2	*6	0	0
M00463	V$POU3F2_01	POU3F2 (N-Oct-5a), POU3F2 (N-Oct-5b)	7	16	2	2	0	8	2	8
M00464	V$POU3F2_02	POU3F2 (N-Oct-5a), POU3F2 (N-Oct-5b)	*25	22	4	4	10	16	6	12
M00472	V$FOXO4_01	FOXO4	2	2	2	0	2	2	0	2
M00474	V$FOXO1_02	FOXO1a	0	30	2	2	6	4	4	6
M00476	V$FOXO4_02	FOXO4	7	*20	4	2	4	4	2	4
M00477	V$FOXO3_01	FOXO3a, FOXO3b	9	*30	2	2	6	10	0	8
M00478	V$CDC5_01	Cdc5	9	16	2	2	4	8	4	6
M00480	V$LUN1_01	LUN-1	0	6	*4	0	0	0	0	0
M00483	V$ATF6_01	ATF6	0	0	0	0	0	2	0	0
M00484	V$NCX_01	NCX	3	*14	0	2	2	4	0	4
M00485	V$NKX22_01	Nkx2-2	5	12	2	2	0	4	0	8
M00490	V$BACH2_01	Bach2	3	4	2	2	0	0	0	4
M00491	V$MAZR_01	MAZR	0	0	0	0	0	0	0	0
M00495	V$BACH1_01	Bach1	2	*14	0	2	2	0	0	6
M00510	V$LHX3_01	LHX3a, LHX3b	5	2	2	0	4	2	2	8
M00512	V$PPARG_01	PPAR-gamma1, PPAR-gamma2	2	4	0	2	2	0	0	0
M00515	V$PPARG_02	PPAR-gamma1, PPAR-gamma2	0	*6	0	0	0	0	0	0
M00516	V$E2F_03	E2F	0	2	0	0	0	0	0	0
M00517	V$AP1_01	AP-1, c-Fos, Fra-1, JunB, JunD	0	4	0	2	0	0	0	2
M00526	V$GCNF_01	GCNF-1, GCNF-2	4	4	0	4	0	0	2	0
M00528	V$PPARG_03	PPAR-gamma1, PPAR-gamma2	0	4	0	0	0	0	0	0

M00532	V$RP58_01	RP58	2	6	0	0	0	2	0	2
M00539	V$ARNT_02	Arnt	0	4	2	0	0	0	0	0
M00615	V$MYCMAX_03	c-Myc, Max	0	*4	0	0	0	0	0	0

HeLa cells

Matrix	Accession Number	Binding factors	MLV	HIV	HIVmIN
M00002	V$E47_01	E47	0	0	0
M00005	V$AP4_01	AP-4	0	0	0
M00006	V$MEF2_01	MEF-2A	12	4	2
M00007	V$ELK1_01	Elk-1	0	0	0
M00017	V$ATF_01	ATF	2	2	0
M00024	V$E2F_01	E2F	12	0	2
M00025	V$ELK1_02	Elk-1	0	0	0
M00026	V$RSRFC4_01	RSRFC4	2	4	4
M00033	V$P300_01	p300	0	0	0
M00034	V$P53_01	p53	4	2	*4
M00037	V$NFE2_01	NF-E2	24	2	0
M00039	V$CREB_01	CREB, deltaCREB	8	0	2
M00040	V$CREBP1_01	ATF-2	*26	0	2
M00041	V$CREBP1CJUN_01	ATF-2, c-Jun	24	0	4
M00045	V$E4BP4_01	E4BP4	*16	4	0
M00050	V$E2F_02	E2F, E2F-1, E2F-2, E2F-3a, E2F-4, E2F-5	4	0	2
M00051	V$NFKAPPAB50_01	NF-kappaB1	0	0	2
M00052	V$NFKAPPAB65_01	RelA	6	2	*4
M00053	V$CREL_01	c-Rel	*6	0	4
M00054	V$NFKAPPAB_01	NF-kappaB, NF-kappaB1, RelA	4	2	0
M00056	V$MYOGNF1_01	NF-1	0	0	0
M00059	V$YY1_01	YY1	14	2	2
M00062	V$IRF1_01	IRF-1	4	6	4

M00065	V$TAL1BETAE47_01	E47, Tal-1beta	4	0	0
M00066	V$TAL1ALPHAE47_01	E47, Tal-1	4	0	0
M00069	V$YY1_02	YY1	*20	0	8
M00070	V$TAL1BETAITF2_01	ITF-2, Tal-1beta	12	0	2
M00071	V$E47_02	E47	0	0	0
M00076	V$GATA2_01	GATA-2	0	0	0
M00077	V$GATA3_01	GATA-3	0	0	0
M00084	V$MZF1_02	MZF-1	0	2	0
M00085	V$ZID_01	ZID	4	0	0
M00095	V$CDP_01	CUTL1	12	0	4
M00096	V$PBX1_01	Pbx1a	0	10	2
M00097	V$PAX6_01	Pax-6	*16	6	2
M00098	V$PAX2_01	Pax-2	0	0	2
M00102	V$CDP_02	CUTL1	14	10	10
M00104	V$CDPCR1_01	CUTL1	2	0	0
M00105	V$CDPCR3_01	CUTL1	4	4	6
M00106	V$CDPCR3HD_01	CUTL1	2	0	2
M00109	V$CEBPB_01	C/EBPbeta	2	2	0
M00113	V$CREB_02	CREB, deltaCREB	0	0	0
M00114	V$TAXCREB_01	CREB, deltaCREB	6	0	0
M00115	V$TAXCREB_02	CREB, deltaCREB	10	0	0
M00116	V$CEBPA_01	C/EBPalpha	2	0	2
M00117	V$CEBPB_02	C/EBPbeta	8	4	2
M00118	V$MYCMAX_01	c-Myc, Max1	0	0	0
M00119	V$MAX_01	Max1	0	0	*2
M00121	V$USF_01	USF1	0	0	2
M00122	V$USF_02	USF1	0	0	0
M00123	V$MYCMAX_02	c-Myc, Max1	0	0	0
M00124	V$PBX1_02	Pbx1a	10	0	4
M00126	V$GATA1_02	GATA-1	14	2	0
M00127	V$GATA1_03	GATA-1	4	2	4

M00128	V$GATA1_04	GATA-1.	6	4	0
M00130	V$FOXD3_01	FOXD3	6	0	2
M00132	V$HNF1_01	HNF-1A	8	6	6
M00133	V$TST1_01	POU3F1	2	8	2
M00134	V$HNF4_01	HNF-4alpha2	4	0	2
M00135	V$OCT1_01	POU2F1	14	2	2
M00136	V$OCT1_02	POU2F1	6	2	8
M00137	V$OCT1_03	POU2F1	2	0	0
M00138	V$OCT1_04	POU2F1	8	0	0
M00143	V$PAX5_01	Pax-5	0	0	0
M00144	V$PAX5_02	Pax-5	0	0	0
M00145	V$BRN2_01	POU3F2	10	6	4
M00146	V$HSF1_01	HSF1 (long)	6	0	0
M00147	V$HSF2_01	HSF2	0	2	0
M00152	V$SRF_01	SRF	8	2	0
M00155	V$ARP1_01	ARP-1	6	0	0
M00156	V$RORA1_01	RORalpha1	2	4	2
M00157	V$RORA2_01	RORalpha2	10	8	2
M00158	V$COUP_01	COUP-TF1, HNF-4alpha2	6	2	2
M00159	V$CEBP_01	C/EBPalpha	0	0	0
M00160	V$SRY_02	SRY	8	8	0
M00161	V$OCT1_05	POU2F1	14	4	4
M00162	V$OCT1_06	POU2F1	8	8	0
M00172	V$AP1FJ_Q2	AP-1, c-Fos, c-Jun	0	2	0
M00173	V$AP1_Q2	AP-1	4	0	2
M00174	V$AP1_Q6	AP-1	0	2	0
M00177	V$CREB_Q2	CREB	6	0	2
M00178	V$CREB_Q4	CREB	8	0	0
M00179	V$CREBP1_Q2	ATF-2	14	0	0
M00183	V$MYB_Q6	c-Myb	0	2	0
M00185	V$NFY_Q6	CP1A, CP1C, NF-Y, NF-YA	0	0	0

M00186	V$SRF_Q6	SRF	4	2	0
M00187	V$USF_Q6	USF1	0	0	0
M00188	V$AP1_Q4	AP-1	4	2	0
M00189	V$AP2_Q6	AP-2alphaA, AP-2gamma	0	0	0
M00190	V$CEBP_Q2	C/EBPalpha	4	2	2
M00191	V$ER_Q6	ER-alpha	0	2	0
M00192	V$GR_Q6	GR-alpha, GR-beta	4	0	0
M00193	V$NF1_Q6	NF-1	2	0	0
M00194	V$NFKB_Q6	NF-kappaB, NF-kappaB1	2	0	0
M00195	V$OCT1_Q6	POU2F1	8	2	2
M00201	V$CEBP_C	C/EBPalpha	8	2	0
M00203	V$GATA_C	GATA-1, GATA-2, GATA-3	6	0	0
M00205	V$GRE_C	GR-alpha	8	2	6
M00206	V$HNF1_C	HNF-1A	6	6	*8
M00208	V$NFKB_C	NF-kappaB, NF-kappaB1, NF-kappaB2	4	4	2
M00209	V$NFY_C	CP1A, NF-Y, NF-YA	2	2	0
M00210	V$OCT_C	Oct-B1, oct-B2, oct-B3, POU2F1, POU2F2, POU2F2(Oct-2.1), POU2F2B, POU2F2C	14	4	10
M00215	V$SRF_C	SRF	*10	0	2
M00216	V$TATA_C	TBP, TFIID	10	6	0
M00220	V$SREBP1_01	SREBP-1a, SREBP-1b, SREBP-1c	4	0	2
M00221	V$SREBP1_02	SREBP-1a, SREBP-1b, SREBP-1c	4	0	0
M00222	V$HAND1E47_01	E47	4	0	4
M00223	V$STAT_01	STAT1alpha, STAT1beta, STAT2, STAT3, STAT4, STAT6	*8	0	2
M00224	V$STAT1_01	STAT1alpha, STAT1beta	4	0	2
M00225	V$STAT3_01	STAT3	*8	0	0
M00231	V$MEF2_02	MEF-2A	*8	*6	2
M00232	V$MEF2_03	MEF-2A	4	2	2
M00233	V$MEF2_04	MEF-2A	12	8	2

M00235	V$AHRARNT_01	AhR, Arnt	4	2	0
M00236	V$ARNT_01	Arnt	0	0	0
M00237	V$AHRARNT_02	AhR, Arnt	2	2	0
M00243	V$EGR1_01	Egr-1	2	0	2
M00245	V$EGR3_01	Egr-3	2	0	0
M00246	V$EGR2_01	Egr-2	2	0	0
M00248	V$OCT1_07	POU2F1	*22	4	16
M00249	V$CHOP_01	C/EBPalpha, CHOP-10	6	2	0
M00251	V$XBP1_01	XBP-1	2	0	2
M00252	V$TATA_01	TBP	*10	0	0
M00256	V$NRSF_01	NRSF form 1, NRSF form 2	6	0	2
M00257	V$RREB1_01	RREB-1	*12	0	2
M00258	V$ISRE_01	ISGF-3	6	4	2
M00260	V$HLF_01	Hlf	4	4	0
M00272	V$P53_02	p53	0	0	0
M00277	V$LMO2COM_01	Lmo2	0	0	0
M00278	V$LMO2COM_02	Lmo2	2	4	0
M00279	V$MIF1_01	MIF-1	*8	2	*4
M00280	V$RFX1_01	RFX1	*10	2	2
M00281	V$RFX1_02	RFX1	2	0	6
M00284	V$TCF11MAFG_01	LCR-F1	*20	6	2
M00285	V$TCF11_01	LCR-F1	0	0	0
M00287	V$NFY_01	NF-Y	0	0	0
M00289	V$HFH3_01	FOXI1	8	6	0
M00290	V$FREAC2_01	FOXF2	*32	*14	10
M00291	V$FREAC3_01	FOXC1	*16	6	8
M00292	V$FREAC4_01	FOXD1	*14	6	6
M00293	V$FREAC7_01	FOXL1	12	8	4
M00302	V$NFAT_Q6	NF-AT1, NF-AT2, NF-AT3, NF-AT4	2	0	0
M00346	V$GATA1_05	GATA-1	16	8	4
M00410	V$SOX9_B1	Sox9	6	0	2

M00412	V$AREB6_01	AREB6	2	0	0
M00413	V$AREB6_02	AREB6	0	0	0
M00414	V$AREB6_03	AREB6	4	2	0
M00416	V$CART1_01	Cart-1	16	2	8
M00418	V$TGIF_01	TGIF	0	0	2
M00419	V$MEIS1_01	Meis-1	0	0	0
M00420	V$MEIS1AHOXA9_01	HOXA9B, Meis-1	10	14	4
M00421	V$MEIS1BHOXA9_02	HOXA9B, Meis-1	6	10	4
M00422	V$FOXJ2_01	FOXJ2 (long isoform)	*16	2	2
M00423	V$FOXJ2_02	FOXJ2 (long isoform)	14	20	12
M00424	V$NKX61_01	Nkx6-1	10	10	0
M00437	V$CHX10_01	Chx10	14	6	2
M00449	V$ZIC2_01	ZIC2	2	4	0
M00451	V$NKX3A_01	Nkx3-1	6	4	2
M00453	V$IRF7_01	IRF-7A	6	6	0
M00454	V$MRF2_01	MRF-2	12	10	2
M00456	V$FAC1_01	FAC1	10	6	2
M00457	V$STAT5A_01	STAT5A	6	2	8
M00459	V$STAT5B_01	STAT5B	*14	4	10
M00460	V$STAT5A_02	STAT5A	*14	0	4
M00463	V$POU3F2_01	POU3F2 (N-Oct-5a), POU3F2 (N-Oct-5b)	8	8	4
M00464	V$POU3F2_02	POU3F2 (N-Oct-5a), POU3F2 (N-Oct-5b)	*38	8	8
M00472	V$FOXO4_01	FOXO4	0	0	2
M00474	V$FOXO1_02	FOXO1a	20	12	2
M00476	V$FOXO4_02	FOXO4	*24	6	2
M00477	V$FOXO3_01	FOXO3a, FOXO3b	28	14	4
M00478	V$CDC5_01	Cdc5	10	10	8
M00480	V$LUN1_01	LUN-1	4	0	2
M00483	V$ATF6_01	ATF6	*8	0	0
M00484	V$NCX_01	NCX	8	2	6
M00485	V$NKX22_01	Nkx2-2	4	4	6

M00490	V$BACH2_01	Bach2	*12	2	4
M00491	V$MAZR_01	MAZR	0	0	0
M00495	V$BACH1_01	Bach1	*34	6	4
M00510	V$LHX3_01	LHX3a, LHX3b	8	0	14
M00512	V$PPARG_01	PPAR-gamma1, PPAR-gamma2	2	0	2
M00515	V$PPARG_02	PPAR-gamma1, PPAR-gamma2	0	0	0
M00516	V$E2F_03	E2F	0	0	0
M00517	V$AP1_01	AP-1, c-Fos, Fra-1, JunB, JunD	*10	0	0
M00526	V$GCNF_01	GCNF-1, GCNF-2	4	8	2
M00528	V$PPARG_03	PPAR-gamma1, PPAR-gamma2	4	0	2
M00532	V$RP58_01	RP58	4	2	2
M00539	V$ARNT_02	Arnt	6	0	2
M00615	V$MYCMAX_03	c-Myc, Max	2	0	0