Charles University

Faculty of Science

Molecular and Cellular Biology, Genetics and Virology

Mgr. Dalibor Miklík

Integration site distribution of expressed proviruses
Distribuce míst integrace exprimovaných provirů

Doctoral thesis

Supervisor: RNDr. Jiří Hejnar, CSc.

Consultant: Mgr. Filip Šenigl, Ph.D.

Laboratory of Viral and Cellular Genetics

Institute of Molecular Genetics of the ASCR, v. v. i.

Prague, 2019

Prohlášení:

Prohlašuji, že jsem závěrečnou práci zpracoval samostatně a že jsem uvedl všechny použité informační zdroje a literaturu. Tato práce ani její podstatná část nebyla předložena k získání jiného nebo stejného akademického titulu.

Declaration:

I hereby certify that I have written this thesis independently and that I have not used other than the cited sources. This thesis has not been submitted for any other degree or purposes.

May 2019, Prague

.....

Mgr. Dalibor Miklík

Preface

Work presented in this thesis follows the work that was started in the 1990's by prof. Jan Svoboda and Jiří Hejnar at the Laboratory of Viral and Cellular Genetics. Their experiments performed with Rous sarcoma virus on rat cells led to the conclusion that the position effect of provirus-surrounding environment affects final expression of the provirus. As time passed, the topic evolved and light was shed on the principles by which the proviral integration site environment affects expression of avian provirus mainly by the work of my supervisor Filip Šenigl. The topic of the integration site effect is still alive and the question: "What is the effect if there's any?" has become more and more pronounced during recent discussions covering the expression activity of retroviruses. Emergence of new publications and new techniques developed to study the effect of proviral environment on proviral expression is a proof of its topicality. With the application of retroviral vectors as tools for delivery and expression of foreign genes or healthy copies of genes damaged by endogenous mutations, the topic of effectivity of expression and possible genotoxic harm of integrated provirus has gained interest of a broader scientific and clinical audience.

The initial work was performed with an avian retrovirus that up to the present has been an important experimental model used in the Laboratory of Viral and Cellular Genetics. Even at the time when the most studied retrovirus is HIV-1 and most used retroviral vectors are those derived from HIV-1 and murine MLV, the avian system is able to bring new insights not only into retroviral biology, but also into as widely studied topic as functionality of human genome chromatin. Using only HIV-1 or MLV, we would probably never be able to define transcriptional start sites as the loci that most strongly support transcription of foreign elements. On the other hand, in this thesis we also show that other retroviruses carry different interesting qualities that shed new light into the topic addressed.

The thesis is based on two already published papers (Senigl et al. 2017, Miklik et al. 2018) as well as on unpublished results that were gained on the evolving path to understanding the association of features identified in the published work with expression of proviruses. Since the data have not yet reached publication, the unpublished results may not be fully conclusive but were introduced into the thesis in the best belief that the data add new information to the story presented in the thesis.

Acknowledgement

Any work is easier when you have a right group of people around. All the people that directly contributed to this work are acknowledged in the text. Here, I would like to thank all the colleagues who helped me to complete the work presented in this thesis. I would like to thank Jiří Hejnar, for giving me the opportunity to work in his laboratory on such an interesting topic and being a nice, cordial and enlightened supervisor; Filip Šenigl for the supervision and for introducing me into the world of retroviral vectors, flow cytometry, for consulting the problems and plans of my projects and letting me develop experiments rising from his own ideas; Miroslav Auxt for teaching me the splinkerette method and introducing me into the magic of after-lunch table football game; "magic" Dana Kučerová for learning me to handle cell cultures and being the one who always had time to help me with anything from cloning to keeping my cells alive, Kateřina Trejbalová who materially and methodologically led me and helped me to enter the world of HIV-related research; Martina Slavková to be involved into my project and helping me to move the thoughts I did not have time to deal with; Daniel Elleder for always having good advice on anything and for helping me with some bioinformatic issues; Tomáš Hron for helping me to find bioinformatic solutions to any of the problems I met for the first time; Helena Fabryová for helping me with anything in the lab and with GAUK administration; Jan Pačes for telling me to "learn MySQL", which rose my self-learning interest in bioinformatics, and to everybody who ever was part of the Laboratory of Viral and Cellular Genetics and helped to keep a good atmosphere in and out the lab. I also want to thank younger students of doctoral programs in our lab that they waited for me to finish the thesis before them. Last but not least, I want to thank my friends outside the lab who supported me in anything I was doing at the time (but still letting me to move the work forward), to my family who supported me even though they had no idea of what I was doing, and I want to thank my lovely wife who motivated me to put the pieces of my work onto this piece of paper.

This study was supported by the Czech Science Foundation (grants No. P502/11/2207, 14-34873S and 15-24776S), Academy of Sciences of the Czech Republic (grant No. AV0Z5050514), Grant Agency of Charles University (GA UK 816216). The work was also institutionally supported by RVO: 68378050.

Abstrakt

Pro zajištění účinné exprese svých genů integrují retroviry provirové kopie svých genomů do genomů infikovaných buněk. Epigenetické procesy však mohou narušit a umlčet expresi integrovaných provirů. Takovéto umlčování sice zpomaluje šíření virové infekce, ale vytváří také reservoir latentních provirů, který v důsledku brání účinné léčbě retrovirových (např. HIV-1) infekcí. Umlčování integrujících se retrovirových vektorů navíc omezuje jejich účinou aplikaci v transgenezi či genové terapii. Cílem této práce je popsat interakce mezi expresí retrovirů a hostitelským (epi)genomikým prostředím v místech integrace provirů.

Pro splnění stanoveného cíle jsme se rozhodli definovat (epi)genomické prostředí provirů, jejichž exprese není zasažena epigenetickým umlčováním. Jako modelové systémy jsme využili odlišné retrovirové vektory odvozené od ptačího sarkomového a leukózového viru (ASLV), myšího leukemického viru (MLV) nebo lidského viru získané imunodeficience typu 1 (HIV-1), jejichž expresní aktivita v lidských buňkách byla sledována. Za účelem popisu charakteristik míst integrace provirů rezistentních vůči umlčování jsme z infikované populace oddělili buňky nesoucí aktivní proviry, identifikovali místa integrace těchto provirů a porovnali charakteristiky takovýh míst s těmi získanými ze směsné, neselektované populace. Pro definici charakteristik přítomných v místě integrace byly aplikovány *in silico* metody. Navzdory značným rozdílům mezi jednotlivými vektory, které byly v práci použity, se nám podařilo identifikovat ty charakteristiky míst integrace, které jsou společné dlouhodobě expresně aktivním provirům použitých vektorů. Takové proviry byly silně asociovány s oblastmi fungujícími jako regulátory exprese – ehancery a promotory aktivních genů. Z našich výsledků vyplývá, že právě oblasti těstně sousedící s místy počátku transkripce jsou těmi, které nejsilněji podporují expresi integrovaných provirů.

V této práci definujeme charakteristiky hostitelského chromatinu přítomného v místech, která jsou permisivní pro expresi retrovirových genů. Předkládané výsledky ukazují, že proviry selektované pro jejich stabilní expresi jsou preferenčně nacházeny v transkibčně aktivním chromatinu poblíž transkribčně regulačních oblastí. Tento fakt by měl být zohledňován při aplikacích retrovirových vektorů.

Abstract

To establish efficient expression of their genes, retroviruses integrate proviral copies into the genomes of the cells they have infected. Epigenetic events, however, silence expression of the integrated proviruses. This silencing protects host cells from harmful viral spread, but also creates a reservoir of latent proviruses that subsequently hinders the cure of retroviral (e.g., HIV-1) infections. Furthermore, the silencing of retrovirus-derived integrative vectors complicates their application in transgenesis and gene therapy. The goal of this thesis is to describe the interaction between retroviral expression and host (epi)genomic environment at the site of proviral integration.

To pursue the goal, we sought to define the (epi)genomic environment of the proviruses, which expression is not affected by the epigenetic silencing. Diverse retroviral vectors derived from avian sarcoma and leukosis virus (ASLV), murine leukemia virus (MLV), and human immunodeficiency virus type 1 (HIV-1) were used as model retroviral systems, and expression stability of the vectors in human cell lines was examined. In order to identify the features unique to integration sites of the active proviruses, we sorted the cells positive for the proviral expression, identified their proviral integration sites, and compared them to proviral integration sites from nonselected populations. *In silico* analytical methods were applied to define the genomic and epigenomic features associated with proviral integration sites. Despite marked differences in the attributes of the vectors used, we identified the features that are common to long-term expressed proviruses. The proviruses were strongly associated with transcription-regulating elements including enhancers and promoters of active genes. We propose that the loci closely associated with transcriptional start sites provide the strongest transcription permissive environment for retroviral expression.

In this thesis, we defined the chromatin environment permissive for the expression of retroviral genes. The results presented here show that the proviruses selected for stable proviral expression tend to be found in transcriptionally active chromatin and closely associated with regulatory sequences. This fact should be considered in approaches utilizing retroviral vectors.

Contents

Abbreviations

AGMRC / agMRC	active gene-matched random control
AIDS	acquired immunodeficiency syndrome
ALV	avian leukosis virus
ASLV	avian sarcoma leukosis virus
BET	bromo- and extraterminal domain
BLAT	BLAST-like alignment tool
bp	base pair
Brd	bromodomain
CA	capsid protein
CA^{W74D}	capsid protein with N74D mutation
CA^{wt}	wild-type capsid protein
CBS	chromatin binding sequence
CCD	catalytic core domain
CPSF6	cleavage and polyadenylation specific factor 6
CTD	C-terminal domain
DNMT	DNA methyl transferase
EGMRC	exact-gene matched random controls
FACS	fluorescence-activated cell sorting
FACT	facilitated chromatin transcription
FIV	feline immunodeficiency virus
FV	foamy virus
GFP	green fluorescence protein
GFP+	GFP-positive
$GFP+^{3dpi}$	GFP-expressing at 3 dpi
$GFP+^{ST}$	Stably expressing GFP
GMRC	gene-matched random control
GRCh37/hg19	February 2009 version of human reference genome assembly
GRCh38/hg38	December 2013 version of human reference genome assembly
H3K36me2/3	histone H3 di-/trimethylation on lysine 36
H3K4me1/2/3	histone H3 mono-/di-/trimethylated on lysine 4
H3K79me2	histone H3 dimethylation on lysine 79
H3K9me2/3	histone H3 di-/trimethylation on lysine 9
HDAC	histone deacetylase
HFV	human foamy virus
HIV-1	human immunodeficiency virus type 1
HMT	histone methyltransferase
HRP-2	hepatoma-derived growth factor related protein 2
HTLV-1	human T-cell leukemia virus type 1
IBD	integrase binding domain
ICTV	International Committee on Taxonomy of Viruses
IN	integrase
IN^{W390A}	integrase with W390A mutation
IN^{wt}	wild-type integrase

iPos	integration position coordinates composed of chromosome, nucleotide and strand
kb	kilo base pair
LAD	lamin-associated domain
LEDGF	lens epithelium-derived growth factor
LTR	long terminal repeat
MA	matrix protein
MLV	murine leukemia virus
MMTV	mouse mammary tumor virus
MPMV	Mason-Pfizer monkey virus
MVV	maedi-visna virus
NC	nucleocapsid
NGS	next-generation sequencing
NLS	nuclear localization signal
NTD	N-terminal domain
PFV	prototype foamy virus
PIC	preintegration complex
PR	protease
RPKM	reads per kilobase per million
RSV	Rous sarcoma virus
RT	reverse transcriptase
SCID	severe combined immunodeficiency
SIV	simian immunodeficiency virus
SU	surface glycoprotein
Tat	transactivator of transcription
tDNA	target DNA
TM	transmembrane protein
TSS	transcriptional start site
TU	transcriptional unit
umMRC	uniquely mapped matched random controls
vDNA	viral DNA
WDSV	Walleye dermal sarcoma virus
wt	wild-type

Literature overview

Introduction into the molecular biology of retroviruses

Retroviruses, or viruses of family *Retroviridae*, are enveloped viruses bearing two copies of genomic RNA in the capsid. After entering a cell, reverse transcription – the process after which retroviruses gained their name – is initiated in the cytoplasm. During the reverse transcription, two copies of the RNA are reversely transcribed into a single molecule of double-stranded viral DNA (vDNA). vDNA carries the whole genomic information of the retrovirus, whose transcription and subsequent translation gives rise to retroviral proteins that are able to assemble into viral particles. After maturation, viral progeny may infect another cell and repeat the replication cycle.

The need of integration of vDNA into the cell genome is unique among viruses. Integration and expression of integrated vDNA, called a provirus, forms the central part of the retroviral replication cycle. Once the integration is completed, the cell becomes permanently infected, with no way of provirus removal. After integration, the only way of stopping the viral spread is either cell death or silencing of proviral expression.

Since this work aims to describe how the process of integration affects subsequent expression of the provirus, the introductory section gives the overview of the facts known about retroviral integration and expression.

Brief taxonomy and model representatives

According to the International Committee on Taxonomy of Viruses (ICTV), family *Retroviridae* is divided into two subfamilies: *Orthoretrovirinae* and *Spumaretrovirinae*. The majority of model viruses come from the *Orthoretrovirinae* subfamily that is divided into six genera: *Alpharetrovirus, Betaretrovirus*, *Gammaretrovirus*, *Deltaretrovirus, Epsilonretrovirus* and *Lentivirus*.

The first retrovirus discovered, Rous sarcoma virus (RSV), is a representative of the *Alpharetrovirus* genus that comprises oncogenic retroviruses of birds. Beside RSV, avian leukosis virus (ALV) and avian sarcoma leukosis virus (ASLV) are used as model viruses of the genera.

The *Betaretrovirus* genus is formed by tumor- and immunodeficiency-causing retroviruses of mammals, of which mouse mammary tumor virus (MMTV) and Mason-Pfizer monkey virus (MPMV) are the best known representatives.

The most abundant genus of the *Retroviridae* family is *Gammaretrovirus*, which is formed by 18 species of exogenous retroviruses and forms the majority of known endogenous retroviruses. Probably the most studied retrovirus of γ-retroviruses is murine leukemia virus (MLV).

The *Deltaretrovirus* genus includes human leukemia-causing retrovirus called T-cell leukemia virus 1 (HTLV-1). Another representative of this genus is bovine leukemia virus (BLV).

The best-known representative of the retroviruses, human immunodeficiency virus 1 (HIV-1), which is the etiologic agent of the pandemic acquired immunodeficiency syndrome (AIDS), is a representative of genus *Lentivirus*. HIV-2 is another human pathogen that emerged independently of HIV-1. Simian immunodeficiency virus (SIV) is an ancestor of HIV infecting apes and a model virus in lentiviral research. Other representatives of lentiviruses are feline immunodeficiency virus (FIV) or maedi-visna virus (MVV) inducing encephalitis and chronic pneumonitis in the sheep.

The last known genera of the *Orthoretrovirinae* subfamily are the *Epsilonretroviridae* genera that contain waterborne piscine retroviruses causing dermal tumors and epidermal hyperplasia. Walleye dermal sarcoma virus (WDSV) is the model ε-retrovirus studied.

The *Spumavirinae* subfamily or foamy viruses (FV) are a group of retroviruses that is different by many aspects in the replication cycle from the *Orthoretrovirinae* subfamily. Spumaviruses were isolated from many different species and were shown to have an ability to infect many tissues *in vitro*. However, no pathology was associated with FVs even though they show the cytopathic effect *in vitro*. The most used laboratory FV is the prototype foamy virus (PFV) formerly known as human foamy virus (HFV). PFV was isolated from a Kenyan man and probably represented zoonotic transmission of simian foamy virus.

The genera of retroviruses significantly differ in replication strategies and molecular mechanisms by which retroviruses accomplish similar goals. The differences in molecular events accompanying integration and expression of retroviruses will be described in the following sections. Attention is focused on three genera that are studied in the Results section: α- (ASLV), γ- (MLV), and lentiviruses (HIV-1). Other retroviral genera are also discussed where relevant.

Genome organization

The retroviral genome present in the virion is comprised of two copies of RNAs. Genomic RNAs are in fact whole genome mRNAs transcribed from proviral DNA. The length of the retroviral genome ranges from about 7 kb up to about 12 kb (Fig. 1A). Both ends of the provirus are made up of tandem repeat sequences called long terminal repeats (LTR). 5' LTR always acts as a promoter, while 3' LTR holds the information about transcriptional termination. The basic gene composition is similar for all retroviruses and comprises three protein-coding genes: *gag*, *pol*, and *env* (as they go from the 5' to 3' end of the retroviral genome, Fig. 1B). Gag and Pol are polyproteins translated from genomic mRNAs and cleaved to particular proteins and peptides by the action of retroviral protease during the virion maturation. The *gag* gene encodes structural components of the viral particle like matrix (MA), capsid (CA), and nucleocapsid (NC). The *pol* gene encodes enzymes important for the retroviral replication cycle like reverse transcriptase (RT), integrase (IN), and protease (PR). The *env* gene encodes surface glycoprotein (SU) and transmembrane protein (TM) that are present in the envelope membrane of the retroviral particle and are responsible for interaction with cellular receptors and entry into the cellular cytoplasm. While Gag and Pol are produced from whole-genome unspliced mRNA, Env can be produced only from spliced mRNA. The genome made of g*ag*, *pol* and *env* genes is characteristic of so-called simple retroviruses such as α- or γ-retroviruses. Other more complex retroviruses like lentiviruses bear more genes called accessory genes, which may play a role in the expression of retroviral genes or act against the host immunity. Acute transforming retroviruses can also bear cellular oncogenes playing a role in tumorigenesis.

Importance of retroviruses in medicine and biotechnology

Retroviruses are important pathogens of human (HIVs and HTLVs) and farm animals like poultry (ALV), sheep (MVV) or cattle (BLV). HTLV-1 is a blood-transmitted retrovirus with highly endemic appearance with estimated 5 – 10 million people being infected globally (Gessain and Cassar 2012). HTLV-1 is considered to be one of the most oncogenic human pathogens causing the fatal adult T-cell leukemia, progressive myelopathy, and other disorders (Tagaya and Gallo 2017). HIV-1 is the causative agent of worldwide epidemics of AIDS. HIV-1 infects CD4+ T-cells causing a decrease of CD4+ T-cell count in the blood of the host and generating the state of immunodeficiency. Application of antiretroviral therapy (ART) decreases the viral load to undetectable levels and restores the number of T-cells in the blood of the patient. However, after withdrawal of ART, a rapid onset of the viral load and the disease is observed. Thus, a latent reservoir resistant to ART and capable of reactivation exists in the infected patients and is the main obstacle to clearing up the infection. Describing the nature and finding the ways to fight the latent reservoir in HIV-1-infected patients is thus a hot topic in the field of retrovirology nowadays.

The natural ability of retroviruses to transport genes and the permanent insertion of the viral genome into the host DNA makes retroviruses an attractive tool for gene transfer technologies. Indeed, retroviral vectors have long been used in gene therapy trials to treat hereditary diseases (reviewed by Galy (2017)).

Still, improvement in the design of retroviral vectors is required as retroviral integration is not controlled as regards site-specific integration and may possess the risk of genotoxicity, while the expression of vectors needs to be as stable as possible for efficient transduction and keeping life-long expression of the transgene.

The detailed knowledge of the processes guiding the integration site selection and expression stability of retroviruses is thus important for both development of strategies for the treatment of retrovirus-caused diseases and advances in gene transfer technologies.

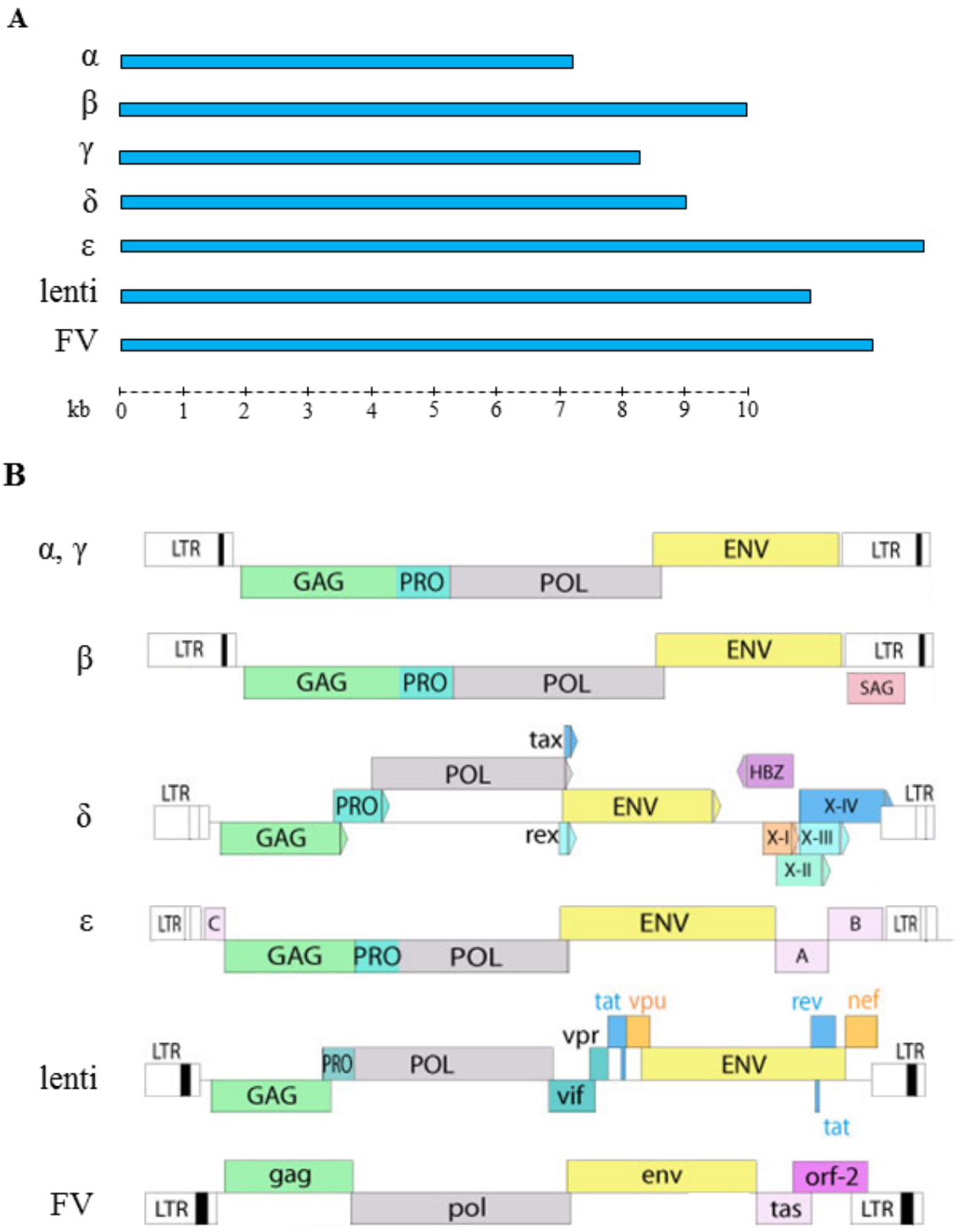

Figure 1. Graphical representation of retroviral genomes. **A.** Graphical representation of proviral gene composition of retroviruses. **B.** Graphical representation of the length of retroviral genomes in kb. Source: ViralZone: www.expasy.org/viralzone, SIB Swiss Institute of Bioinformatics.

Integration

Integrase I: structure and enzymatic activity

Integration of vDNA into target genomic DNA (tDNA) is facilitated by the enzyme called integrase (IN). The current state of knowledge in the field of structural and functional features of IN has been well reviewed by Grawenhoff and Engelman (2017) and Engelman and Cherepanov (2017).

The main function of IN is covalent insertion of vDNA into tDNA. Integration of vDNA is performed by two enzymatic activities of IN: 3'-processing and strand transfer reaction (Fig. 2A). During 3'-processing, IN removes nucleotides at the 3' end of vDNA, exposing the hydroxyl group of invariant CA dinucleotide. In the strand transfer reaction, IN facilitates ligation of processed 3' ends of vDNA to the target site on tDNA. The entire process leading to covalently linked vDNA and tDNA is called concerted integration and can be accomplished by purified IN *in vitro*. Cellular machinery then completes the process of integration by removing 5' overhangs, filling the gaps and ligating 5'ends to the tDNA. As a result, a provirus surrounded by short target site duplications of 4 to 6 bp is formed.

IN is a recombinase encoded by the *pol* gene and translated from genomic RNA as a part of the Gag-Pol polyprotein. IN consists of three structural domains: N-terminal domain (NTD), catalytic core domain (CCD), and SH3-like C-terminal domain (CTD) with DNA binding activity. NTD is formed by a helix-turn-helix motif, binds two pairs of Zn^{2+}, and is responsible for recognition of the long terminal repeats (LTRs) at the ends of vDNA and multimerization of IN molecules (Khan et al. 1991, Burke et al. 1992, Bushman et al. 1993, Zheng et al. 1996). CCD consists of conserved D, DX_{35} E catalytic triad that together with Mg^{2+} ions catalyzes the integration reaction (Drelich et al. 1992, Engelman and Craigie 1992, Kulkosky et al. 1992, Bushman et al. 1993). CTD adopts the SH3 fold and is involved in tDNA binding (Engelman et al. 1994) and IN multimerization (Andrake and Skalka 1995, Jenkins et al. 1996). IN of some retroviral genera, namely spuma- , ε-, and γ-retroviruses, contain an additional domain preceding the NTD - NTD extension domain (NED) that is involved in vDNA binding (Hare et al. 2010, Ballandras-Colas et al. 2016).

Until 2010, the full structure of IN was missIn 2010, the Cherepanov group published the structures of PFV intasome consisting of tetrameric IN and vDNA (Hare et al. 2010, Maertens et al. 2010). Later, the same group published the PFV intasome *in flagrante delicto* – sequential crystallographic mapping of 3'-processing and strand transfer reactions (Hare et al. 2012). However, as shown later, the tetrameric structure observed with PFV is not universal and the intasome architecture varies among retroviral genera (Fig. 2B). Since 2016, the intasome structures of retroviruses from various genera were revealed. MMTV and RSV were proved to build octameric intasomes (Ballandras-Colas et al. 2016, Yin et al. 2016). The revolution in cryo-electron microscopy enabled structural biologists to enlighten the structure of lentiviral intasomes (Ballandras-Colas et al. 2017, Passos et al. 2017). MVV and HIV-1 intasomes are formed by a core tetramer, but higher-order multimerization seems to be required for the intasome function as a plethora of IN multimer species were described. The IN tetramer observed in PFV intasomes seems to be a conserved intasome core for all intasome quaternary structures observed. The ability of IN to build up higher multimers seems to be dependent on the length of the CCD-CTD linker. PFV IN includes a CCD-CTD linker of 49 amino acids; the analogous parts of MMTV and RSV INs are composed of 8 amino acids. CCD-CTD linkers of MVV and HIV-1 are of intermediate length of 19-22 amino acids. So far, no intasome structures for ε- and γ-retroviruses have been described. The lengths of IN CCD-CTD linker being 55 and 61 amino acids suggest that γ- and ε-retrovirus intasomes might be assembled by tetramers, as observed in the case of PFV. However, mutation analysis conducted in RSV IN suggests that the C-terminal tail of IN is responsible for the formation of octamers *in vitro* and that truncated INs forming only the tetrameric intasome core keeps the enzymatic activity at the level of octameric intasome (Pandey et al. 2017). Thus, the role of multimeric intasomes *in vivo* remains to be elucidated.

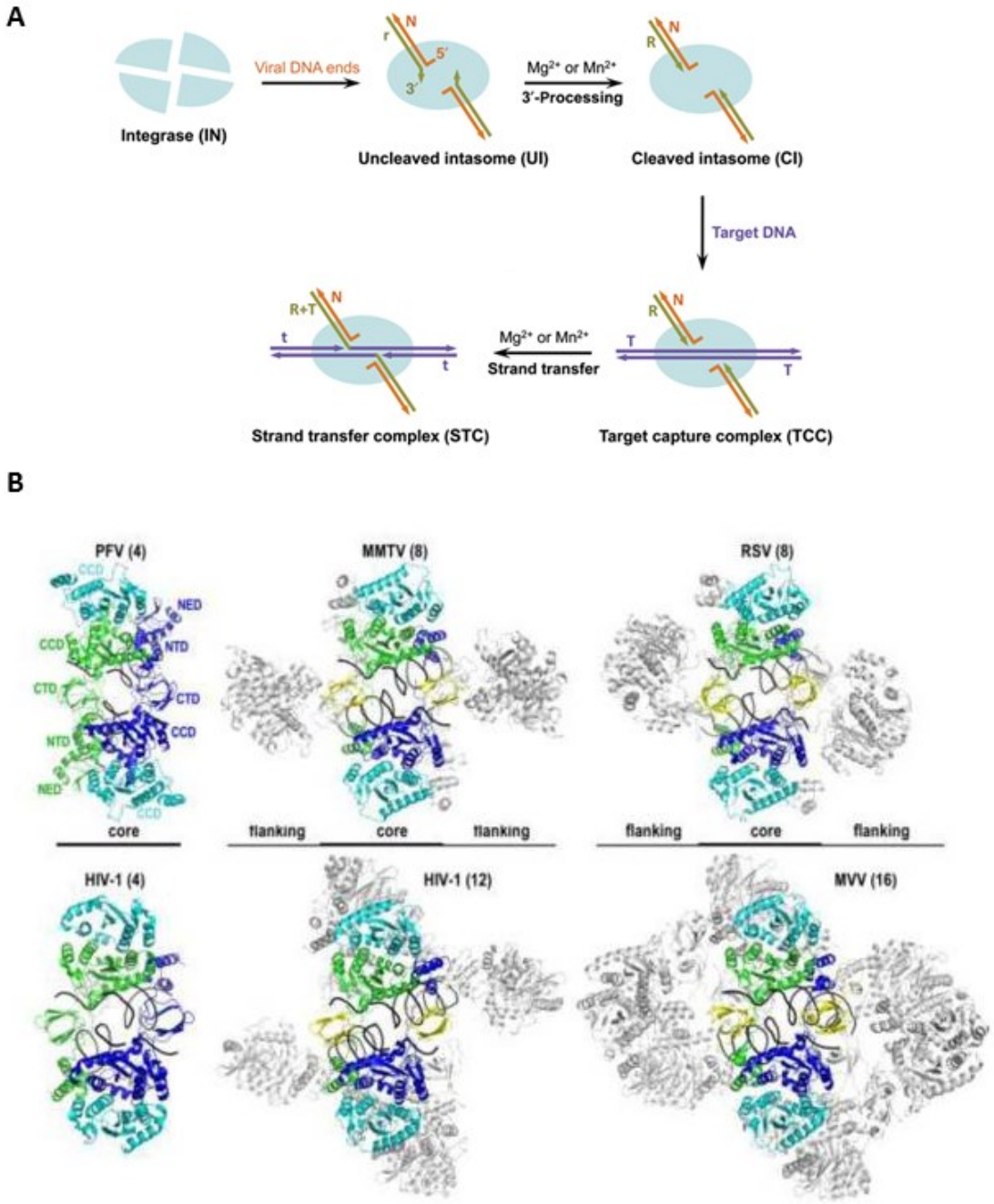

Figure 2. Enzymatic and structural features of retroviral integration. **A.** Step-wise depiction of the process of orchestrated integration. **B.** Structures of core intasomes of different retroviruses. Core parts are depicted as colored structures. The number in parentheses indicates the number of integrase molecules displayed. PFV – prototype foamy virus, MMTV – mouse mammary tumor virus, RSV – Rous sarcoma virus, HIV-1 – human immunodeficiency virus type 1, MVV – maedi-visna virus. Source: **A**. Hare et al. (2012), **B**. Engelman and Cherepanov (2017).

Integrase II: non-enzymatic functions

While the main function of IN is demonstrated by the ability to perform concerted integration, IN was shown to significantly influence or directly mediate many steps in the retroviral replication cycle facilitated by a number of interactions with viral or cellular components.

As early as in the released virion, IN was proposed to mediate appropriate virion maturation through binding viral RNA (Kessl et al. 2016). This interaction is supposed to ensure proper positioning of viral RNA in the virion and thus proper maturation of the viral particle. Although reverse transcriptase is capable of reverse transcription *in vitro*, IN was shown to play a role in reverse transcription *in vivo* through direct interaction with reverse transcriptase (Wu et al. 1999, Lai et al. 2001, Zhu et al. 2004, Dobard et al. 2007, Jurado et al. 2013, Tekeste et al. 2015).

On the outer surface of intasome, IN interacts with a number of proteins. The complex of intasome and other proteins is called preintegration complex (PIC). PIC composition influences diverse steps in the retroviral replication cycle spanning from nuclear import of PIC to PIC tethering to chromatin.

Nuclear entry

At first, PIC needs to transport vDNA to the nucleoplasm. Because the size of PIC is above the limit for passive diffusion through the nuclear pores, PIC can reach host chromosomes by active transport through the nuclear pores or by mitosis-dependent disassembly of the nuclear envelope. Both cases were observed among the retroviruses.

MLV infection and integration is mitosis dependent (Fischinger et al. 1975, Roe et al. 1993, Lewis and Emerman 1994). This is strengthened by the observation that p12, the cleavage product of the Gag protein, plays a role in attachment of PIC to the histones of mitotic chromosomes and is phosphorylated in mitosis (Elis et al. 2012, Brzezinski et al. 2016, Wanaguru et al. 2018).

RSV was first observed to infect only proliferating cells (Humphries and Temin 1972, Humphries and Temin 1974). However, in synchronized cell culture, RSV proviruses were detected during S-phase before mitosis (Humphries et al. 1981), the nuclear localization signal (NLS) of RSV IN was identified (Kukolj et al. 1997, Kukolj et al. 1998), and the ability of RSV to transduce interphase cells was demonstrated (Hatziioannou and Goff 2001, Katz et al. 2002). Even though it was confirmed that active transport of ASV PIC is executed by a soluble cellular factor (Andrake et al. 2008). This factor or group of factors are yet to be discovered. Although RSV is able to enter the interphase nucleus, the expression of RSV in the cell-cycle arrested cells is decreased compared to cycling cells, which might relate to a limited effectivity with which the PIC is transported through the nuclear pores (Hatziioannou and Goff 2001).

From the very first experiments with lentiviruses, there was no doubt about their ability to infect non-dividing cells. HIV-1 was shown to productively infect terminally differentiated non-dividing cells such as macrophages and cell-cycle arrested cells (Bukrinsky et al. 1992, Lewis et al. 1992, Li et al. 1993, Lewis and Emerman 1994). Contrary to RSV, HIV-1 IN was first reported to lack NLS and being unable to drive active transport through nuclear pores (Kukolj et al. 1997). Nevertheless, functional NLSs were found in the MA protein of HIV-1, which is a structural building block of the HIV-1 capsid, and accessory viral protein Vpr (Bukrinsky et al. 1993, Heinzinger et al. 1994, von Schwedler et al. 1994). Finally, also IN was described to contain NLS and being capable of nuclear import (Gallay et al. 1997, Bouyac-Bertoia et al. 2001). vDNA was also suggested to play a role in translocation through nuclear pores (Zennou et al. 2000). Despite the plethora of factors that were identified to drive the PIC nuclear import, capsid protein (CA) is nowadays considered to be the major lentiviral factor involved in the nuclear import of PIC (Yamashita and Emerman 2004, Yamashita et al. 2007). While the specific pathway of HIV-1 PIC nuclear import remains to be elucidated, CA is known to facilitate the import through direct interaction with Nup153 – protein forming nuclear part of the nuclear pore basket (Buffone et al. 2018).

Another virus that was demonstrated to have the ability of entering the interphase cell nucleus is MMTV. Konstantoulas and Indik (2014) showed that an MMTV-derived vector transduces the cell-cycle arrested cells with the efficiency comparable to HIV-1 yet using a different path than the one known for HIV-1. The viral effector responsible for nuclear import and the path used by MMTV yet remains to be described.

The ability of spumaviruses to enter the nucleus of non-dividing cells seems to be limited, as the efficiency of transduction of PFV-derived vectors is clearly dependent on cycling of host cells (Patton et al. 2004, Trobridge and Russell 2004). Even though some transduction was observed in cell-cycle arrested culture, the transduction is attributed to the cells that are not arrested in cycling (Trobridge and Russell 2004). NLS was found in both IN (Hossain et al. 2014) and Gag (Mullers et al. 2011). However, in accordance with the observation made by Trobridge and Russell (2004), PFV Gag was observed to accumulate on centromeres during interphase and accessing the nucleus during mitosis via the action of the Gag chromatin-binding domain (Tobaly-Tapiero et al. 2008). The literature thus suggests that despite containing NLS, PFV PIC is not able to adopt the nuclear import machinery. However, PIC seems to be stable in arrested cells, with the ability to transduce cells after releasing the cell-cycle break.

Chromatin tethering

After reaching the host cell chromosomes, PIC needs to attach to the chromatin and subsequently tDNA, where integration can be performed. As shown in recent years, IN alone is not capable of reaching the final destination by itself and needs the help of viral or cellular proteins.

The first factor discovered to maintain the function of the PIC tethering factor was lens epithelium-derived growth factor / p75 (LEDGF/p75) interacting with IN of HIV-1 (Cherepanov et al. 2003). LEDGF/p75 was immediately shown to enhance the activity of IN and maintain the attachment of IN to chromosomes (Maertens et al. 2003, Llano et al. 2004, Emiliani et al. 2005). Moreover, IN-LEDGF/p75 interaction is conserved and specific to lentiviruses (Cherepanov 2007). LEDGF/p75 is encoded by the *PSIP1* gene and is a ubiquitously expressed transcriptional coactivator member of the hepatoma-derived growth factor family. LEDGF/p75 bears the ability to bind modified histones, specifically the histone H3 di- and trimethylated on lysine 36 (H3K36me2/3), together with nucleosomal DNA, using the N-terminal part of the protein containing the conserved PWWP domain and AT hook (Llano et al. 2006, Eidahl et al. 2013, van Nuland et al. 2013). On the C-terminal part, LEDGF/p75 interacts with CCD interphase of the IN dimer through the integrase binding domain (IBD) (Cherepanov et al. 2005, Vanegas et al. 2005, Busschots et al. 2007). When the interaction with LEDGF/p75 is abolished either by single amino acid mutations (Q168A in Emiliani et al. (2005)) or by knocking-down the LEDGF/p75 expression, the integration and thus proviral expression is disrupted. However, there are other cellular factors that may potentially substitute for LEDGF/p75 in lentiviral integration, such as LEDGF/p52 that is a splice isoform of the *PSIP1* gene or hepatoma-derived growth factor related protein 2 (HRP-2) that contains both IBD and PWWP domains. Still, LEDGF/p52 lacks IBD and therefore cannot interact with lentiviral IN (Maertens et al. 2003, Shun et al. 2007). HRP-2 is able to interact with HIV-1 IN (Cherepanov et al. 2004) and in the context of LEDGF/p75 knock-out may drive residual replication of HIV-1 (Schrijvers et al. 2012). Although HIV-1 laboratory-adapted strains can replicate in LEDGF/p75 knock-out cells, the replication is highly delayed with even more pronounced replication defect in clinical isolates of HIV-1 (Schrijvers et al. 2012). Moreover, selection for inhibition-resistant mutants corroborate the exclusive role of LEDGF/p75 in the tethering of HIV-1 PIC to the chromatin (Hombrouck et al. 2007). This makes LEDGF/p75 the cellular master regulator of lentiviral integration. The idea of blocking the lentiviral replication cycle through LEDGFP/p75-IN interaction disruption led to the development of small molecules acting as non-catalytic inhibitors of the integration called LEDGINs (Christ et al. 2010). Downstream of the LEDGF/p75-mediated chromatin tethering, additional factors may influence the HIV-1 integration effectivity such as facilitated chromatin transcription (FACT) complex nucleosome remodeling (Matysiak et al. 2017) or IN CTD attachment to the amino-terminal tail of nucleosomes (Benleulmi et al. 2017).

MLV cannot infect non-dividing cells due to the lack of ability to enter the interphase nucleus of the host cell. In order to effectively target the cellular nucleus, MLV needs to tether to metaphase chromosomes. The factor identified to fulfil this function is p12 – cleavage product of viral Gag protein that is the constituent of MLV PIC (Prizan-Ravid et al. 2010, Elis et al. 2012, Wight et al. 2012). p12 comes to the cell as a part of the viral core and stabilizes the capsid in early phase of infection by direct interaction of p12 NTD with CA (Wight et al. 2014). In mitosis, p12 is phosphorylated and activated to bind nucleosomes on metaphase chromosomes with CTD while preserving the interaction with the capsid by its CTD (Brzezinski et al. 2016, Brzezinski et al. 2016, Wanaguru et al. 2018). In postmitotic nucleus, p12 dissociates from chromatin (Elis et al. 2012) and PIC-chromatin association is maintained by the interaction of PIC with the cellular bromo- and extraterminal domain (BET) proteins (Borrenberghs et al. 2018). Interaction of PIC with bromodomain (Brd) proteins 2, 3 and 4 is facilitated directly by the unstructured C-terminal tail of MLV IN and CTD of BET proteins (De Rijck et al. 2013, Gupta et al. 2013, Sharma et al. 2013). By its N-terminal bromodomain, BET proteins bind to acetylated histones, thus presenting the chromatin biding platform to the PIC. *In vitro* IN-BET protein interaction enhances IN catalytic activity. *In vivo* IN-BET-chromatin interaction promotes post-mitotic changes in IN oligomerization (Borrenberghs et al. 2018).

The Gag product was identified as chromatin tether also in the case of PFV integration (Tobaly-Tapiero et al. 2008, Mullers et al. 2011, Lesbats et al. 2017). PFV Gag tethers to chromatin by direct interaction of its chromatin-binding sequence (CBS) at the C-terminal region with nucleosomes.

ASLV IN interacts with the heterodimeric FACT complex (Winans et al. 2017). The FACT complex acts as a general histone chaperone complex, and in contrast to HIV-1 IN-FACT interaction, interaction of the FACT complex with ASLV IN shows to be critical for the ASLV chromatin tethering and integration.

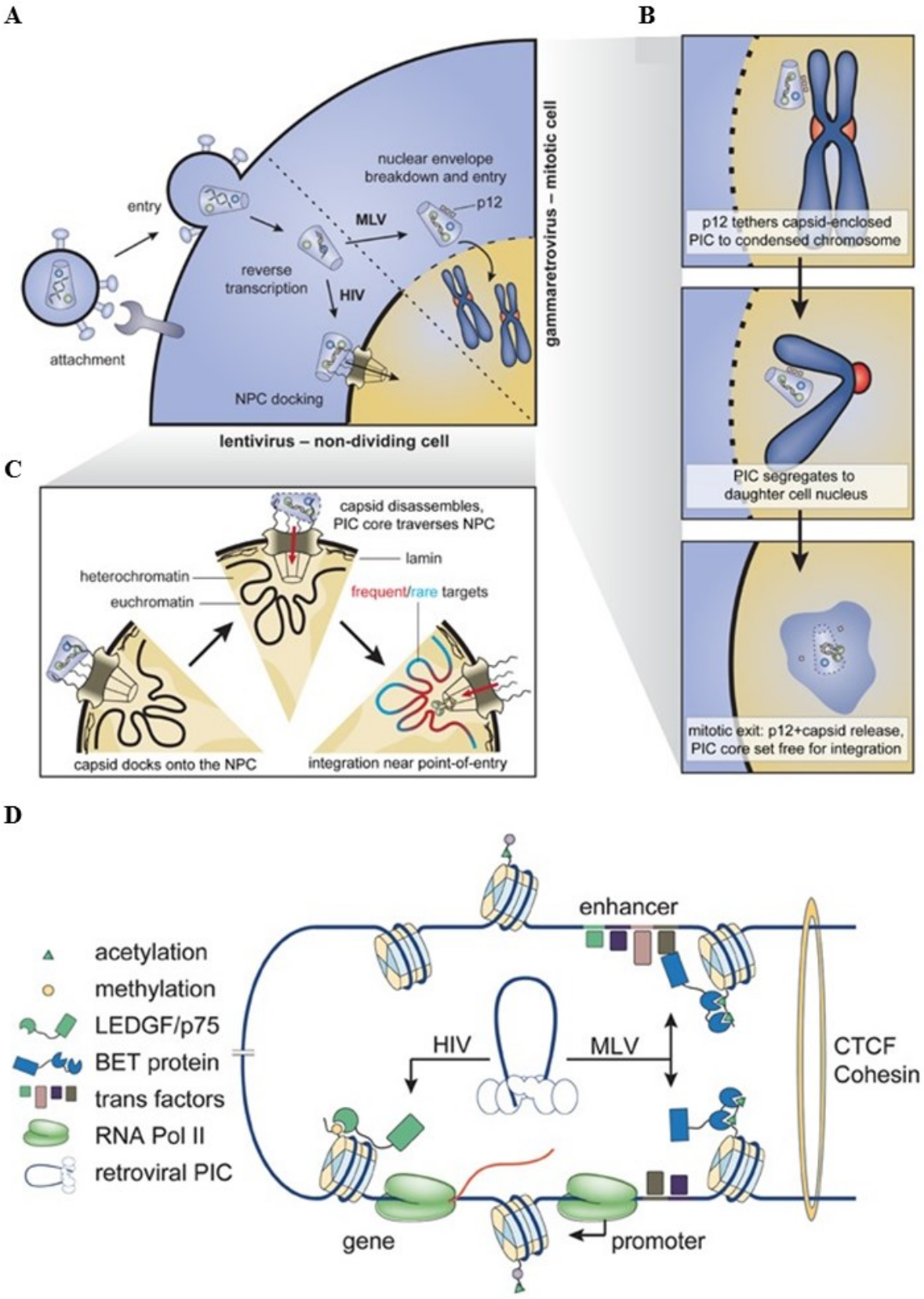

Figure 3. Nucleus entry and chromatin tethering of PIC. **A.** Distinct pathways used by HIV and MLV to enter the host nucleus. HIV is actively transported through nuclear pores, while MLV encounters host chromatin after cell-cycle nuclear breakdown. **B.** MLV PIC tethers to host cell chromatin through the binding of p12 protein of MLV. **C.** Active transport of HIV PIC through the nuclear pore. HIV targets for integration may be represented

by the chromatin regions placed closely to the nuclear pores. **D.** Host cell, IN-binding factors function as tethers to the sites of specific chromatin environment. HIV is targeted to H3K36me3 sites through interaction of H3K36me3-binding protein LEDGF/p75. MLV IN interacts with proteins of the BET family, which bind to acetylated histones. The picture was adopted from Demeulemeester et al. (2015).

Integration site selection

The cellular genome is not homogenous but is bound by a plethora of proteins marked by a number of posttranslational modifications that together form so-called chromatin. Functionally, the genome can be divided into parts that play different roles in the genome structure or gene activity (Ernst and Kellis 2010). Chromatin can be basically differentiated into two categories of active and repressed chromatin. Even before techniques allowed differentiation of the features defining diverse chromatin states, it was uncovered that the distribution of proviruses is not random and differs among retroviral genera. The boom of uncovering the differential integration preference started after the release of human genome assembly (Lander et al. 2001, Venter et al. 2001) and the accessibility of next-generation sequencing techniques. Retroviral integration distribution is genome-wide and is the outcome of the specificity of chromatin tethers that retroviruses hitchhike on their way to the chromatin.

HIV-1 was the first retrovirus announced to integrate in the nonrandom way, preferring active genes as a target (Elleder et al. 2002, Schroder et al. 2002). Targeting of transcriptionally active parts of the genome was shown to be characteristic and specific to lentiviruses (Hematti et al. 2004, Mitchell et al. 2004, Crise et al. 2005, Kang et al. 2006, MacNeil et al. 2006). The preference of lentiviruses for active genes is robust, as the gene targeting was observed to be independent of the type of infected cell line, tissue (Liu et al. 2006), or the route of entry (Barr et al. 2006). The same preference was also observed in cycling cells as well as differentiated, arrested, or quiescent cells (Barr et al. 2006, Ciuffi et al. 2006, Vatakis et al. 2009). Using epigenomic data, HIV-1 integration showed to be closely associated with sites of transcriptionally active chromatin marked by acetylation of histones H3 and H4 and H3K4 methylations (Wang et al. 2007, Santoni et al. 2010). Immediately after the discovery of HIV-1 IN-LEDGF/p75 interaction, the role of LEDGF/p75 in gene targeting of HIV-1 was proposed (Ciuffi et al. 2005). The genomic distribution of chromatin-associated LEDGF/p75 indeed corresponds to the areas downstream from active promoters (De Rijck et al. 2010), and the disruption of LEDGF/p75 expression or modification of its chromatin specificity changes the integration profile more to intergenic regions (Marshall et al. 2007, Shun et al. 2007, Ferris et al. 2010, Gijsbers et al. 2010). Moreover, disruption of the IN-LEDGF/p75 interaction by small inhibitors also alters the profile of proviral integration sites (Feng et al. 2016, Vranckx et al. 2016). LEDGF/p75, however, is not the sole factor playing a role in the lentiviral integration site selection. Interacting with HIV-1 CA protein, cleavage and polyadenylation specific factor 6 (CPSF6), subunit of the cleavage factor complex playing a role in pre-mRNA 3'-end processing, was identified to participate in the chromatin targeting of PIC. In the context of HIV-1 infection, CPSF6 was at first identified as antiviral protein, as the retention of CPSF6 stabilizes the capsid and blocks the nuclear entry through the nuclear pores (Lee et al. 2010, Lee et al. 2012, De Iaco et al. 2013). In recent research, CPSF6 was, however, identified to support nuclear import of PIC via the interaction with CA and to influence the nucleoplasm trafficking of PIC (Chin et al. 2015). Moreover, CPSF6 was proposed to target HIV-1 integration into gene-rich regions, transcriptionally active intron-dense genes (Rasheedi et al. 2016, Sowd et al. 2016). Sowd et al. (2016) proposed the model where CPSF6 drives integration to the transcriptionally active chromatin, while LEDGF/p75 conducts the distribution in gene bodies, preferentially away from transcriptional start sites (TSSs, +1 nucleotide of RNA). CPSF6 was also proposed to target the integration of HIV-1 into the genes that are transcriptionally active upon activation of target T-cells (Zhyvoloup et al. 2017). The role of CPSF6 and LEDGF/p75 in the integration was also examined in the context of 3D nuclear space. HIV-1 was observed to naturally prefer euchromatin at the periphery of the nucleus (Albanese et al. 2008, Di Primio et al. 2013, Quercioli et al. 2016). The role of cellular factors, however, seems to be controversial. While

Marini et al. (2015) attribute nuclear periphery and out-of-lamin-associated domains (LADs) chromatin targeting to LEDGF/p75 and nucleporin Nup153, Achuthan et al. (2018) not only show that nuclear periphery is not the preferred area for integration, but also point out the role of CPSF6 as a factor guiding the PIC of HIV-1 deeper to the nucleoplasm and away from the repressive environment of LADs. Di Primio et al. (2013) and Quercioli et al. (2016) observed something of both when they described peripheral localization of retroviral DNA that was preserved when the infection was performed in cells knocked-down for LEDGF/p75 but was lost in cells with chimeric heterochromatin-targeting LEDGF/p75. Another contradictory observation of Di Primio et al. (2013) was that peripheral distribution observed at 48 hours after infection was lost in 13 days after infection. In any case, all studies presented nonrandom distribution of lentiviral integration that was affected by LEDGF/p75 or CPSF6. The exact role and interplay of LEDGF/p75 and CPSF6, however, remains to be elucidated.

In vivo genome targeting of MLV integration came into spotlight after application of MLV-derived vectors in gene therapy of children suffering from severe combined immunodeficiency (SCID). In otherwise successful therapy (Hacein-Bey-Abina et al. 2002), two out of 10 treated children developed uncontrolled clonal T-cell expansion. In these patients, integration into the proximity of the *LMO2* gene promoter leading to aberrant expression of the protooncogene was observed (Hacein-Bey-Abina et al. 2003). MLV was indeed shown to preferentially target active TSSs (Wu et al. 2003, Hematti et al. 2004, Mitchell et al. 2004, Berry et al. 2006). Targeting of active promoters, however, explained only a small portion of MLV integrations. Taking advantage of the high number of integration sites and progress in epigenomics helped to discover that the major targets of MLV integration are enhancers (De Ravin et al. 2014, LaFave et al. 2014). Interestingly, MLV integration sites covered only about 2 % of the human genome. BET proteins interacting with MLV IN were identified as factors responsible for this striking preference (De Rijck et al. 2013, Gupta et al. 2013, Sharma et al. 2013). This integration site selection can be not only disrupted by disturbing the IN-BET proteins interaction, but the vulnerability of γ-retroviral integrase to modifications allows retargeting the integration to the sites of the desired environment (El Ashkar et al. 2014, El Ashkar et al. 2017, Nam et al. 2019).

Retroviruses of other genera of which the global integration pattern was investigated do not show any striking patterns as observed for lentiviruses and γ-retroviruses. PFV was shown to preferentially avoid active genes and integrate into the vicinity of TSSs and CpG islands, yet with lower frequency than MLV (Trobridge et al. 2006, Lesbats et al. 2017). Interestingly, when PFV Gag interaction with nucleosomes was disrupted, PFV showed a striking preference for centromeric regions (Lesbats et al. 2017). ASLV was described to integrate randomly, with mild non-significant accumulation of proviruses in the gene bodies. MMTV and HTLV-1 integrate randomly without any preference observed (Derse et al. 2007, Konstantoulas and Indik 2014).

Unintegrated genomes

After nuclear entry, not all vDNAs end up as integrated proviruses. In fact, proviruses, whose number peaks at about 48 h after infection, comprise about 20 % of total vDNA present in infected cells (Butler et al. 2001, Brussel and Sonigo 2003, Munir et al. 2013). The vDNA species that fail to integrate into the host genome thus form the dominant pool of vDNA at the early times after retroviral infection. Although PIC-incorporated vDNA enter the nucleus in the form of linear molecule, circular species of retroviral DNA were observed in the nucleus of infected cells (Guntaka et al. 1976). The circular forms of vDNA were observed to contain junctions of 2-LTR or single LTR sequences, pointing to the different paths involved in the genesis of circular vDNA (Shoemaker et al. 1980). In fact, 2- and 1-LTR circles are very abundant forms of unintegrated vDNA species (Munir et al. 2013) and are preferentially formed in the cell nucleus by the action of DNA repair host factors that cause ligation of LTR ends or DNA recombination (Kilzer et al. 2003).

Circular vDNA species were thought to be dead-end byproducts of retroviral infection usable just for the detection of successful nuclear entry of the PIC. However, some observations highlight the role of

the unintegrated vDNA in retroviral infection. 2-LTR circles were suggested to give rise to new integrated proviruses after the removal of strand-transfer inhibitors of IN (Thierry et al. 2015), probably thanks to the ability of the IN to cleave the canonical sequence of 2-LTR junctions (Delelis et al. 2005, Delelis et al. 2007, Zhang et al. 2014), forming a new linear substrate for the integration. Unintegrated vDNA was also shown to be transcribed and contribute to the expression of retroviral genes (see the Expression from unintegrated genomes chapter).

Expression of retroviral genome

Expression of proviral genes can be regulated at both the transcriptional and posttranscriptional level. Transcription of the provirus is driven by a retroviral promoter present in 5' LTR. After integration, the provirus becomes part of the cellular genome and is treated like one of the host genes. Retroviruses possess their own promoters with cis-acting enhancers that are bound by cellular transcription factors to activate or downregulate PolII-dependent transcription. Retroviruses can also encode their own insulators or regulators of transcription. At the posttranscriptional level, the splicing and export of retroviral mRNA can be regulated. After transcription, the primary transcript can be multiply spliced to produce Env or accessory proteins. For instance, HIV-1 was reported to produce over 50 different functionally relevant splice variants during the infection (Emery et al. 2017), and aberrant splicing events accompany the crosspecies restriction of RSV in mammalian cells (Lounkova et al. 2014). For the export of mRNA, retroviruses use the cellular export machinery to transport the mRNA to the cytoplasm. However, some complex retroviruses such as β-, δ-, lentiviruses use the accessory proteins encoded by their genome to regulate the export of full-length genomic mRNAs into the cytoplasm (Felber et al. 1989, Malim et al. 1989, Toyoshima et al. 1990, Indik et al. 2005, Mertz et al. 2005), where the mRNAs are translated.

Regulation of transcription from retroviral promoter

Most of the regulation of gene expression happens in the stage of transcription initiation. The provirus promoter is located in 5' LTR (Yamamoto et al. 1980, Fuhrman et al. 1981). LTR is divided into three parts: U3, R, and U5. The proviral TSS is found at the edge of U3 and R segments. Retroviral promoters belong to the group of TATA-box promoters in which the TATA-box is located upstream to the TSS, thus in the U3 segment. The U3 part of LTR is also the site of proviral enhancer, where additional transcription factors are bound.

RSV and ALV bear quite simple LTRs that have the general ability to drive strong cell type-independent transcription (Gorman et al. 1982). The enhancer segments are mainly formed by the CCAAT box sequences (Greuel et al. 1990), but other sequences important for the enhancer activity of the U3 segment were described (Zachow and Conklin 1992). Despite the general activity of RSV enhancers as a whole and the binding of generally expressed enhancer factors to the U3 part of RSV LTR (Houtz and Conklin 1996), particular enhancer segments were shown to be specifically important in different cell types such as B cells or fibroblasts (Curristin et al. 1997, Swamynathan et al. 1997).

MLV is, similarly as RSV, a simple retrovirus in which the LTR-driven transcription is not cell type restricted. However, the ability of different MLVs to induce different types of disease was assigned to the U3 region of MLV LTR (Chatis et al. 1983, Lenz et al. 1984, Li et al. 1987), highlighting the cell-type-specific activity of different enhancers. Compared to RSV, the U3 part of MLV LTR is longer, containing direct repeats and GC-rich sequences that are bound by ubiquitously expressed transcription factor Sp1 (Baum et al. 1997, Wahlers et al. 2002). Interestingly, the presence of a GC-rich Sp1-binding site in MLV LTR correlates with the high expression of retroviral genes.

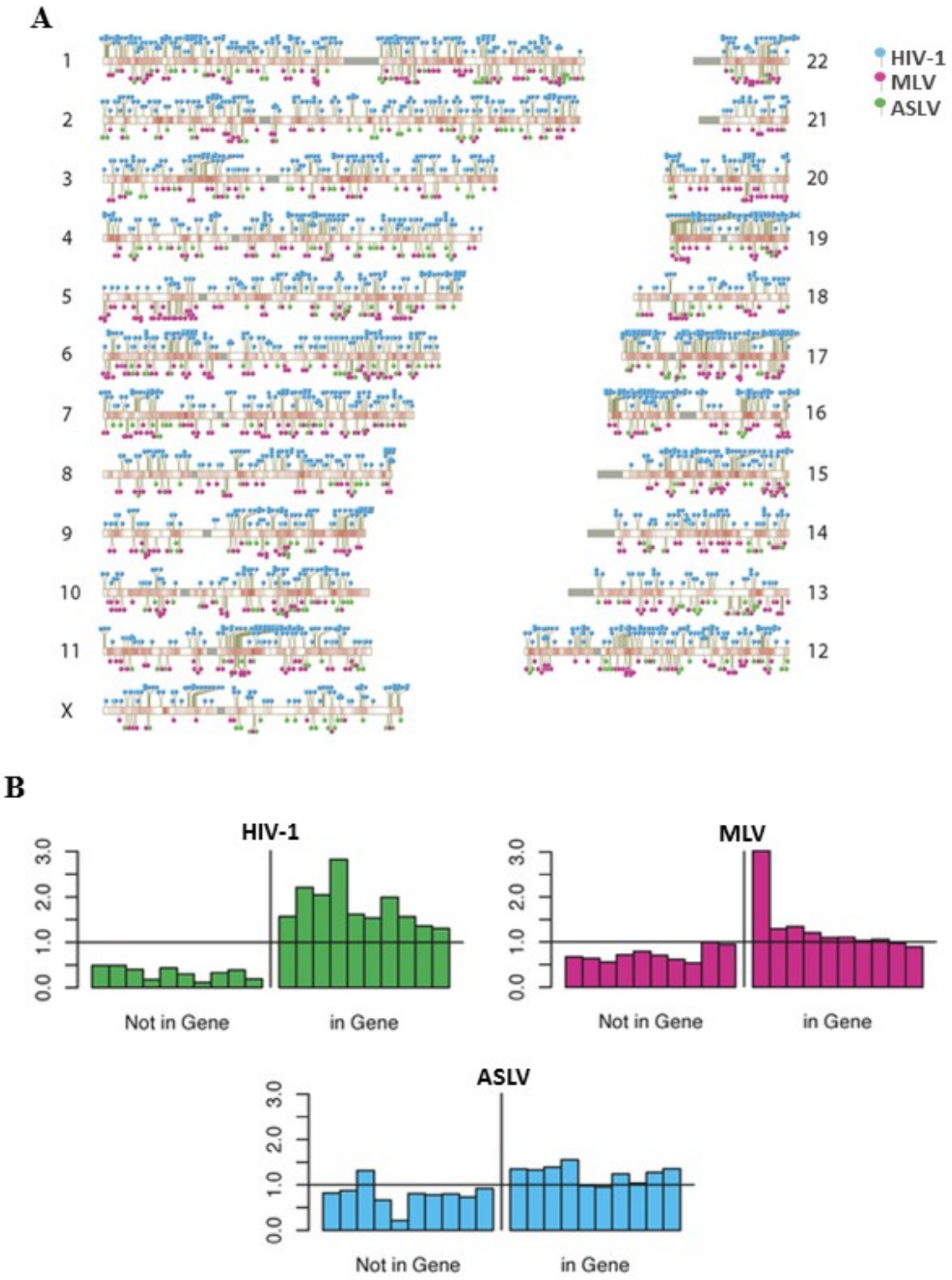

Figure 4. Retroviral integration is genome-wide but preferential for some features. **A.** Genome-wide map of HIV-1, MLV and ASLV proviral integration sites across human chromosomes. **B.** Distribution of HIV-1, MLV and ASLV proviral integration sites in genes. HIV-1 preferentially integrates into active genes and MLV prefers TSS surroundings. Unlike the two retroviruses, ALSV does not display any of the mentioned preferences and is only slightly enriched in gene bodies.

The rest of retroviruses belong to the group of complex retroviruses, in which more complex traits of transcriptional regulation are present. Some of complex retroviruses (δ-, lentiviruses, and FV) were identified to encode accessory proteins regulating their own transcription (Sodroski et al. 1985, Rosen et al. 1986, Keller et al. 1991), while other retroviruses (β- and ε-retroviruses) adapted to the hormonal regulation of the host to transcribe and express their genes in defined space and time (Varmus et al. 1973, von der Ahe et al. 1985, Zhang et al. 1999, Hronek et al. 2004, Morabito et al. 2008). Moreover, additionally to the conventional LTR promoter, some retroviruses (β-, δ-retroviruses, and FV) evolved internal promoters driving the transcription of some retroviral genes (Lochelt et al. 1993, Arroyo et al. 1997, Van Driessche et al. 2016).

Since the discovery of HIV, the regulation of HIV-1 transcription has been a hot topic of retroviral research (a comprehensive review of HIV-1 transcriptional regulation can be found in Ne et al. (2018)). The LTR of HIV-1 is approximately 650 bp long, of which the U3 region occupies about 450 bp. The U3 segment of HIV-1 5' LTR is divided into three regulatory parts: the modulatory region, the enhancer, and the core promoter. The modulatory region is located at the 5' end of U3, occupies the majority of the U3 region (about 350 bp), and is bound by numerous transcription factors that by positive or negative effects modulate the activity of HIV-1 promoter in a context-dependent manner (Cooney et al. 1991, Canonne-Hergaux et al. 1995, Henderson et al. 1995, Horiba et al. 2007, Vemula et al. 2015). Interestingly, the negative regulatory element (NRE) that is similar to the cellular interleukin receptor two is part of the modulatory region of HIV-1 LTR (Smith and Greene 1989). The enhancer region follows the modulatory region and is defined as the region with ability to enhance the transcription of heterologous promoters (Rosen et al. 1985). The most important factor binding to this region is NF-κB, an important transcriptional activator of immune cells. NF-κB is present in activated CD4+ T-cells and activates the LTR-driven transcription of HIV-1. However, other activating transcription factors such as NFAT or Ets1 bind to the enhancer element and activate the LTR promoter activity (Sieweke et al. 1998, Cron et al. 2000). NF-κB binding sites are also bound by HsPB1, which has an inhibitory effect on transcription of HIV-1 (Chaudhary et al. 2016). The core promoter of HIV-1 spans about 250 bp and consists of a GC-rich sequence with three binding sites for Sp1 upstream of the TATA box (Jones et al. 1986). The presence of Sp1 sites is crucial for HIV-1 transcription (Miller-Jensen et al. 2013). Outside of the U3 region, transcription factors influencing the basal activity of HIV 5' LTR were also found to bind to the U5 region (Rabbi et al. 1997). Although the basal transcriptional activity can be achieved by the action of transcription factors, elongation of transcription is very inefficient, with the majority of transcripts terminating at about 60 bp downstream to TSS. To enhance the transcription elongation, HIV-1 transactivator of transcription (Tat) needs to be produced from doubly spliced subgenomic mRNA. Tat then binds to the hairpin-like structure formed at the short transcript, where it enhances the processivity of PolII by bringing the P-TEFb complex that enhances transcription elongation of PolII (Kao et al. 1987). Tat thus represents the positive feedback loop agent and is the crucial factor for effective HIV-1 transcription. Tat was also shown to drive the phenotypic diversity in the population of HIV-1 proviruses (Weinberger et al. 2005). Interestingly, DNA-binding protein ZASC1 binding to the core promoter region upstream to TAR was shown to direct the Tat-P-TEFb complex to LTR of HIV-1 (Bruce et al. 2013). Although the general architecture of LTR is maintained, differences among HIV-1 subtypes leading to different counts of transcriptional factors binding sites were described (Mbondji-Wonje et al. 2018). The variability in LTR also correlates with the clinical outcome of the infection (Nonnemacher et al. 2016, Qu et al. 2016).

Expression from unintegrated genomes

Expression of retroviral genes can be achieved not only from the proviral DNA, but also from unintegrated genomes. Unintegrated vDNA was observed to be loaded by histones immediately after the nuclear entry and subsequently marked by the histone modifications (Wang et al. 2016). Histone modifications are mostly of repressive character, which is a result of NP220 protein binding to cytidine-rich clusters in the U3 region of retroviral LTR (Zhu et al. 2018). According to Zhu et al. (2018), NP220

brings histone deacetylases (HDACs) HDAC1, HDAC4, histone methyltransferase (HMT) SETDB1, and the HUSH complex to the unintegrated genomes of MLV, which cause the epigenetic silencing of MLV unintegrated genomes. HIV-1 unintegrated vDNA was also shown to be targeted by NP220 and bringing HDACs, leading to deacetylation of histones loaded to unintegrated DNA. No binding of NP220 was shown for ASLV unintegrated genomes.

Expression from integration-defective retroviruses or in the presence of integrase inhibitors is weaker than from integration-competent retroviruses. In the case of HIV-1, expression of Nef, Tat and weak expression of Gag can be achieved by unintegrated vDNA, which results in the modulation of target T-cells or spread of viral infection in the presence of integrase inhibitors (Cara et al. 1996, Wu and Marsh 2001, Trinite et al. 2013, Chan et al. 2016, Meltzer et al. 2018). Expression from both HIV-1 and MLV unintegrated vDNA can be enhanced by the treatment by HDAC inhibitors, showing that epigenetic silencing of expression from unintegrated vDNA may be a general phenomenon (Schneider et al. 2012, Pelascini et al. 2013, Chan et al. 2016).

Despite the limited natural expression of integration-defective retroviruses, a new type of viral vectors, the non-integrating retroviral vectors, are being developed for biomedical applications (Shaw and Cornetta 2014). RNA and protein production from unintegrated genomes also needs to be taken into account when reporting the expression analysis of proviral activity (Bonczkowski et al. 2016, Chan et al. 2016).

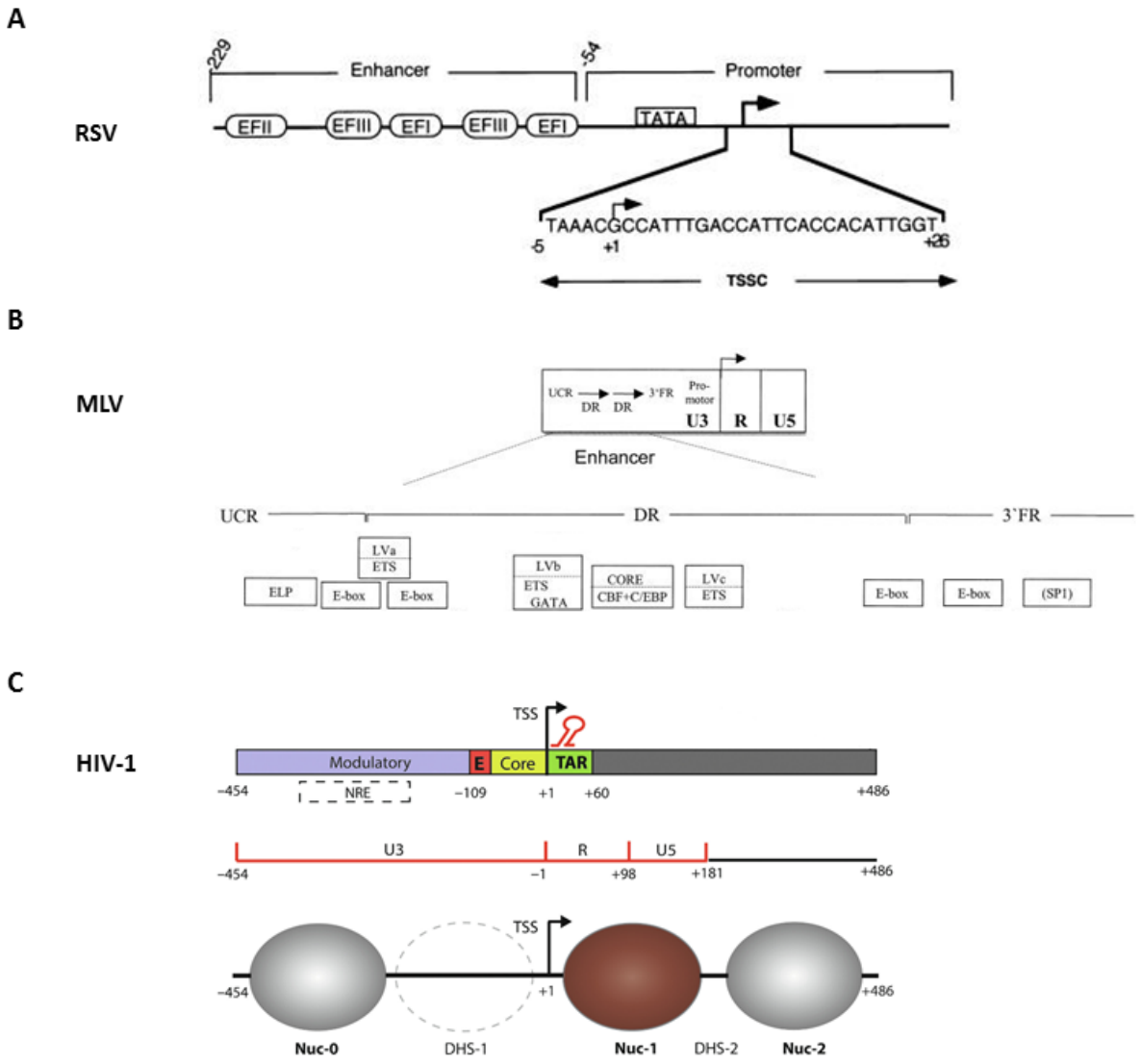

Figure 5. Structure of retroviral promoters. Retroviral LTRs are sites of proviral enhancers and transcription initiation (at +1 nucleotide). **A.** RSV promoter with marked TATA box and cis-acting elements EFI, EFII, and EFIII acting as enhancer factor binding sites that have been shown to be crucial for viral expression. The picture is adapted from Mobley and Sealy (1998). **B.** U3-placed transcription enhancers of MLV. Sites of identified transcription factor binding sites are depicted. U3 LTR of MLV contains two direct repeat regions containing transcription factor binding sites. The picture was adopted from Wahlers et al. (2002). **C.** HIV-1 promoter with marked functional parts of LTR. The bottom part shows the positioning of nucleosomes around TSS (+1 nucleotide). The picture was adapted from Ne et al. (2018).

Silencing of proviral expression

Despite successful integration, proviral transcription can be undetectable, leading to silent infection of the host cell. The loss of expression can be caused by mutations in proviral DNA due to the error-prone process of reverse transcription (Roberts et al. 1988, Mansky and Temin 1995, Yasukawa et al. 2017) leading to large deletions or nonsense mutations in vDNA (Ho et al. 2013). In the case of HIV-1, it was observed that only a few percent of proviruses are intact and thus able of replication (Ho et al. 2013, Bruner et al. 2019). However, in the population of latently infected cells, intact proviruses have been found, showing that epigenetic silencing is also playing its role.

In mammalian cells transformed by RSV, the morphological revertants were observed with high frequencies, leading to the hypothesis of epigenetic silencing of proviral transcription (Wyke and Quade 1980, Chiswell et al. 1982, Hejnar et al. 1994). DNA methylation was showed to play a major role in the silencing of RSV expression in mammalian cells (Chiswell et al. 1982, Katz et al. 1983, Searle et al. 1984, Roguel et al. 1987, Hejnar et al. 1999, Senigl et al. 2012), and cellular protein Daxx was identified as a factor targeting the DNA methylation toward the provirus (Shalginskikh et al. 2013). The significant role of DNA methylation in RSV silencing in mammalian cells is strengthened by the observation that insertion of the core of the CpG-island into the viral LTR enhances the proviral expression by the protection of 5'LTR from DNA methylation. However, the pool of silenced proviruses of ASLV in human cells was also observed being hypomethylated at 5'LTR and thus silenced by a DNA methylation-independent mechanism (Senigl et al. 2012). DNA methylation was observed to be present at the 5'LTRs of MLV and HIV-1, albeit its role seems to be rather in locking the proviruses in the silent state (Niwa et al. 1983, Blazkova et al. 2009). The initial epigenetic silencing for the most of the cases is reported to be caused by histone modifiers removing the acetyl groups from histones H3 and H4 (HDACs) or positioning the repressive marks by HMTs, like di- or trimethylation of histone H3 on lysine 9 (H3K9me2/3).

Despite its general expression in many cell types, the expression of MLV is efficiently restricted in embryonic cells, where specific antiretroviral silencing mechanisms are involved. The restriction is caused by TRIM28 recruitment by DNA-binding protein ZNF809, which binds specifically to the primer-binding site of the provirus (Wolf and Goff 2007, Wolf and Goff 2009). MLV proviruses were showed to be silenced and maintained in silent state by H3K9 HMTs (Matsui et al. 2010, Leung et al. 2011). More proteins interacting with TRIM28, however, were shown to be important for the primer-binding site-dependent silencing in embryonic cells (Wolf et al. 2008, Wang et al. 2014).

HIV-1 and other lentiviruses are known to establish the population of transcriptionally silenced proviruses early after infection (Chavez et al. 2015). Promoters of transcriptionally silenced proviruses are known to be occupied by nucleosomes positioned in a repressive way that inhibits transcription initiation (Verdin et al. 1993). Upon activation of proviral transcription, this repressive positioning is disrupted by chromatin remodeling complex SWI/SNF recruited to the 5' LTR by both Tat-independent and Tat-dependent mechanism (Angelov et al. 2000, Henderson et al. 2004, Treand et al. 2006, Mizutani et al. 2009). Interestingly, biochemically distinct SWI/SNF complexes exist, of which one called BAF inhibits the HIV-1 transcription by positioning nucleosome nuc-1 to the repressive position, while the other one called PBAF, recruited to the 5' LTR by acetylated Tat, removes the nuc-1 from the repressive position, allowing for effective transcription of the proviral genome (Rafati et al. 2011). Transcriptionally active 5' LTR is associated with acetyltransferases, marked by acetylated histones (Van Lint et al. 1996, Sheridan et al. 1997, Lusic et al. 2003), and the loss of acetylation at the histones occupying 5'LTR is associated with silencing of proviral transcription. HDACs were shown to be recruited to the 5' LTR by transcriptional factors such as YY1 or NF-κB p50 (Coull et al. 2000, Pannell et al. 2000, He and Margolis 2002, Williams et al. 2006). Besides deacetylation, the silenced HIV-1 proviral promoter is associated with the factors marking 5' LTR with H3K9 methylation (du Chene et al. 2007, Marban et al. 2007). Inhibitors of epigenetic modifiers, such as HDAC inhibitors, show the ability to reactivate epigenetically silenced proviruses. However, DNA methylation, despite no correlation with proviral silencing (Pion et al. 2003), may accumulate at 5' LTR of the silenced HIV-1 promoter over time (Trejbalova et al. 2016) and promote resistance of the silenced provirus to the reactivation compounds (Blazkova et al. 2009).

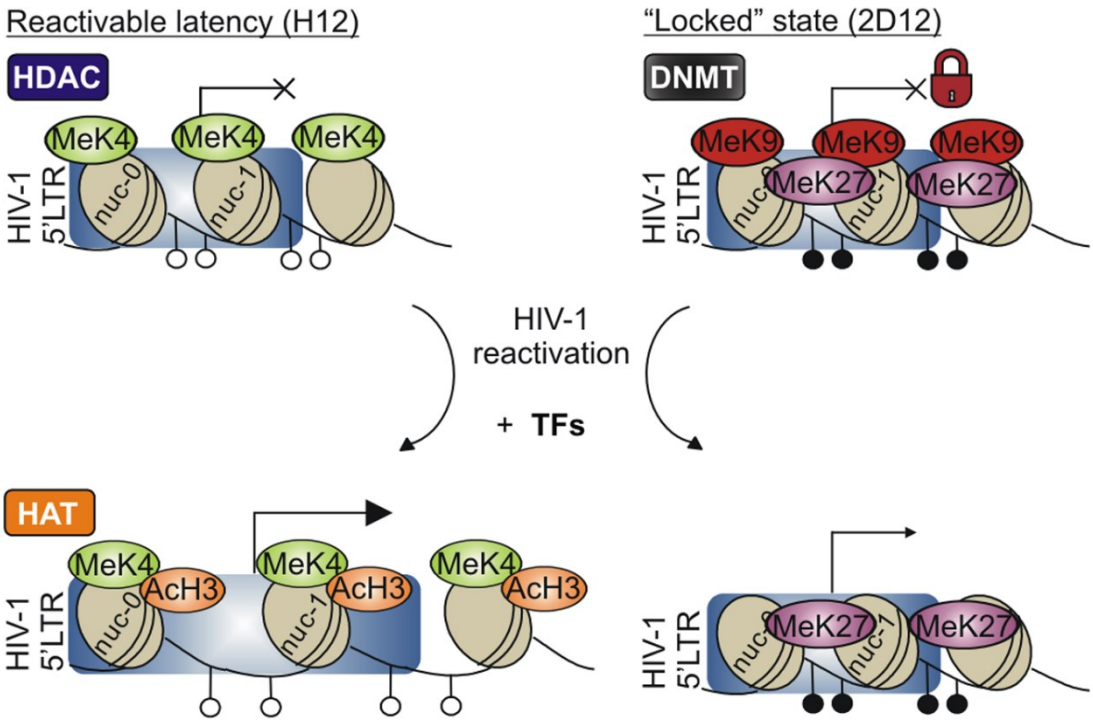

Figure 6. Models of epigenetic silencing of HIV-1 LTR. On the left side, silenced but reactivable proviruses contain deacetylated histones at LTR, which are rapidly acetylated under reactivation of proviral expression. On the right side, proviral LTR is locked in silenced state, when LTR-occupying histones are marked by H3K9me3, H3K27me3, and DNA methylation. The picture is adopted from Blazkova et al. (2009).

Integration site environment and proviral expression

During the studies of the transformation of mammalian cells by avian retroviruses, the observation has been made that while most of the proviral population undergoes morphological reversion, some of the proviruses keep the transformed phenotype (Akroyd et al. 1987, Green et al. 1990, Fincham and Wyke 1991, Lang et al. 1993, Hejnar et al. 1994). The theory that the position effect of the proviral surroundings affects the expression of proviral genes was proposed. Indeed, the proviruses with stable expression were observed to be integrated nearby active promoters marked with H3K4me3 modification (Senigl et al. 2012). Senigl et al. (2012) also observed that proviruses found in the gene bodies were DNA-methylated and silenced in a *de novo* methyl transferase 3b (DNMT3b)-dependent manner, while the intergenic proviruses were DNA methylation-free and the silencing was DNMT-independent.

Early studies done on HIV-1 suggested that the integration site plays an important role in Tat-independent transcriptional activity of the provirus, but Tat transactivation masks the differences introduced by differential chromatin environment (Jordan et al. 2001). The integration site was also found to be a factor influencing the size of the transcriptional bursts from the HIV-1 promoter (Singh et al. 2010). Silenced proviruses were found to be associated with centromeric heterochromatin, intergenic regions, and highly transcribed genes (Jordan et al. 2003, Lewinski et al. 2005). When integrated into genes, the transcription of proviruses was shown to be repressed by transcriptional interference and the convergent transcription of targeted transcriptional unit (Han et al. 2008, Lenasi et al. 2008, Shan et al. 2011). However, no common genomic or epigenomic features were found to be associated with the silenced proviruses across the cell models of latency (Sherrill-Mix et al. 2013, Dahabieh et al. 2014). On the other hand, the expressed proviruses and proviruses sensitive to reactivation from the silent state tend to be more frequent in active genes and closer to the sites enriched in histone modifications marking active TSSs, enhancers, and gene bodies than their unreactivable counterparts (Chen et al. 2017, Battivelli et al. 2018). Expressed proviruses also tend to be further away from the LADs. Quantitative

analysis done by Battivelli et al. (2018) shows that only a few percent of silenced proviruses can be reactivated. Also, retargeting of natural gene-centric integration of HIV-1 by the LEDGINs increases the portion of unreactivable proviruses (Vranckx et al. 2016). The present studies thus show that there might be a connection between the genomic environment of HIV-1 and the activity of proviral expression.

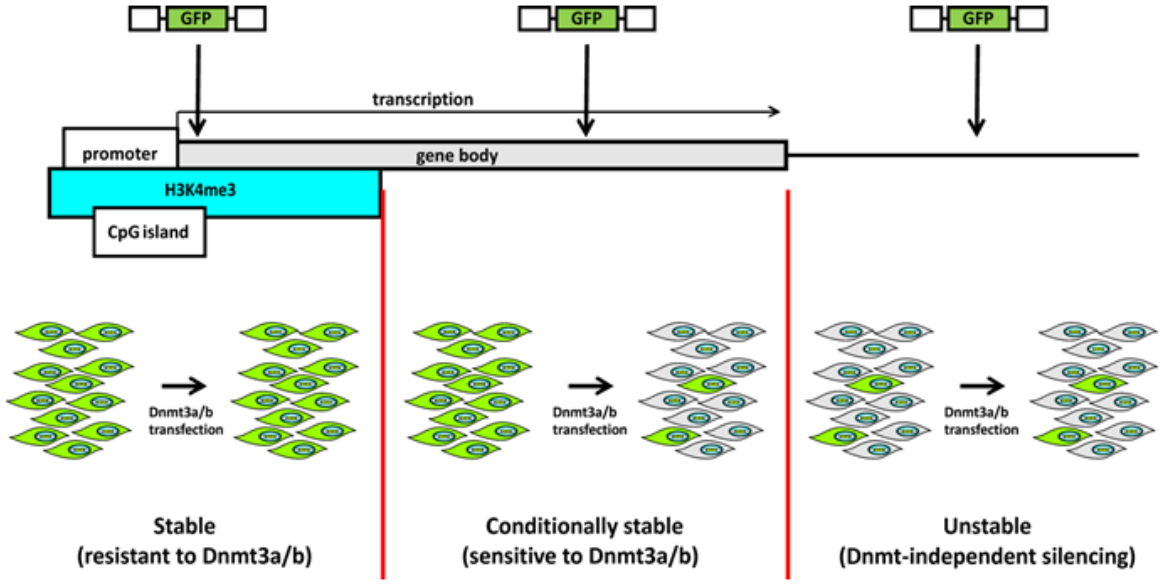

Figure 7. Example of integration site-dependent proviral expression. ASLV proviruses are silenced in a DNA methylation-dependent manner when integrated into the bodies of transcribed genes. When integrated outside genes, proviruses are also silenced, but in a DNA methylation-independent manner. Stably expressed proviruses were found in transcribed genes close to TSSs inside H3K4me3 marked areas. The picture is adapted from Senigl et al. (2012).

Aims

The aim of the thesis is characterization of the features associated with stable proviral expression. The main goal lies in the definition of host cell genomic and epigenomic features associated with stably expressed proviruses. The hypothesis stating that certain features are selected with expressed proviruses, meaning that proviral expression state is integration site-dependent, is tested in this thesis.

Specific Aims

- Compare stability of expression and association of genomic and epigenetic features with stably expressed proviruses of ASLV-derived vectors with wt LTR and LTR modified by the insertion of CpG island core sequence.
- Examine the stability of LTR-driven expression of retroviral vectors derived from mammalian retroviruses and compare it to the expression stability of ASLV.
- Examine whether there is general genomic or epigenetic marker of loci permissive for stable proviral expression or the permissive loci markers are retrovirus-specific.

Materials and Methods

Construction of retroviral and other vectors

All retroviral vectors described here were constructed as single-round replication-deficient vectors expressing the gene for enhanced green fluorescent protein (GFP) under the control of retroviral long terminal repeat (LTR).

Sequences of all primers used in the work presented here are shown in Table 1.

Table 1: Oligonucleotide sequences used in the studies presented

Name	Sequence
GFP-Cherry-End_F	GCATGGACGAGCTGTACAAG
GFP-Cherry-End_R	CTTGTACAGCTCGTCCATGC
GFP-Cherry-Start_F	GTGAGCAAGGGCGAGGAG
GFP-Cherry-Start_R	CTCCTCGCCCTTGCTCAC
HIV-MID10-LTR2_F	TCTCTATGCGCAGACCCTTTTAGTCAGTGTGGAAAATC
HIV-MID1-LTR2_F	ACGAGTGCGTCAGACCCTTTTAGTCAGTGTGGAAAATC
HIV-MID2-LTR2_F	ACGCTCGACACAGACCCTTTTAGTCAGTGTGGAAAATC
HIV-MID9-LTR2_F	TAGTATCAGCCAGACCCTTTTAGTCAGTGTGGAAAATC
HMSpAa	CGAAGAGTAACCGTTGCTAGGAGAGACCGTGGCTGAATGAGACTGGT GTCGACACTAGTGG
HMSpBb_MseI	TACCACTAGTGTCGACACCAGTCTCATTTTTTTTTTTTCAAAAAAA
HMSpBb_NlaIII	CATGCCACTAGTGTCGACACCAGTCTCATTTTTTTTTTTTCAAAAAAA
HMSpBb_Sau3AI/DpnII	GATCCCACTAGTGTCGACACCAGTCTCATTTTTTTTTTTTCAAAAAAA
HMspBb-ANS	CTAGCCACTAGTGTCGACACCAGTCTCATTTTTTTTTTTTCAAAAAAA
Link1	ACCGTTGCTAGGAGAGACCGT
Link2	AATGAGACTGGTGTCGACACTAGTGG
MLV-IN-W390A_F	CTGACAGCGCGCGTTCAACGCTCTCAAAAC
MLV-IN-W390A_R	AACGCGCGCTGTCAGTCTAGAGGATGGTC
MLV-LTR1_F-Bio	(Biotin)-TTCCATGCCTTGCAAAATGGCGT
OLinkA-T	CGAAGAGTAACCGTTGCTAGGAGAGACCGTGGCTGAATGAGACTGGT GTCGACACTAGTGGT
OLinkB-PhoAm	(Phosphorilated)-CCACTAGTGTCGACACCAGTCTCATT-(Amino modifier C6)
pGEM-JLat10.6-3LTR-InFu_R	GAATTCACTAGTGATTGCTTTGGAATGATTTGTTGC
pGEM-pNL4-PRD25A_F	CCGCGGGAATTCGATTACTTACACTAGGAAAGGC
pNL4_PRD25A_vpr_F	ATGGACACTAGAGCTTTTAGAGG
pNL4-vpr_InFu-R	AGCTCTAGTGTCCATTCATTGTATG
spAG3-2IEDD-R	CTCCTCGTGCTGGTACGCTCCCTA
spinAG3-2IEDD-R	ACCTTCCTAATGGCGCCTTCCTAATAA
spinPCR_MLV_R	TGGCGTTACTTAAGCTAGCTTGCC
spinPCR-AG3-R	CGGTGCTTTTTCTCTCCTTGTAAGGCA
spinPCR-HIV-R	CTGCCAATCAGGGAAGTAGCCTTGTGTG
spPCR_MLV_R	TTCCATGCCTTGCAAAATGGCGT
spPCR-AG3-R	CCACCTTACTTCCACCAATCGGCATG
spPCR-HIV-R	TATCTGATCCCTGGCCCTGGTGTGTAG

ASLV-derived vectors

ASLV-derived vectors used here were constructed by Filip Šenigl in the Laboratory of Viral and Cellular Genetics for the purposes of previously conducted studies. Construction of the ASLV-derived GFP-expressing vector (pAG) was described in the publication by Senigl et al. (2012). Construction of the CpG island-modified ASLV vector (pAG-2IE) was described in the publication by Senigl et al. (2008). Briefly, a tandem of two internal elements (IE) from the hamster aprt CpG island was amplified from the previously described pRNIG2-2IE vector and inserted into a *de novo* created unique KasI restriction site in the U3 region of 3'LTR (position -89 respective to the transcription start site) of the AG vector.

MLV-derived vector

MLV-derived vector pLG was constructed as a part of the study conducted by Kalina et al. (2007). pLG was created by insertion of the GFP-encoding sequence into the pLPCX vector (Clonetech). Notably, LTR sequences of the pLG vector differ since 5' LTR is derived from Moloney murine sarcoma virus (MoMuSV), whereas 3' LTR comes from Moloney murine leukemia virus (MoMuLV).

HIV-1-derived vectors

The minimal vector derived from HIV-1 was obtained by cloning of the pEV731 (LTR-Tat-IRES-GFP) vector proviral sequence from a latent H12 cellular clone (Jordan et al. 2001) into the pGEM®-T Easy vector (Promega) by Denisa Kovářová from the Laboratory of Viral and Cellular Genetics.

The full-length HIV-1 vector (pNL4-R7, HIV^{fl} in the main text) was constructed from two parts. The 3' part was derived from the J-Lat 10.6 clone bearing provirus of the HIV-R7/E–/GFP vector (Jordan et al. 2003). The 3' proviral sequence was amplified from J-Lat 10.6 genomic DNA by a Vpr-specific forward primer (pNL4_PRD25A_vpr_F) and an integration site-specific reverse primer (pGEM-JLat10.6-3LTR-InFu_R). The 5' part of the vector was derived from the pNL4-PRD25A molecular clone that was a kind gift from Jan Weber from the Virology Research-Service Team at IOCB. The 5'part was amplified by a plasmid-specific forward primer (pGEM-pNL4-PRD25A_F) and a Vpr-specific primer (pNL4-vpr_InFu-R). Both amplicons were joined and cloned into the pGEM®-T Easy vector (Promega) in a one-step reaction with In-Fusion HD Cloning kit (Clonetech). As a result, an NL4-derived vector encoding GFP in place of the *nef* gene, frameshift mutation in *env* gene, and enzyme-inactivation D25A substitution in PR was created.

The NL4-RG ("red-green") vector was constructed from a pNL4-R7-derived vector where most of the *gag* sequence (nucleotides 869 – 2026 of the pNL4-R7 vector, here called pNL4-dGagGFP) is substituted for the EGFP sequence flanked by a T2A peptide-coding sequence (this vector is not part of this study). EGFP was substituted by the mCherry-coding sequence. mCherry was amplified from the AIDmCherry CSII vector (provided by Filip Šenigl) by the GFP-mCherry_Start_F + GFP-mCherry_End_R primer pair. Other two fragments were amplified from the pNL4-dGagGFP vector using pGEM-pNL4-PRD25A_F + GFP-mCherry_Start_R and GFP-mCherry_End_F + pNL4-vpr_InFu-R primer pairs. The 3' fragment of the vector was amplified from pNL4-R7 by the pNL4_PRD25A_vpr_F + pGEM-JLat10.6-3LTR-InFu_R primer pair. All four fragments (LTR, mCherry, 3' *pol*, 3' pNL4) and pGEM®-T Easy vector (Promega) were linked in a one-step reaction using In-Fusion HD Cloning kit (Clonetech).

MLV IN^{W390A} plasmid

The pMLV-GagPol-IN^{W390A} vector was constructed from the pMLV-GagPol vector by whole-vector amplification with the MLV-IN-W390A_F + MLV-IN-W390A_R primer pair. After amplification, residual pMLV-GagPol was digested by DpnI restriction enzyme and the amplified vector was circularized by In-Fusion HD Cloning kit (Clonetech).

Cell culture and virus propagation

K562 cell line

K562 human myeloid lymphoblastoma cell line was maintained in RPMI 1640 supplemented with 5% newborn calf serum and 5% fetal calf serum (ASLV and comparative MLV, HIV-1 experiments) or solely with 10% fetal calf serum (MLV IN mutatnt experiment) and penicillin/streptomycin (100 mg/ml each, Sigma) in a 5% CO_2 atmosphere at 37°C.

T-cell-derived cell lines

Jurkat, MOLT-4, CEM and HPB cell lines were maintained in RPMI 1640 supplemented with 10% fetal calf serum and penicillin/streptomycin (100 mg/ml each, Sigma) in a 5% CO_2 atmosphere at 37°C.

HEK293T cell line

HEK293T cell line was maintained in D-MEM/F12 (Sigma) with 5% newborn calf serum, 5% fetal calf serum (both Gibco BRL) and penicillin/streptomycin (100 mg/ml each, Sigma) in a 5% CO_2 atmosphere at 37°C.

AviPack cell line

AviPack cell line was derived from the chicken DF-1 cell line as a packaging cell line for avian retroviruses (Plachy et al. 2010) and maintained in D-MEM/F12 medium (Sigma) supplemented with 5% newborn calf serum, 5% fetal calf serum, 1% chicken serum (all Gibco BRL), and penicillin/streptomycin (100 mg/ml each, Sigma) in a 3% CO_2 atmosphere at 37°C.

ASLV virus propagation

Both ASLV-derived vectors were propagated as described previously (Senigl et al. 2012). Briefly, 10^7 AviPack cells plated on a 150 mm Petri dish were cultured and cotransfected with 50 µg of pAG3 and 10 µg of pVSV-G (Clontech) plasmids by calcium phosphate precipitation 24 h after platFresh culture medium supplemented with 100 mM glucose was added 24 hours post transfection and viral stocks were collected twice 48 h and 72 h post transfection. These viral stocks were clarified by centrifugation at 200 x g for 10 min at 4°C, supernatants were collected and centrifuged at 23 000 rpm for 150 min at 4°C in rotor SW28, Beckman Optima100 (Beckman). The pellet was resuspended in the culture medium containing 5% newborn calf serum, frozen, and stored in -80°C. Titration of the infectious virus stock was performed by its serial dilution and subsequent infection of DF-1 cells. Two days post infection (dpi), the number of GFP+ cells was counted under a fluorescent microscope.

MLV virus propagation

To propagate MLV-derived vectors, 10^7 HEK293T cells plated on a 150 mm Petri dish were cotransfected with 6.5 µg pLG, 10 µg of pMLV-GagPol or pMLV-GagPol-IN^{W390A}, and 3.5 µg pVSV-G by calcium phosphate precipitation 24 h after platFresh culture medium supplemented with 100 mM glucose was added 24 hours post transfection and viral stocks were collected twice 48 h and 72 h post transfection. These viral stocks were clarified by centrifugation at 200 x g for 10 min at 4°C, supernatants were collected and centrifuged at 23 000 rpm for 150 min at 4°C in rotor SW28, Beckman Optima100 (Beckman). The pellet was resuspended in the culture medium containing 5% newborn calf serum, frozen and stored in -80°C. Titration of the infectious virus stock was performed by its serial dilution and subsequent infection of K562 cells. Three dpi, the number of GFP+ cells counted using an LSR II cytometer (Becton-Dickinson).

HIV-1 virus propagation

To propagate minimal HIV-1-derived vectors, HEK293T cells were cotransfected by 10 μg of pEV731, 10 μg of psPAX2 (Clontech) and 10 μg of pVSV-G (Clontech) using calcium phosphate precipitation. Viral stocks were collected 48 h after transfection, frozen, and stored in -80°C. Titration of the infectious virus stock was performed by serial dilution and subsequently used to infect K562 cells. Three dpi, the number of GFP+ cells was counted by an LSRII flow cytometer.

To propagate full-length vectors (pNL4-R7 and pNL4-RG), HEK293T cells seeded on a 6-well plate were cotransfected by 0.6 μg transfer vector, 0.9 μg of psPAX.2, and 0.5 μg of pVSV-G by X-tremeGENE HP DNA Transfection Reagent (Sigma-Aldrich) according to manufacturer's instructions. Viral stocks were collected 76 h after transfection, frozen, and stored in -80°C. Titration of the infectious virus stock was performed by serial dilution and subsequent transduction of Jurkat cells. Three dpi, the number of GFP+ cells was counted by an LSRII flow cytometer.

Cell line transduction

Cell lines were transduced by VSV-G-pseudotyped retroviral vectors at low multiplicity to prevent occurrence of multiple proviruses per cell. Multiplicity was determined by the percentage of GFP+ cells in the transduced population at 3 dpi by flow cytometry or during cell sortGenerally, cells were transduced for 30 minutes in a total of 500 μl of transduction medium constituted by serum-free cultivation medium and aliquot of viral stock. Cells were then transfered to a cultivation dish and serum-containing cultivation medium was added. Next day, transduction medium was replaced by fresh cultivation medium in which cells were cultivated until expression measurement or cell sorting.

ASLV transduction of K562 cell line

The amount of 3 x 10^6 cells of the K562 cell line was transduced by pAG or pAG-2IE vectors. In transductions where GFP+ cells were later sorted, the percentage of GFP+ cells at 3 dpi did not exceed 0.5 %. Higher multiplicity of infection was used for the cellular population used for isolation of NS proviruses. In this experiment, the number of GFP+ cells was not examined by flow cytometry. Transduction by a pAG-2IE vector was performed by Filip Šenigl In total, five independent transduction experiments were conducted with the pAG vector.

MLV transduction of K562 cell line

In the experiment performed to examine GFP+ST proviruses, 3 x 10^6 cells of the K562 cell line were transduced by pLG vector. The percentage of GFP+ cells at 3 dpi did not exceed 0.5 %. Transduction was performed by Filip Šenigl One transduction experiment for production of GFP+ST cellular clones was performed.

In the experiment performed to examine proviral populations of the pLG vector equipped with IN variants, three independent transductions are presented. To examine vector expression in a mixed population, 1 x 10^5 cells were transduced in parallel by each vector. In other experiments involving sorting of GFP+ cells, 3 x 10^6 cells of the K562 cell line were transduced. In the experiment where single-cell sorting was performed, GFP+ cells did not exceed 0.5 % at 3 dpi. In the experiment where polyclonal GFP+ populations were established, the percentage of GFP+ cells in the populations at 3 dpi did not exceed 5 %.

HIV-1 transduction of K562 cell line

Three different transduction experiments where 3 x 10^6 cells of the K562 cell line were transduced by pEV731 vectors are presented. In transductions where GFP+ cells were later sorted, percentage of GFP+ cells at 3 dpi did not exceed 0.5%. Higher multiplicity of infection was used for the cellular population used for isolation of NS proviruses. In this experiment, number of GFP+ cells was not examined by flow cytometry.

HIV transduction of T-cell derived cell lines

To compare the expression stability in a T-cell-derived cell line, the Jurkat cell line was transduced in parallel with the K562 cell line, where 1 x 10^5 cells were transduced by the pEV731 vector.

Cell sorting and flow cytometry analysis

At 3 dpi, cells were collected, spun down and resuspended in 1 ml of serum-free RPMI. Right prior to sorting, cells were passed through a CellTrix® 50 µm filter and Hoechst 33258 was added. Sorting of GFP+ cells in single-cell or "bulk" mode was always performed at 3 dpi with an Influx cell sorter (Becton-Dickinson). When the single-cell mode was applied, cells were sorted into a flat-bottom 96-well plate with about 100 µl of cultivation medium per well. In the bulk mode, cells were sorted into 15 ml centrifuge tubes (TPP) with about 8 ml of cultivation medium and subsequently spun down, resuspended in fresh cultivation medium, and transferred to a cultivation plate.

Flow cytometry analysis was performed in an LSR II cytometer (Becton-Dickinson). Prior to measurement, 200 µl of cell suspension of each sample was collected and placed into a well of a U-shaped 96-well plate. The 96-well plate was spun down for 3 min at 500 g. Cellular pellet was resuspended in 50 µl of Hoechst 33258-containing phosphate-buffered saline (PBS). Measurements were performed using a High-Throughput Sampler using 10 µl of each sample. The measurement threshold was set to 10,000 Hoechst-negative cells. Analysis of cytometric data was performed with FlowJo software. Gates were always set to the nontransduced negative control.

Cell sorting and flow cytometry analysis of pAG3-2IE and pLG vector-transduced cells were performed by Filip Šenigl

Integration site isolation and sequencing

DNA isolation

The genomic DNA was isolated from collected cells by phenol-chloroform extraction. First, cells were spun down for 10 minutes at 200 g and the cell pellet was frozen or directly carefully resuspended in lysis buffer (1% of SDS and 0.25M EDTA), and proteinase K was added in final concentration of 1 mg/ml. The solution was incubated at 55°C for at least 3 h or overnight. RNAse H at final concentration 0.3 µg/µl was added and the sample was incubated for an additional 1 h at 37°C. The lysis buffer volume of the phenol-chloroform mixture (1:1, pH 7.9) was added to the sample and gently mixed, and the sample was spun down at 16,200 g for 5 min. The upper water phase was carefully transferred to a new tube and 1x the volume of sample of cold 96% ethanol was added, and the sample was mixed by shaking and spun down for 5 minutes. The pellet was once washed by 80% ethanol, spun down, and dried at 37 °C. Finally, the pellet was resuspended in 50 µl of TE buffer and stored in 4 °C. Concentration of isolated DNA was established by measurement in a NanoDrop 1000 Spectrometer (Thermofisher).

DNA was prepared from individual ASLV and HIV-1 GFP+ST clones after their examination by flow cytometry at 60 dpi. MLV clones whose expression was examined were frozen on a 96-well plate by Filip Šenigl Later, the clones were thawed, and the clones meeting the criteria for GFP+ST clones were mixed and DNA was purified from the mixture of clones. Other samples (NS and GFP+^{3dpi}) were harvested as polyclonal populations.

Splinkerette PCR

To identify the genomic sequence neighboring LTR of the integrated provirus, we adapted the splinkerette PCR method (Uren et al., 2009). In this method, DNA is digested by restriction endonuclease and splinkerette adaptors are ligated to the digested DNA. After additional digestion by a restriction enzyme utilized to preclude amplification of inner sequences, an unknown piece of genomic DNA located between LTR and ligated splikerette adaptor is amplified in nested PCR reactions using pairs of LTR and adaptor-specific primers.

We digested 2 µg of purified genomic DNA in 20 µl reaction by vector-suitable restriction endonucleases (all of which were purchased from New England Biolabs). DNA of ASLV-derived vector-transduced cells was digested by DpnII or MseI. DNA from MLV-transduced samples was digested by NlaIII. For HIV-1 samples, NlaIII, MseI, or a mix of SpeI, NheI, and XbaI (SNX mix) was used. Overnight digestion was performed, after which, when possible, the endonuclease mix was inactivated by high temperature.

Splinkerette adapters were created by annealing of HMspAa and restriction endonuclease-specific HMspBb olinucleotides. The adaptor mix containing both oligonucleotides (25 µM of each) and 1x Restriction buffer M (Boehringer Mannheim) was heated at 94°C for 3 minutes and then cooled to 21°C by 1°C per 15s. Five µl of the adaptor mix was then added to 20 µl of the ligation mix containing 3 µl of the digestion mix, ligase buffer (1x, New England Biolabs), and 400 U of T4 DNA ligase (New England Biolabs). The mixture was incubated overnight at 15°C, after which the reaction was terminated by incubation at 65°C for 20 minutes.

To eliminate amplification of the inner proviral sequences, a second round of restriction endonuclease was applied. Endonuclease-specific buffer, BSA, and 10 U of restriction endonuclease were added to the ligation reaction and filled up to the volume of 50 µl by water. ASLV-transduced samples were digested by Bsu36I, MLV samples by ClaI, and HIV-1 samples by PvuII. Overnight digestion was performed, after which, when possible, the reaction was heat inactivated. DNA was then purified by a QIAquick PCR Purification Kit (Qiagen).

Primary PCR was performed with Link1 and LTR-specific spPCR primers as follows: 94°C for 3 min, 2 cycles of 94°C 15 s, 68°C 30 s, 72°C 2 min and 31 cycles of 94°C 15 s, 62°C 30 s, 72°C 2 min and final polymerization 72°C for 5 min. The secondary PCR used Link2 and LTR-specific spinPCR primers with the program settings: 94°C 3 min, 30 cycles- 94°C 15 s, 60°C 30 s, 72°C 2 min and final 72°C 5 min.

Sequencing of integration sites

After the splinkerette protocol was applied to DNA isolated from cellular clones, individual samples were examined in agarose gel, and single DNA bands were cut out and purified by a QIAEX II Gel Extraction Kit (Qiagene) and sequenced by a Sanger sequencing-providing lab from spinLTR primer.

Polyclonal samples (all NS and GFP+^{3dpi} samples and sample of mixed MLV GFP+ST clones) were run in agarose gel where a smear in the range of 200 – 800 bp of amplified DNA was cut out and purified by a QIAEX II Gel Extraction Kit (Qiagene). Purified DNA was cloned into pGEM®-T Easy vector (Promega), and XL1-Blue competent cells were then heat-shock transformed by the vector mix and seeded on X-Gal-containing plates. Single "white" colonies were picked and moved to a 96-well plate filled with solid bacterial medium. The whole plate was then send to the GATC company to perform Sanger sequencing of particular colonies from a common bacterial plasmid M13-specific primer.

The entire preparation of samples originating from cells transduced by pAG3-2IE vectors were performed by Miroslav Auxt and Filip Šenigl.

Library preparation for next-generation sequencing of MLV-transduced cells

The splinkerette PCR method adapted for next-generation sequencing (NGS) was applied for identification of genomic DNA-LTR junctions of NS and GFP+^{3dpi} proviral populations of K562 cells transduced with pLG vectors with IN^{wt} and IN^{W390A} variants. Briefly, genomic DNA was subjected to random fragmentation and adaptors were ligated to the end-repaired fragments. Sequences between LTRs and adaptors were amplified by nested PCR. Notably, in the first round of the PCR, biotinylated LTR-specific primers are used to later enrich for target sequences. The resulting fragments of 100 to 200 bp in length were cut out from agarose gel and subjected to sequencing at Ion Proton (Thermofisher) platform. The library for NGS was prepared by Martina Slavková.

The protocol follows in detail: According to manufacturer's instructions, 4 µg of genomic DNA was treated by 8 µl of NEBNext® dsDNA Fragmentase® (New England Biolabs) in 80 µl reaction. After 20 min, 40 µl aliquots were taken and 10 µl of 0.5M EDTA was added to each sample to stop the reaction. Fragments longer than 100 bp were purified by application of AMPure magnetic beads (Beckman Coulter) at the 1:0.4 ratio. The eluted sample was incubated with end-repair mixture containing T4 DNA polymerase (3 U, New England Biolabs), T4 polynucleotide kinase (10 U, New England Biolabs), and Taq DNA polymerase (10 U, New England Biolabs) for 20 min at 20°C followed by incubation for 25 min at 72°C.

Adapters were prepared by annealing OlinkA-T and OlinkB-PhoA. The OlinkB-PhoA oligonucleotide contains C6-amino modifier at the 3‘ end and phoshate at the 5‘ end. Thirty-fold pmol excess of adapter to DNA was used and 0.9 µl of Quick Ligase (New England Biolabs) was added to the end-repair mixture and incubated at 25°C for 20 min. To prevent inner sequence amplification, the sample was digested with 5U of ClaI (New England Biolabs) in 50 µl reaction. After overnight digestion, the reaction was heat inactivated (65°C, 20 min) and DNA was purified by a High Pure PCR Cleanup Micro Kit (Sigma-Aldrich), from which 15 µl of each sample was retrieved.

In the first round of nested PCR, linear amplification with a biotinylated MLV-LTR1_F-Bio primer was performed using 100 ng of template sample (95°C for 2 min, then 12 cycles of 95°C for 15 s, 62°C for 30 s and 72°C for 1 min, and then 72 °C for 10 min), after which the Link1 primer was added and standard PCR amplification was performed (95°C for 2 min, then 27 cycles of 95°C for 15 s, 62°C for 30 s and 72°C for 1 min, and then 72°C for 10 min). Samples were purified by a High Pure PCR Cleanup Micro Kit (Roche) and biotinylated amplicons were purified by streptavidin-coated beads (Dynabeads MyOne Streptavidin C1, Invitrogen). The second round of nested PCR was performed with template-coated streptavidin beads and barcoded MLV-LTR2_F primers (MID1 for IN^{wt} NS, MID2 for IN^{W390A} NS, MID9 for IN^{wt} GFP+^{3dpi} and MID10 for IN^{W390A} GFP+^{3dpi}) and Link2 primers (95°C for 2 min, then 35 cycles of 95°C for 15 s, 60°C for 30 s and 72°C for 1 min, and then 72°C for 10 min). From each sample, 100 – 200 bp fragments were cut out from agarose gel and purified by a QIAEX II Gel Extraction Kit (Qiagen).

After extraction, samples were pooled and forwarded to Ion Proton sequencing provided by SEQme s.r.o as a part of ShareSeq service.

Integration site sequence mapping

Mapping of sequences obtained by Sanger sequencing

Sequences of clonal and polyclonal populations of ASLV, MLV, and HIV-1-transduced cells obtained by Sanger sequences were mapped individually using a web-based BLAST-like alignment tool (BLAT) found at the University of California Santa Cruz (UCSC) Genomic Institute web page (https://genome.ucsc.edu/cgi-bin/hgBlat). Prior to mapping, sequences containing the sequence of LTR end and lacking inner proviral sequences were identified and LTR sequences were removed. If a splinker sequence was found in the sequence, this was also removed. The final sequence was then mapped against the February 2009 version of human reference genome assembly (GRCh37/hg19). From the resulting hits, hits were selected if the alignment started at the beginning of the sequence mapped (query sequence) and if the best hit provided the unique, high score of alignment compared to the second best hit. All query hits not fulfilling the criteria were not used in subsequent analysis. For uniquely mapping hits, coordinates comprising the chromosome, nucleotide number, and orientation of the first, LTR-proximal nucleotide were saved. The orientation coordinate value was subsequently changed to the opposite value to act as a marker of provirus orientation. These coordinates were called integration position (iPos).

Mapping of MLV integration site sequences obtained by NGS

Prior to mapping, reads were preprocessed by Unix-adapted tools. FASTQ/A Barcode splitter from FASTX-Toolkit was used to split reads into MID-specific groups according to the barcode with *--mismatches 2* option. Reads in each group were renamed to contain a MID-specific identifier and were pooled to a single file. The Cutadapt tool with options *-g CCTACAGGTGGGGTCTTTCA -m 20* was applied to cut LTR sequences. The "--discard-untrimmed" option was used to obtain LTR-containing sequences. Cutadapt was also used to remove reads with the inner proviral sequence amplified (with *-a TTCCCCCCTTTTTCTGGAGA --discard-trimmed* options) and to trim the adaptor sequence if present (with *-a ACCACTAGTGTCGACACCAG -m 20* options).

Preprocessed reads were run as query against the GRCh37/hg19 genome with the BLAT tool using *-tileSize=11 -stepSize=9 -minIdentity=85 –maxIntron = 5 –minScore = 27 –dots = 1000 -out=psl -noHead* options that were adapted from INSPIIRED pipeline (Berry et al. 2017). Unique hits were selected using R scripts utilizing GenomicRanges and GenomicStrings Bioconducor packages. Hits accepted were allowed to start within 3 bp from the LTR-proximal read start. Reads with multiple hits were accepted if the second best score was not higher than 10 % of the best score. iPos was generated for each accepted hit. iPos was further accepted as real integration site if it was covered by more than two hits. If more iPos were next to each other, iPos were merged into a single iPos with the highest count of hits (reads) in the merged group.

Processing of HIV-1 integration sites from Zhyvoloup et al. (2017)

Coordinates of integration sites were obtained from supplementary data of Zhyvoloup et al. (2017) and remapped to the December 2013 version of human reference genome assembly (GRCh38/hg38).

Random-position control generation

Random controls at AG vs AG-2IE study

As a control of random targeting of features for a low number of integration sites, we generated a set of 200 biologically unbiased genomic positions. Human chromosomes were virtually joined (from chromosome 1 to chromosome X) and 200 random positions in range 1 – 3,031,042,417 (chromosome Y was omitted) were generated. Genomic coordinates were then obtained by mapping of random positions to chromosomal joint-genomic positions.

Uniquely matched random controls (umMRCs)

To create a set of random genomic positions with position-to-restriction enzyme-recognized site distribution, the restrSiteUtils_1.2.8 R package was used (Berry 2017). A set of 200 matched random positions per integration site was generated. Genomic sequences covering ranges from a random position to a restriction enzyme-recognized site were extracted from GRCh37/hg19 assembly using the Biostrings R package for each matched random position (Pagès et al. 2017). Sequences were mapped to GRCh37/hg19 assembly using BLAT, and sequences with single full-length match with ≥ 98% identity were accepted as uniquely mapped matched random controls (umMRCs). Three umMRCs per integration site were randomly selected and used as controls in subsequent analysis.

Feature-matched random controls (agMRCs, GMRCs, AGMRCs, EMRCs)

To analyze the effect of feature targeting on targeting other features, feature-matched random controls were generated. Generally, random positions that matched particular iPos concerning the distance to a particular feature were selected. All groups of feature-matched MRCs were generated by R scripts.

In the study comparing GFP+ST proviral integration sites of ASLV, MLV and HIV-1, active gene-matched MRCs (agMRCs) were generated. One thousand of umMRCs per integration site were generated. For each umMRC, the distance to active RefSeq Genes which TSSs associated with a Tss chromatin segment was counted in the same manner as performed for integration site analysis. For each integration site, the three umMRCs with the most similar distance to an active gene when compared to the integration site were randomly selected.

In the analysis of HIV-1 data of Zhyvoloup et al. (2017), different groups of gene-targeting MRCs were generated. To each iPos, one random position was generated using the *random* tool from bedtools utilities. To match random positions to iPos with the distance, first, random positions within the matched feature were generated using the *shuffle* tool from bedtools with *-incl* option referring to intra-feature BED-formatted genomic sites. Inter-feature regions were generated using the *complement* tool from samtools, and random sites matching iPos with the distance to the feature were generated using R script randomly choosing (R *sample* function) positions with the exact to-feature distance as matched iPos.

Evaluation of integration site association with features

MySQL-based distance measurements

In the studies comparing expression and proviral integration sites of ASLV-derived vectors (pAG3 and pAG3-2IE), distances and frequencies of the features were calculated using the MySQL Workbench 6.2 software.

R-based distance measurements

In the study comparing distribution of ASLV, MLV and HIV-1, distances of proviral integration sites to features and frequency of feature targeting was managed by R scripting using the Bioconductor GenomicRanges package (Lawrence et al. 2013).

Data source

Genomic and epigenomic data for *in silico* analysis of features associated with integration sites were obtained from UCSC golden path (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/database/). Data for CAGE-peak positions [FANTOM Consortium and the RIKEN PMI and CLST (DGT)] were obtained from web-based data source http://fantom.gsc.riken.jp/5/.

Transcriptional units

All integrations mapped into the RefSeg Genes were considered as intragenic. The absolute distance to TSS marks the distance to the closest TSS in the RefSeq Genes track. In the relative distance to TSS, intergenic integration distance to TSS equals to the absolute distance to TSS. For integrations inside RefSeq Genes, the distance to the nearest TSS of a particular RefSeq Gene targeted by the integration was calculated.

RNA-seq data for the K562 cell line

Seven RNA-seq datasets for the K562 cell line (ERR310212, SRR090233, SRR346063, SRR521457_1, SRR644784, SRR901899 and SRR901900) were obtained from Sequence Read Archive (SRA) at the NCBI website (http://www.ncbi.nlm.nih.gov/Traces/sra/). Reads were mapped to human genes using CLC Genomics Workbench 6.5.1 with default settings. RefSeq Genes were classified into groups by the mean of reads per kilobase per million (RPKM), with the NA group containing RefSeq Genes with the mean RPKM lower than 1 and RefSeq Genes with no match in CLC Genomics Workbench 6.5.1 database and Q1-Q4 groups containing RefSeq Genes with the mean RPKM equal to/higher than 1 classified into the mean RPKM quartile groups, with Q1 being the first quartile containing RefSeq Genes with the lowest RPKM.

Selection of dominant transcripts from Jurkat RNA-seq data

Raw Jurkat RNA-seq data published by Zhyvoloup et al. (2017) were downloaded from NCBI SRA accession PRJNA321856. Reads were mapped on GRCh38/hg38 using TopHat2. The mapped reads were filtered to remove PCR duplicates and select uniquely mapped reads using the samtools *view* function with *-bq 4 -F 0x400* options. The mapped reads were associated with RefSeq Genes transcripts using the Cufflinks tool. Transcript coordinates, name ("name" column), RPKM, and associated RefSeq Gene name ("name2" column) were extracted using the awk tool. R script was then used to select a single dominant transcript for each RefSeq Gene. For DMSO-treated samples, the ERR2097196 SRA dataset was discarded from analysis since RPKM values of transcripts did not correlate with the rest of the DMSO-treated dataset. Transcripts with RPKM > 0 were selected. For each RefSeg Gene ("name2" column), one associated transcript ("name" column) with highest RPKM value was selected. In the following analyses, RefSeq Genes were selected according to the RPKM value of the dominant transcript.

Histone modification peaks

Histone modification peak datasets for the K562 cell line from Broad Histone track were obtained from UCSC golden path. Peaks with signals higher than the median of the signal of particular histone modification peak dataset were selected for further analysis. The distance to the nearest peak of particular modification was calculated for each integration.

Chromatin segments

Chromatin segments for the K562 cell line were obtained from the UCSC Genome Segments track. Segments were grouped by the itemRgb field. Mnemonics for the segments are adapted from Hoffman et al. (2013) and explained in Table 2. For global analysis, we merged the related segments. The active chromatin segment group contains 18 segments including Tss, TssF, PromF, PromP, Enh, EnhF, EnhW, EnhWF, DnaseD, DnaseU, FaireW, Gen5, Elon, ElonW, ElonWF, Gen3, H4K20, and Low. The regulatory segment group consists of 11 segments including Tss, TssF, PromF, PromP, Enh, EnhF, EnhW, EnhWF, DnaseD, DnaseU, and FaireW.

Table 2: Chromatin segment mnemonics

Mnemonic	Rationale
Tss	Active promoter, TSS/CpG island region
TssF	Active promoter, flanking TSS/CpG islands
PromF	Promoter flanking
PromP	Inactive/Poised promoter, highly conserved
Enh	Candidate strong enhancer, open chromatin
EnhF	Candidate strong enhancer, flanking open chromatin
EnhWF	Candidate poised/weak enhancer; flanking open chromatin of candidate enhancers
EnhW	Candidate weak enhancer and open chromatin
DNaseU	Primarily UW DNase, weaker open chromatin sites
DNaseD	Primarily Duke DNase, candidate regulatory elements in more likely repressive locations
FaireW	Modest Faire/Control enrichments, potential CNV
CtcfO	Distal CTCF/Candidate insulator with open chromatin
Ctcf	Distal CTCF/Candidate insulator without open chromatin
Gen5'	Transcription transition, highly expressed genes towards 5' end
Elon	Transcriptional elongation, stronger H3K36me3, more exonic
ElonW	Transcriptional elongation, weaker H3K36me3
Gen3'	Transcription 3' end of genes, highly expressed; more exonic
Pol2	Pol2-specific locations, mostly in genes but a substantial portion in intergenic locations
H4K20	Transcription, primarily H4K20me1, more intronic
ReprD	Polycomb repression with Duke DNase sites/promoter and conservation enrichment (except HELA)
Repr	Strong Polycomb repression
ReprW	Weaker Polycomb repression
Low	Low signal proximal to active elements
Quies	Heterochromatin/Dead Zone
Art	Potential CNV or repetitive artifacts

Active genes

Different groups of active genes were created. A RefSeq Gene was considered to be active if its TSS was either within the H3K4me3 peak or within the Tss chromatin segment or within the distance of 500 bp from the nearest CAGE peak TSS [FANTOM Consortium and the RIKEN PMI and CLST (DGT)]. Three groups of active genes/TSSs were analyzed separately.

Chart generation and statistics

Charts were produced using the R (integration site distribution charts) or Excel (expression charts) software. Statistical tests were performed with the R software.

Results

Integration sites and expression of ASLV

The rate and kinetics of provirus silencing of ALSV-derived vectors in human cells

To tackle the question of the role of integration site environment in proviral expression, we took advantage of previously described enhanced silencing of ASLV-derived vectors transducing mammalian cells (Wyke and Quade 1980, Chiswell et al. 1982, Hejnar et al. 1994, Senigl et al. 2008, Senigl et al. 2012). Here we described the rate and kinetics of provirus silencing observed after transduction of human K562 cells with ASLV-derived vectors. We compared the silencing of two replication-defective, ASLV-derived GFP-transducing (AG) vectors (Fig 8A). Both vectors, AG and AG-2IE, differ solely by insertion of two internal elements (IE) from a CpG island into the U3 segment of AG-2IE. The IE comprises the core sequence of CpG island from the hamster adenine phosphoribosyl transferase gene along with a tandem of high-affinity Sp1 binding sites; the effect of this insertion on the expression of ASLV-derived vectors has been previously described (Senigl et al. 2008). Both vectors were produced by cotransfection of the AviPack packaging cell line (Fig. 8B). The K562 cell line is an ENCODE Tier1 cell line providing extensive epigenomic data for subsequent analyses and represents a valuable gene therapy model.

K562 cells were transduced with VSV-G-pseudotyped vectors at a low multiplicity of infection (MOI < 0.01), which was necessary to minimize the probability of multiple proviral integrations per cell (Fig. 8C). GFP-positive (GFP+) cells with transcriptionally active proviruses were separated by fluorescence-activated cell sorting (FACS) three days post infection (3 dpi) and single-cell-derived clones were established. After expansion, the clones were cultivated for two months or longer with a FACS count of GFP+ cells at 30 and 60 days. Two sets of clones, 2,128 clones that contained expressed AG proviruses and 558 clones with expressed AG-2IE proviruses, were expanded and cultivated. Clonal analysis of the reporter expression confirmed the high rate of provirus silencing in the AG vector. Cellular clones consisting of at least 90 % of GFP+ cells were pronounced to carry the provirus stably expressing GFP (GFP+ST). At 60 dpi, only 3.5 % (74 of 2,128) of AG clones contained GFP+ST, whereas the majority of clones tended to undergo rapid silencing (Fig. 9A). This behavior was in sharp contrast to much less effective and slow provirus silencing in AG-2IE clones, with 29 % (164 of 558) of the GFP+ST clones at 60 dpi. The striking contrast of silencing in AG and AG-2IE vectors was apparent already by 30 dpi.

In previous studies, Hejnar et al. (2001) and Senigl et al. (2008) demonstrated that the anti-silencing effect of IE insertion lies in protection from DNA methylation, but does not increase LTR-driven expression. Here, we compared the mean fluorescence intensity (MFI) of GFP in GFP+ST clones from both AG and AG-2IE groups. GFP+ST clones containing AG proviruses exhibited higher GFP MFI in comparison with non-selected GFP+ cells at the beginning of clonal expansion (Fig. 9B). AG-2IE proviruses, however, exhibited lower variability and approximately the same MFI in GFP+ST clones and non-selected cells. This can be explained by the more autonomous expression of AG-2IE proviruses, which was less influenced by position effects in direct contrast to AG proviruses, whose expression is mostly determined by position effects and where the high expression correlates with stability. Our clonal analysis independently confirmed the anti-silencing but not transcription-enhancing effect of the IE insertion.

The higher autonomy of AG-2IE proviruses is also supported by the course of GFP expression silencMost AG clones proceeded to silencing through a gradual decrease of GFP intensity, and in the transient state, there was a broad variability of fluorescence intensities in individual cells (Fig. 9C). In contrast, most transcriptionally active AG-2IE proviruses in unstable clones simply switched to a silenced state, and the distribution of fluorescence intensities was bimodal without intermediate states.

We conclude that the insertion of the CpG island core sequence partly protected retroviral vectors from provirus silencing and released the dependence of vectors on the local position effects towards the vector expression.

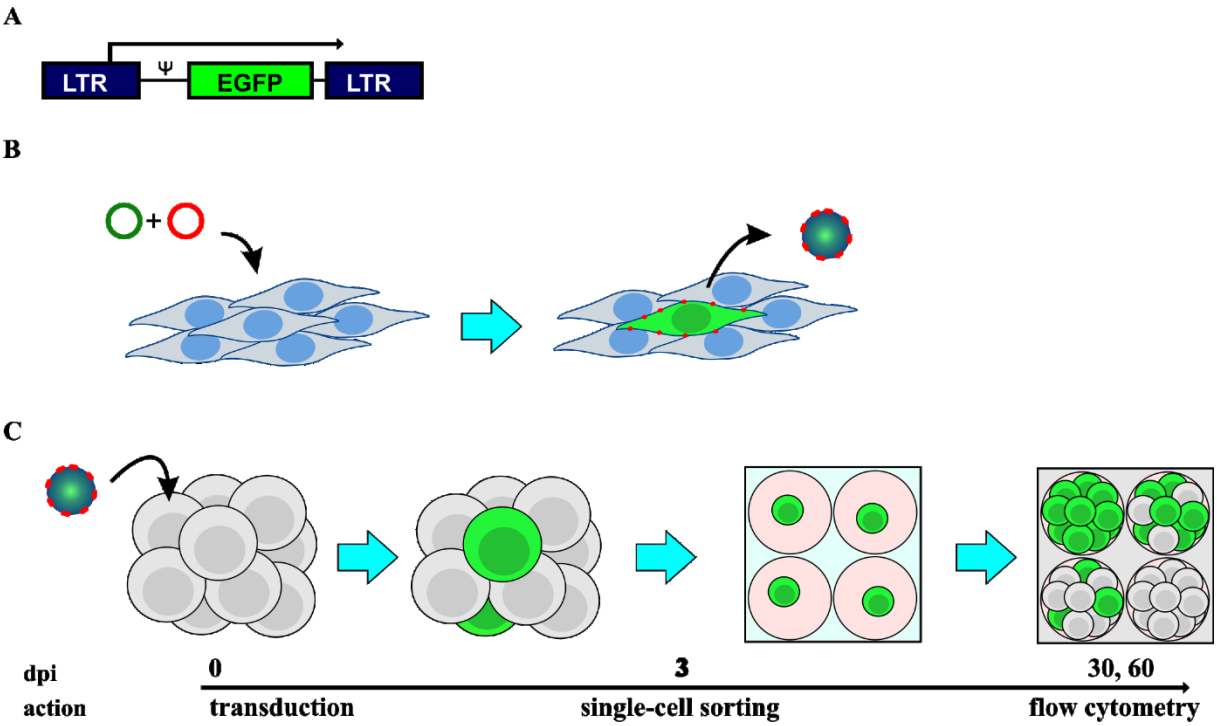

Figure 8. Outline of the experimental approach. **A.** Scheme of ASLV-derived minimal vector expressing the EGFP marker under the control of LTR. LTR – long terminal repeat, Ψ – packaging signal, EGFP – enhanced green fluorescent protein. Arrow marks the transcription provirus. **B.** The stock of ASLV retroviral vector is produced by the packaging cell line (AviPack) cotransfection with transfer and envelope (VSV-G) vectors. **C.** Target cells (K562) transduced by low MOI. GFP+ cells are single-cell sorted at 3 dpi. From cells sorted, cellular clones are grown. The percentage of each clone is measured by flow cytometry at 30 dpi. Expression of highly active clones is followed until 60 dpi, which is the last point of measurement. Clones showing at least 90 % of GFP+ cells at 60 dpi are reported as GFP+ST clones, and proviral integration sites of the clone-containing provirus are sequenced. MOI – multiplicity of infection, dpi – days post infection, GFP+ST – GFP-stably expressing.

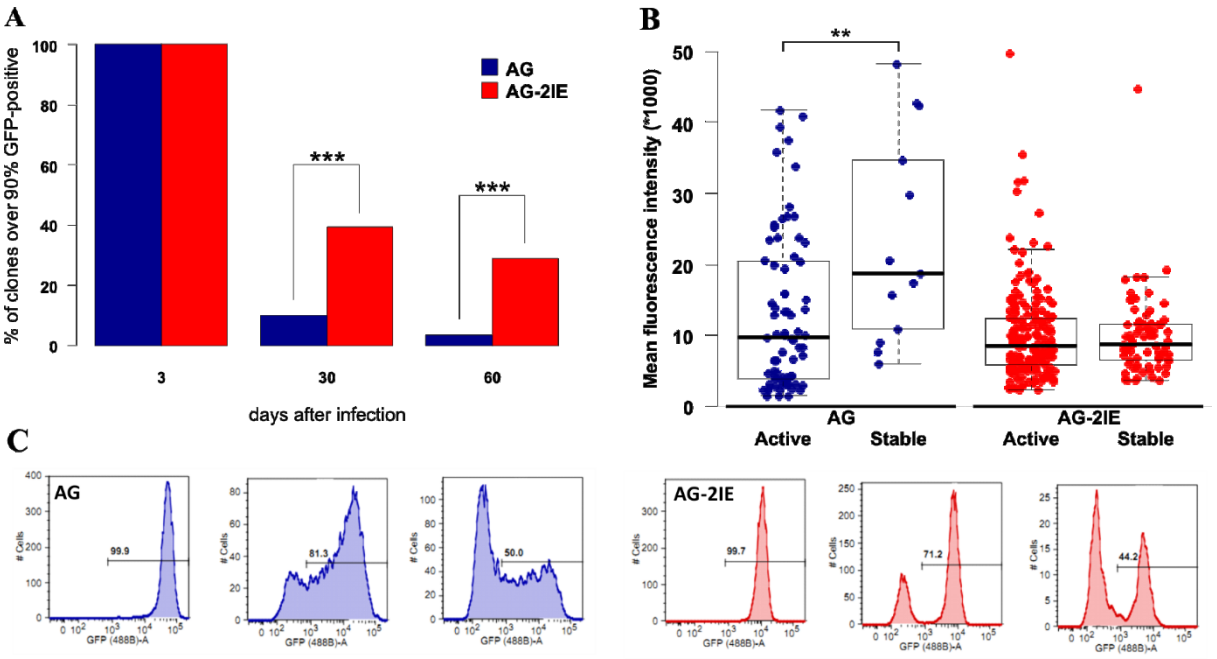

Figure 9. Expression stability of AG and AG-2IE ASLV-derived vectors in human cells. **A.** Percentage of GFP+ST clones at different time-points. At 3 dpi, all cells sorted are GFP+. At 30 and 60 dpi, the percentage depicts the portion of GFP+ST clones of all clones measured after sortStatistical significance was tested by the Fisher's exact test. **B.** MFI of clones expressing GFP at 3 dpi (Active) and GFP+ST clones. Each dot represents MFI of a single clone. Statistical significance was tested by the Wilcoxon test. **C.** Examples of different clonal expression profiles observed during silencing of GFP expression. Bars represent the area of GFP+ cells. MFI – mean fluorescence intensity. * - p value < 0.05 ** - p value < 0.01, *** - p value < 0.001.

Genome-wide mapping and characterization of provirus integration sites

Genomic DNA from individual GFP+ST clones with AG or AG-2IE proviruses was isolated and digested with specific restriction enzymes. The genome-viral DNA junction was amplified using the splinkerette PCR technique (Uren et al. 2009). In addition to individual clones that contained single proviral integrations, we cloned the proviral integration sites in en masse splinkerette PCR from cells that were transduced with either AG or AG-2IE vectors without selection for GFP activity (non-selected control, NS) or sorted for GFP-positivity at 3 dpi (active 3 dpi controls, GFP+^{3dpi}). We sequenced the junction DNA fragments and identified integration sites using BLAT in the human genomic assembly GRCh37, version hg19. In total, after neglecting the small number of equivocal integrations, we identified 90 NS sites of AG proviruses and 82 NS sites of AG-2IE proviruses, 124 sites of GFP+^{3dpi} AG and 63 AG-2IE proviruses, and, finally, 46 sites of GFP+ST AG and 58 AG-2IE proviruses. For better comparison with previous studies, we generated 200 random integration sites by *in silico* targeting the human genome. This set of integration sites was used in all subsequent analyses as a random control. The comparison of the random with NS set of integrations also defines the possible bias of the technique given by the distribution of restriction recognition sequences. We further analyzed provirus integrations from the point of view of targeting transcriptional units (TUs), transcriptionally active TUs in K562 cells, and TSSs.

We observed that the NS proviruses of AG and AG-2IE had integrated within TUs at the frequency of 57 % and 48 %, respectively, which is slightly higher than the 39 % that was observed in the random set, corroborating the slight preference of ASLV integrase for genes as described previously in the studies of Narezkina et al. (2004) and Barr et al. (2005). The percentage of GFP+^{3dpi} proviruses found in TUs had increased to 65 % and 64 % for AG and AG-2IE proviruses, respectively. The enrichment of GFP+ST proviruses within TUs at 60 dpi was even higher, 74 % and 79 % for AG and AG-2IE

proviruses, respectively (Fig. 10A). This data demonstrates that insertion of the IE element into LTR does not influence the integration preference and long-term selection of transcriptionally active proviruses increases the rate of genic/intergenic integration. We can assume that integration into TU increases the chance of the provirus to be transcriptionally active over time and resistant to epigenetic silencWe also analyzed the proportion of insertions oriented sense or anti-sense according to the transcription of targeted TUs (Fig. 10B). Although approximately equal at the beginning of the experiment, the selection for transcriptional activity of AG proviruses led to statistically insignificant preponderance of sense integrations. AG-2IE proviral integrations displayed approximately the same ratio of both orientations.

Next, we analyzed the level of transcription of targeted TUs. We combined transcription data of all TUs targeted by AG and AG-2IE proviruses from publicly available RNAseq data, which was retrieved from the Sequence Read Archive (SRA) for the K562 cell line. According to reads per kilobase per million (RPKM), TUs were divided into five groups according to the transcriptional level – a non-active (NA) group and four quartiles of active TUs (Q1-Q4) (Fig. 10C). In the set of NS integration sites, the integrations into transcriptionally active TUs prevailed, which was in accordance with previous findings (Mitchell et al. 2004, Narezkina et al. 2004, Barr et al. 2005) showing preferential integration of retroviruses into active chromatin. We did not observe any striking difference between AG and AG-2IE proviruses. After short-term selection for GFP+^{3dpi}, the prevalence of transcriptionally active targeted TUs increased to approximately 70 %. Among the AG proviruses, there was an increase in integrations into Q4 TUs, the TUs with the strongest transcription intensity. In contrast, integrations into Q1 TUs, the group comprising weakly transcribed TUs, were underrepresented. This effect of short-term selection was not observed in AG-2IE proviruses. Long-term selection for stable expression of AG proviruses led to strong overrepresentation of transcriptionally active TUs, particularly those of Q4 and Q3. Only 2 % of GFP+ST AG proviruses resided in non-transcribed TUs and 40 % in Q4 TUs. In summary, we have demonstrated that ongoing selection for transcriptional activity of proviruses leads to increased representation of proviruses integrated into transcribed TUs. GFP+ST proviruses are found almost exclusively in the transcribed TUs, particularly in those with highest transcription levels. This can be explained by a protective anti-silencing effect of the genomic environment in transcriptionally active TUs. Insertion of IE elements partly releases this dependence, and some GFP+ST AG-2IE proviruses can also be found in non-transcribed or weakly transcribed TUs.

The epigenetic environment varies along the TUs, which may influence the stability of the provirus expression. Integration close to TSSs has already been documented for MLV (Wu et al. 2003) and correlates with the expression of ASLV proviruses (Plachy et al. 2010, Senigl et al. 2012). We therefore analyzed the position of integration sites within the targeted genes and in relation to the adjacent TSS in our sets of AG and AG-2IE proviruses. First, we analyzed the absolute proviral distance to the closest TSS (Fig. 10D). Short-term selection forming the population of GFP+^{3dpi} proviruses selected for the integrations concentrated around the TSSs with the median distance around 10 kb in both AG and AG-2IE proviruses. After long-term selection, the GFP+ST AG proviruses were found closer to TSSs with a median distance of 2 kb ($p = 8.2^{-6}$, Wilcoxon–Mann-Whitney Rank Sum Test). AG-2IE GFP+ST proviruses were integrated at a median distance of 6.5 kb to TSSs, resembling the integration pattern of shortly selected proviruses ($p = 0.35$, Wilcoxon–Mann-Whitney Rank Sum Test). Taking into account the distribution of integrations along targeted TUs, we calculated the relative proviral distance to TSS (Fig. 10E). Unlike the absolute distance to TSS, the relative distance indicates the distance to a particular TSS belonging to the targeted TU for intra-TU integrations. For inter-TU integrations, the relative distance is equal to the absolute distance to the nearest TSS. The relative proviral distance to TSS showed that with further positive selection of expression, proviruses were concentrated around TSSs with a mild bias for integrations within TUs downstream to TSSs. This bias is most striking for AG GFP+ST proviruses (Fig. 10F), where 70 % of all integrations were found within 15 kb downstream of TSSs, from which one third of the proviruses (33%) were integrated within 2 kb downstream from TSSs (Fig. 10G). On the other hand, AG-2IE GFP+ST proviruses that have spread into more distal parts

downstream from TSSs resembled the integration pattern that was observed with shortly-selected proviruses. We conclude that proviruses selected for stable expression are predominantly found close to TSSs. Obviously, the vicinity to TSS is favorable for provirus transcription and protects the proviruses from transcriptional silencProviruses that are equipped with protective IEs are less dependent on the need to be located within close proximity of TSS.

CpG islands, sequences rich for CG dinucleotides, are an important part of vertebrate genomes (reviewed by Deaton and Bird (2011)). We analyzed the distribution of proviral integrations in relation to the closest CpG islands in a similar way as TSSs (Fig. 11A). The random set of integration sites displayed a longer absolute distance to CpG islands than to TSSs (medians of 104 kb versus 62 kb, respectively), which reflected the fact that CpG islands were less frequent in the genome. The distance to the CpG islands showed a pattern similar to that observed with distances to TSSs, i.e., with further selection for expressional stability the proviruses of both vectors were found closer to the CpG island. The correspondence of the distances to TSSs and CpG islands shows that most of TSSs that are located close to active and stably active proviruses are associated with CpG islands (Fig. 11B).

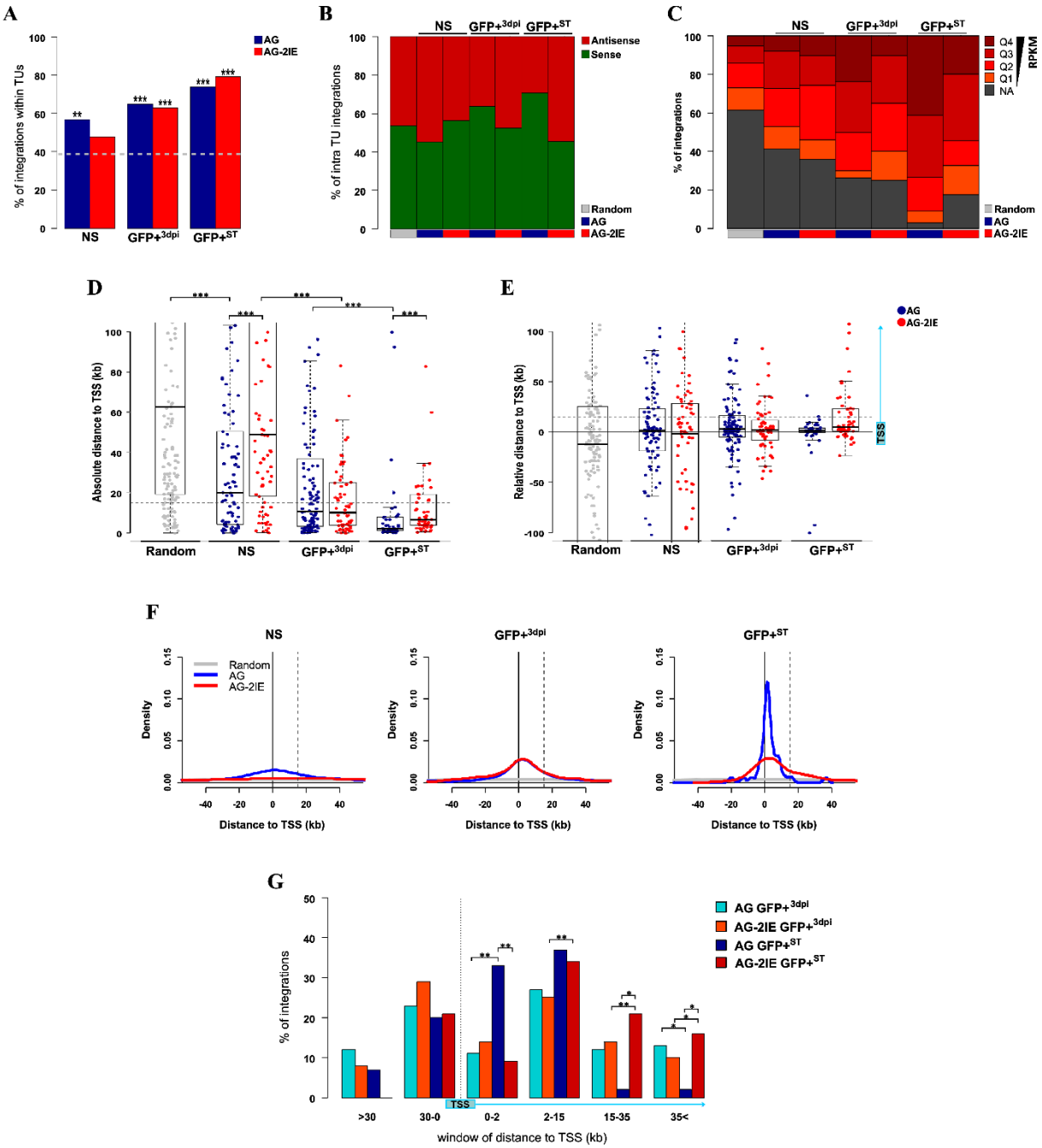

Figure 10. Genomic features at proviral integration sites. **A.** Frequency of TU targetGray line represents the frequency of random positions. Statistical significance was tested by Fisher's exact test. Asterisks mark the significance of difference against a set of random positions. No statistical significance was observed between selection categories of the same vector or between vectors inside the selection category **B.** Relative orientation of proviruses in targeted TUs. The sense category contains proviruses with same orientation of transcription as targeted TUs. **C.** Transcriptional activity of integration targeted TUs. TUs were divided into five categories according to RPKM with the NA group containing TUs with mean RPKM lower than 1 and TUs with no match in the database and Q4 group containing TUs with highest mean RPKM. Frequencies of proviral integration sites found in RefSeq Genes belonging to the categories are depicted. **D.** Boxplot representation of distribution of distances of the proviral integration site to the nearest TSSs. Statistical significance was tested by Wilcoxon test. **E.** Boxplot representation of distribution of proviral integration site distances around TSSs. Negative values represent proviral integration sites outside TUs – distances represent the distances to the nearest TSSs. Positive values represent proviral integration sites inside TUs – distances represent the distances to the TSSs of targeted TUs. **F.** Density representation of proviral integration sites around TSS like in panel **E. G.** Frequency of proviral integration sites in the windows covering the marked distance to the TSSs. Statistical significance was tested by

Fisher's exact test. TU – transcriptional unit, TSS – transcriptional start site. * - p value < 0.05, ** - p value < 0.01, *** - pvalue < 0.001.

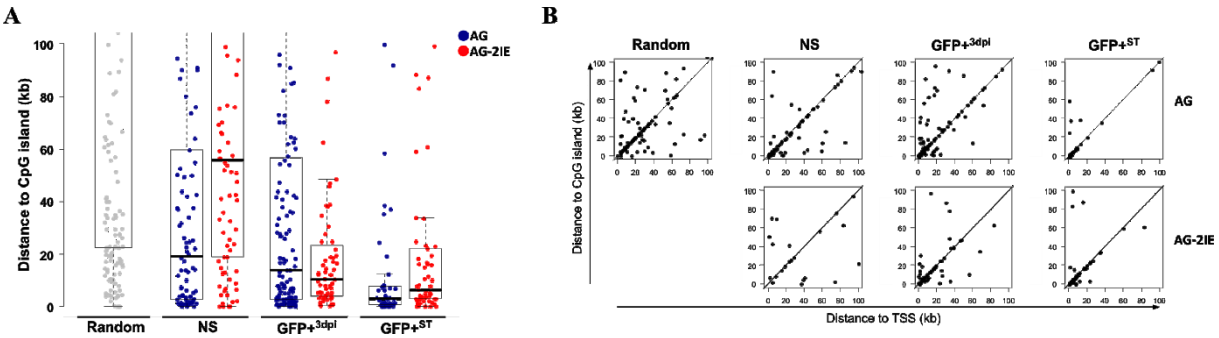

Figure 11. Distribution of proviral integration sites relative to CpG islands. **A.** Boxplot representation of the distribution of distances of proviral integration sites to the nearest CpG island. **B.** Correlation of the distance of proviral integration sites to the nearest TSSs and CpG islands.

Active histone modifications at the sites of integration in GFP+ST proviruses

As described previously (Senigl et al. 2012), active proviruses are overrepresented in the regions enriched in H3K4me3 markers, which are characteristic for active TSSs. To obtain better insight into the epigenomic landscape of provirus integration and expression, we analyzed the distribution of our proviral integrations with respect to their distance to the peaks of different histone modifications. In Fig. 12, we present results of the analysis of histone modifications that are associated with both, transcriptionally active or suppressive chromatin. Methylation on H3K4 (H3K4me1/2/3) and acetylated histones (H3K9ac, H3K27ac) are characteristic of the regions occupied by active promoters and enhancers. Proviruses of both AG and AG-IE were found closer to the peaks of these modifications as the selection for expression stability went on (Fig. 12A). The median distance of the AG GFP+^{3dpi} to the peaks of H3K4me1/2/3, H3K9ac, H3K27ac ranged from 5 to 6 kb. AG GFP+ST proviruses were found in close proximity to these peaks, with the median distances ranging from 0.3 to 2 kb. GFP+ST proviruses of both vectors were significantly closer to H3K4me2/3 and H3K9ac than their GFP+^{3dpi} counterparts. However, for the H3K4me1 and H3K27ac modifications, the hallmarks of enhancer regions, the distance did not change significantly between GFP+^{3dpi} and GFP+ST. Taking into account only AG proviruses, the most significant shift of median distance was observed for the H3K4me3 modification ($p = 1.3^{-4}$, Wilcoxon-Mann-Whitney Rank Sum Test).

Other histone modifications correlating with active chromatin were also analyzed. H3K79me2 and H4K20me1 (Fig. 12B) are more lateral modifications of histones that can be found downstream of TSSs of active TUs (Wang et al. 2008). However, the role of the modifications in transcriptional activation or repression is unclear. According to the presence of H3K79me2 downstream of active TSSs, AG, and AG-2IE proviruses showed a gradual decrease of the median distance to the modification peaks with progressing selection for expression. AG and AG-2IE GFP+ST proviruses accumulated at a median distance of 0 and 5 kb to H3K79me2 peaks, respectively. AG GFP+ST proviruses also showed a significant decrease in the distance to H4K20me1 compared to AG GFP+^{3dpi}. Moreover, H4K20me1 is the only histone modification we examined to which AG GFP+ST proviruses were significantly closer than AG-2IE GFP+ST proviruses. The distance of GFP+ST proviruses to these histone modifications shows that at least at the particular environment, H3K79me2 and H4K20me1 form transcriptionally permissive chromatin.

Other epigenomic features that correlate with transcription are histone isoforms, some of which are also enriched at specific and narrow genomic loci. We examined the distribution of proviral distance to peaks

of the H2A.Z histone variant. H2A.Z has been found in association with active TSSs and enhancers occupied by various transcription factors (Soboleva et al. 2011). Both AG and AG-2IE NS proviruses were already found close to the peaks of H2A.Z at a mean distance of ca 20 kb. Even after short-term selection for proviral expressional activity, GFP+^{3dpi} proviruses of both AG and AG-2IE accumulated proviral integrations close to H2A.Z peaks (median of ca 6 kb), but no decrease in distance was observed in the group of GFP+ST proviruses (Fig. 12A). This suggests that ASLV might have a slight preference for H2A.Z-enriched areas and that integration to these areas might be important for the transcriptional activity of proviruses of either vector. However, other features seem to be important for the long-term transcriptional stability.

We also examined the distance of proviruses to the modifications characteristic for transcriptionally suppressed chromatin (H3K9me3, H3K27me3). The peaks of suppressive histone modifications H3K9me3, H3K27me3 were found at a median distance of 40 to 50 kb from the AG proviral integrations, respectively, and the long-term selection for expressional activity did not lead to any accumulation of proviruses observed closer to the histone modifications associated with active transcription (Fig. 12C).

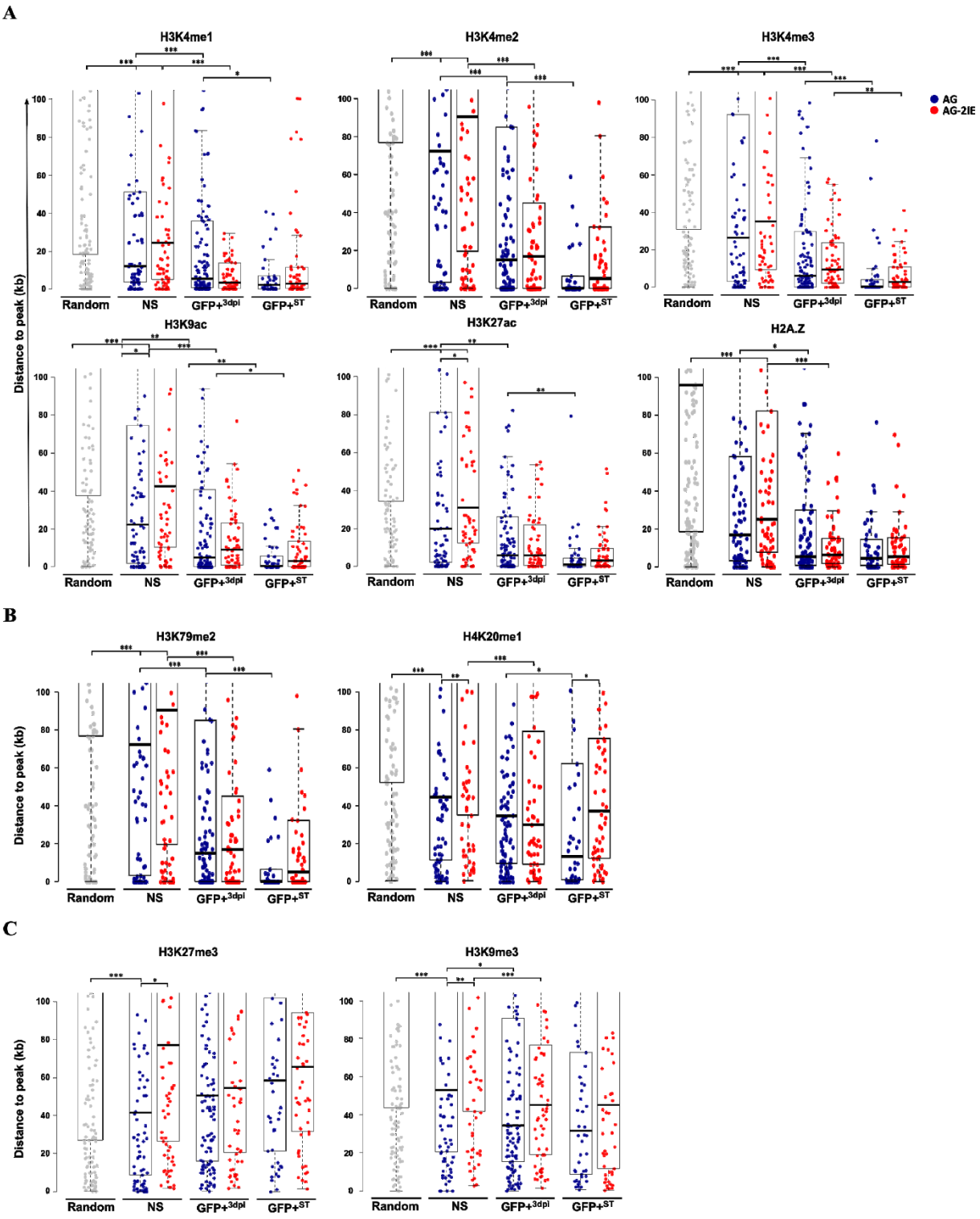

Figure 12. Distribution of proviral integration sites to peaks of histone modifications. Boxplot representation of the distribution of distances of proviral integration sites to the nearest peak of histone modification representing **A.** active regulatory segments, **B.** putative active chromatin, or **C.** repressive chromatin. Statistical significance was tested by Wilcoxon test. * - p value < 0.05, ** - p value < 0.01, *** - p value < 0.001.

Proviral integration and transcriptional stability in functional chromatin segments

The integrative combination of epigenetic marks has been used for genome-wide annotation of functional chromatin states and non-coding functional elements in the human genome across multiple cell types. Two independent chromatin state annotation algorithms, ChromHMM (Ernst and Kellis 2010) and Segway (Hoffman et al. 2012), based mostly on the results from ChIPseq assays, served for ENCODE-wide annotation of functional chromatin segments and regulatory elements (Hoffman et al. 2013). To describe proviral integration sites with regard to the function of their chromatin regions, we calculated the percentages of integrations within chromatin segments categorized by the ChromHMM (Fig. 13) database.

The randomly generated integration sites showed the approximate proportion of certain functional chromatin segments in the human genome with the majority being non-transcribed heterochromatin or polycomb-repressed chromatin. The proportion of targeted promoters and enhancers was found to be less than 10 % in the random set of integration sites (Fig. 13A). Both AG and AG-2IE experimentally determined NS proviruses preferred the merged active chromatin segments at the expense of non-transcribed ones ($p = 6.1 \times 10^{-6}$ and $p = 6.4 \times 10^{-5}$ for AG and AG-2IE, respectively, Fisher's Exact Test for Count Data, Fig. 13B). Even the short-term selection of GFP+^{3dpi} proviruses further increased the proportion of active chromatin among the targeted segments ($p = 9.5 \times 10^{-4}$ and $p = 2.7 \times 10^{-6}$ for AG and AG-2IE, respectively. Fisher's Exact Test for Count Data). Selection for GFP+ST proviruses did not lead to any increase of proviruses located in active chromatin segments, but resulted in selection of proviruses in active regulatory segments, i.e., promoters and enhancers, where we observed more than 50 % of stable integrations ($p = 1.9 \times 10^{-4}$ and $p = 4.0 \times 10^{-2}$ for AG and AG-2IE, respectively, Fisher's Exact Test for Count Data). The most striking observation was the enrichment GFP+ST AG proviruses in promoters (ca. 40%, $p = 4.0 \times 10^{-5}$, Fisher's Exact Test for Count Data).

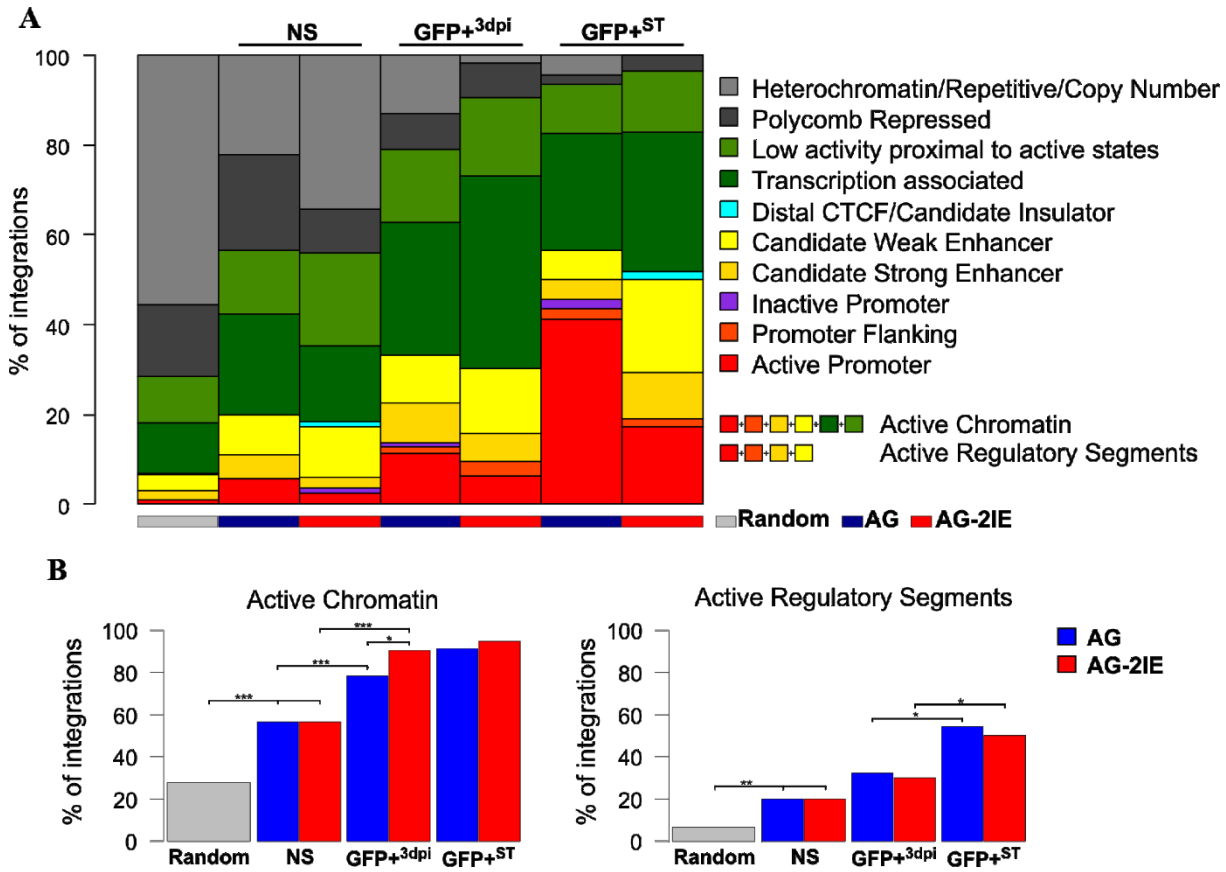

Figure 13. Proviral integration site frequency in marked chromatin state segments. **A.** Frequency of proviral integration sites in marked chromatin segments. **B.** and **C.** Frequency of proviral integration sites in joined chromatin segments representing **B.** Active chromatin or **C.** Active regulatory segments. Segments were joined according to the legend presented in panel **A.**. Statistical significance was tested by Fisher's exact test. * - p value < 0.05, ** - p value < 0.01, *** - p value < 0.001.

Proviral expression and integration sites of GFP+ST proviruses of HIV-1 and MLV

Data on ASLV-derived vectors reveal that in the suppressive environment of mammalian cells, GFP+ST proviruses tend to be integrated close to active TSSs. Modification of ASLV LTR with short regulatory elements that stabilize expression of the original vector released GFP+ST proviruses from TSS-proximal areas. However, proviruses of the modified vector selected for long-term expressional activity show non-random distribution of integration sites skewed toward active TUs and close to enhancers. The preference toward the features such as TSSs, enhancers, or active TUs is known for natural integration of some mammalian retroviruses such as lentiviruses or γ-retroviruses (Elleder et al. 2002, Schroder et al. 2002, Wu et al. 2003, Mitchell et al. 2004, De Ravin et al. 2014, LaFave et al. 2014). To look for the common features of the integration sites of GFP+ST proviruses across the retroviral genera, we compared the distribution of integration sites of proviruses of ASLV, HIV-1, and MLV.

Stability of proviral expression

To further extend the data on the distribution of stably active proviruses, we used vectors derived from HIV-1 and MLV (Fig. 14) in the same clonal assay used in previous study with ASLV-derived vectors (Fig. 8C). Minimal non-replicative retroviral vectors encoding GFP under the transcriptional control of LTR were used. The HIV-1 minivirus also contained the functional *tat* gene, which is necessary for normal expression from HIV-1 LTR (Jordan et al. 2001). VSV-G-pseudotyped miniviral vectors were used to transduce the K562 cell line, single-cell sorted for GFP+ cells at 3 dpi and expanded to clonal populations, which were FACS-examined at 30 and 60 dpi. Data obtained with HIV-1 and MLV vectors were compared to those obtained previously with the AG vector, here marked as ASLV dataset.

The percentages of GFP+ cells were determined in 378 HIV-1- and 239 MLV-transduced clones at 30 dpi. Unlike ASLV that was effectively silenced, more than 80 % of HIV-1- and MLV-transduced clones (262 and 239, respectively) maintained stable expression at 30 dpi (Fig. 15). The clones with stable provirus expression 30 dpi were cultured for an additional 30 days and the percentage of GFP+ cells was re-calculated at 60 dpi. Whereas almost all MLV-transduced clones (202) maintained provirus expression, the numbers of GFP+ST clones in the HIV-1 sample further decreased to 49 % (136, Fig. 15). Thus, MLV provirus expression was deemed to be long-term stable. HIV-1 provirus expression was observed to be stable during short cultivation (30 dpi), but displayed gradual late silencing when cultured for a longer period of time. In any case, both vectors were one order of magnitude more stable than the vector derived from ASLV.

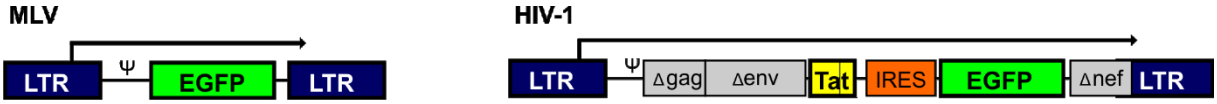

Figure 14. Graphical representation of retroviral vectors. Minimal retroviral vectors derived from MLV and HIV-1 and expressing the EGFP marker under the control of retroviral LTRs are shown. LTR – long terminal repeat, Ψ – packaging signal, EGFP – enhanced green fluorescent protein, Δgag – *gag* gene with deletion, Δenv – *env* gene with deletion, Tat – double exon form of transactivator *tat* gene, IRES – internal ribosomal entry site. Arrow marks the transcription of provirus.

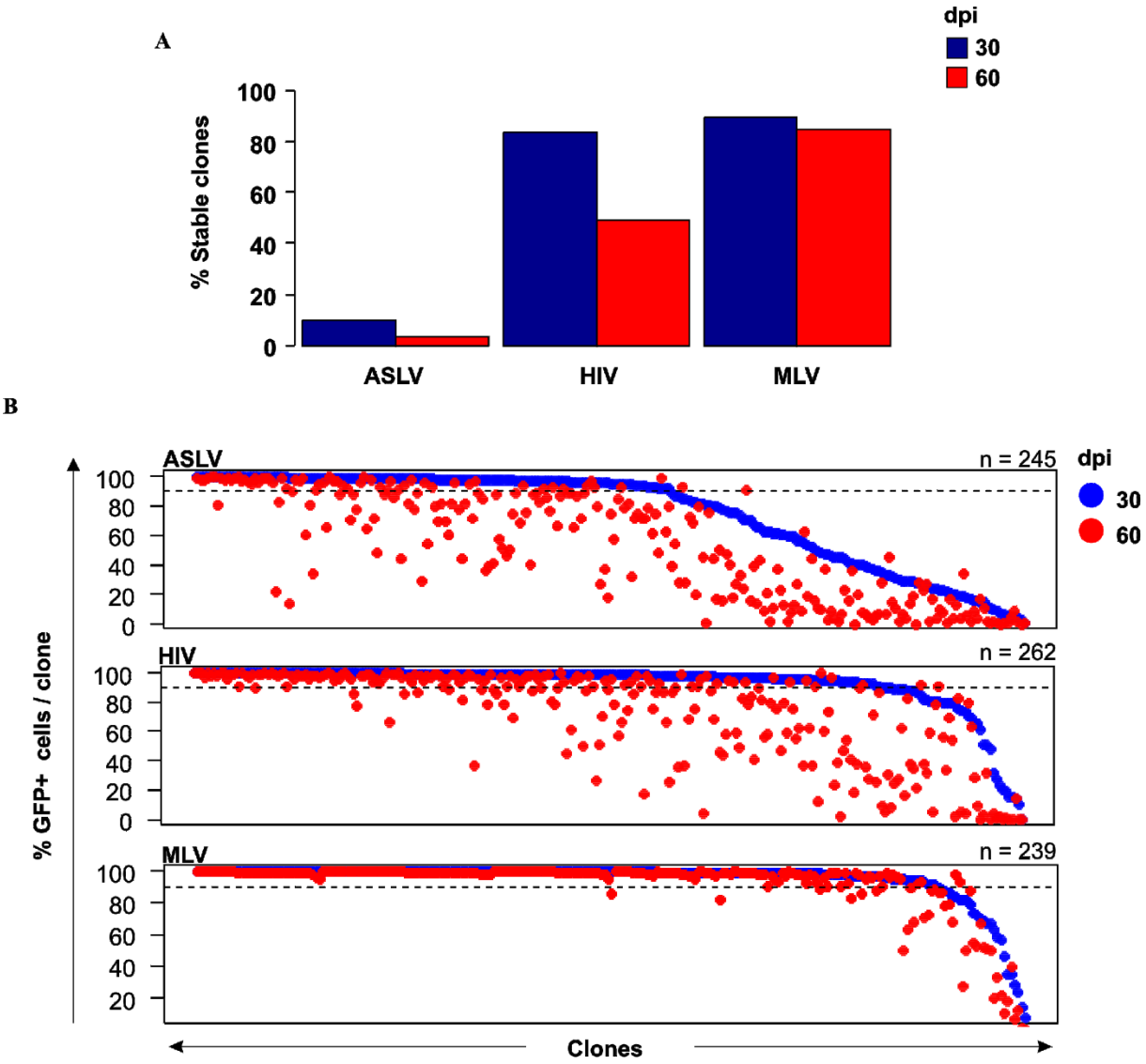

Figure 15. Comparison of expression stability of ASLV, MLV, and HIV-1 vectors. **A.** Frequency of GFP+ST clones. Frequency is counted as a portion of all clones gained after the sorting that contained ≥ 90 % GFP+ cells at a given time-point. **B.** Fraction of GFP+ cells in clones at given time-points. Clones are lined up at x-axis next to each other by the fraction of GFP+ cells at 30 dpi (blue dots). A red dot represents the fraction of GFP+ cells of a particular clone at 60 dpi. Values representing one cellular clone are aligned in one column. The dashed line represents 90 % of GFP+ cells, which is the threshold for GFP+ST clones. n – number of cellular clones in the chart.

Integration and matched random control site generation

GFP+ST clones derived from HIV-1, and MLV-transduced cells were subjected to splinkerette PCR in order to amplify and sequence the provirus integration sites. We identified 45 HIV-1 and 32 MLV integration sites, which we compared to 46 ASLV integration sites and sets of *in silico* prepared random sites. This time, each integration site sample was given its set of random controls, which we here refer to as the uniquely mapped matched random controls (umMRC). For each integration site in the sample, three umMRCs were generated. These three random sites match particular integration distances by the distance to the nearest recognition sites of a given restriction enzyme used in the integration site cloning protocol. The genomic sequences spanning the range given by the recognition site and the random site had to fulfill the criteria for uniquely mapped sequences (see Materials and Methods section).

Gene targeting and distance to TSS

First, we analyzed the frequency of integration into transcription units and the orientation of proviruses relative to transcription of targeted RefSeq Genes (Fig. 16A). HIV-1 is known to target active genes (Elleder et al. 2002, Schroder et al. 2002), while MLV was not observed to possess any preference for genes (Wu et al. 2003, De Ravin et al. 2014, LaFave et al. 2014). HIV-1 proviruses with stable expression were found in RefSeq Genes with the frequency of 80 %, which was the same as the frequency of ASLV GFP+ST proviruses found in RefSeq Genes. Both samples displayed a significant increase compared to umMRCs. On the other hand, the frequency of MLV GFP+ST proviruses found in RefSeq Genes was about 60 %, and thus lower than that of ASLV and HIV-1 and comparable to the respective umMRCs. Interestingly, the orientation of proviruses relative to the direction of targeted RefSeq Gene transcription showed different patterns for all three vectors (Fig. 16B). Compared to umMRC, ASLV GFP+ST proviruses showed a preponderance of proviruses with sense orientations to RefSeq Gene transcription ($p = 0.0225$, Fisher's Exact Test for Count Data). HIV-1 proviruses displayed an equal proportion of proviruses in sense or antisense orientations to endogenous transcription. The majority of MLV proviruses were found in antisense orientation compared to that of targeted RefSeq Genes, albeit giving a weak significance compared to umMRCs ($p = 0.0489$, Fisher's Exact Test for Count Data).

Next, we counted the distance of GFP+ST proviruses to the TSSs of the RefSeq Genes (Fig. 16C). All groups of proviruses accumulated significantly closer to TSSs in comparison to respective umMRCs, with medians of distances well below 20 kb. We also checked the distribution of proviruses around TSSs (Fig. 16D). The distribution of GFP+ST proviral integration sites of ASLV and MLV was observed to be centered close to TSSs, while HIV-1 GFP+ST proviruses were centered more inside gene bodies, in accordance with the HIV-1 preference for active TUs. We also observed a similar accumulation close to CpG islands for ASLV and MLV, and, to a lesser extent, for HIV-1 (Fig. 16E). While ASLV and MLV proviruses distribution tended to center in close proximity of the TSSs, HIV provirus distribution centered at a distance of 12 kb downstream to TSSs.

GFP+ST proviruses of ASLV, HIV, and MLV were thus shown to differ in their distribution toward the gene-associated features. ASLV proviruses accumulated inside RefSeq Genes close to TSSs and mostly in sense orientation to the RefSeq Gene transcription. GFP+ST proviruses of HIV-1 as the ones of ASLV also accumulated inside RefSeq Genes, but in longer distances to TSSs with no bias for either intragenic orientation observed. MLV GFP+ST proviruses, on the other hand, showed no overrepresentation inside RefSeq Genes, and those found in RefSeq Genes were found mostly in the anti-sense orientation towards the targeted RefSeq Gene transcription. However, MLV and ASLV GFP+ST proviruses shared striking accumulation at regions proximal to TSSs, although their distribution around TSSs differed.

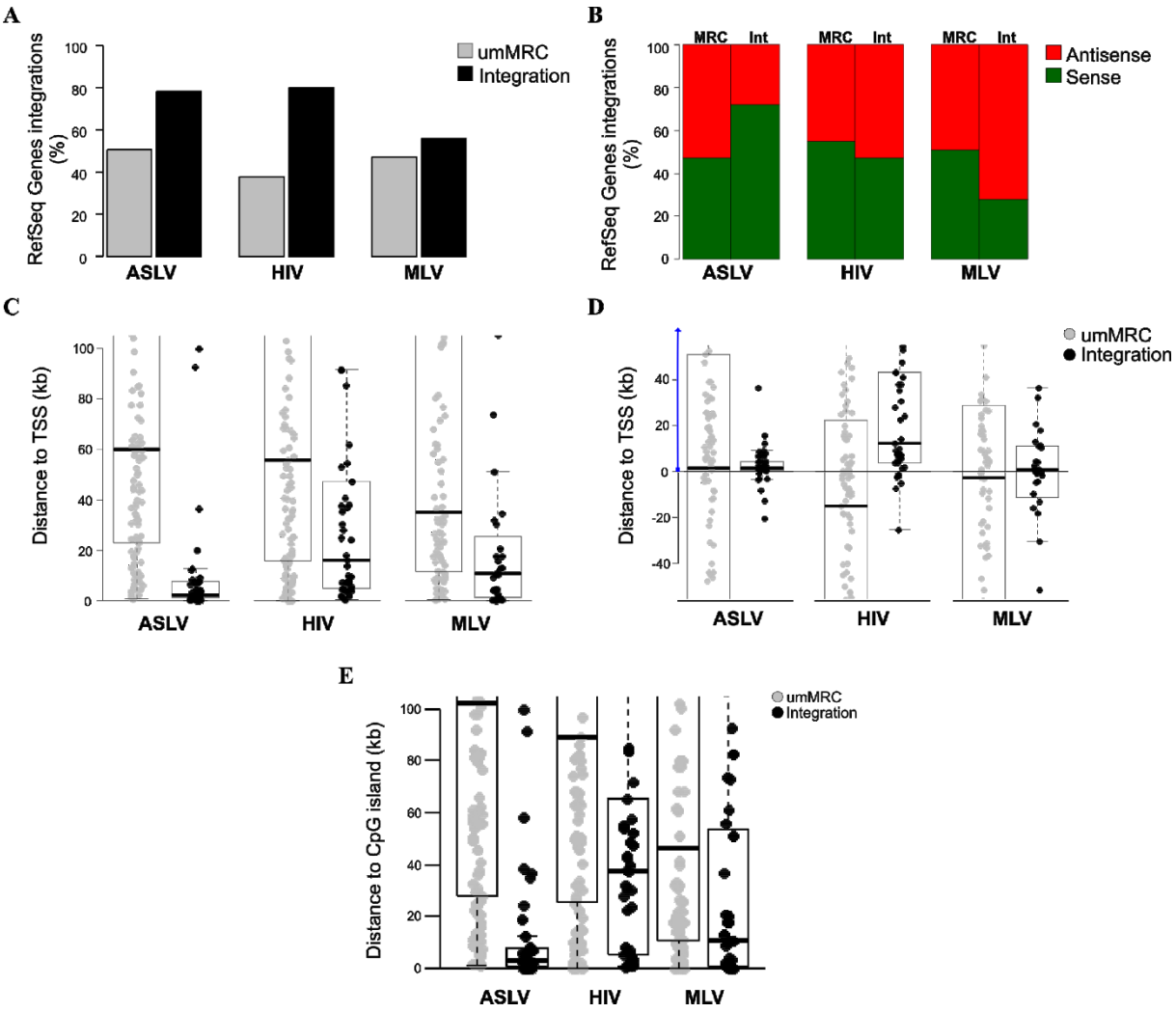

Figure 16. Distribution of proviral integration sites according to genomic features. **A.** Frequency of proviral integration sites in RefSeq Genes. **B.** Relative proviral orientation according to the transcription of targeted RefSeg Gene. Sense orientation marks proviruses with the same orientation as is the orientation of the targeted RefSeq Gene. **C.** Boxplot representation of the distribution of distances of proviral integration sites to the nearest TSSs. **D.** Boxplot representation of the distribution of proviral integration site distances around TSSs. Negative values represent proviral integration sites outside TUs – distances represent the distances to the nearest TSSs. Positive values represent proviral integration sites inside TUs – distances represent the distances to the TSS of the targeted TU. **E.** Boxplot representation of the distribution of distances of proviral integration sites to the nearest CpG island.

GFP+ST proviruses associate with active genes

The environment that is permissive for provirus expression can be correlated with the activity of targeted TUs. Therefore, we used a subset of RefSeq Genes that exhibited high transcriptional activity and whose TSSs associated with features defining active promoters: peaks of H3K4me3, Tss chromatin segment defined by ChromHMM (Ernst and Kellis 2012), or peaks of transcription defined by cap analysis of gene expression (CAGE, (Kanamori-Katayama et al. 2011)).

Approximately 20 % of umMRCs targeted RefSeq Genes defined here as active showing about 50 % decrease in the frequency compared to targeting of all RefSeq Genes. On the other hand the frequencies of proviral integrations in active RefSeq Genes were comparable to the frequencies of targeting of all RefSeq Genes (Fig. 17A). Notably, while no significant difference of all RefSeq Genes targeting was observed between MLV proviral integrations and respective umMRCs, the MLV proviral targeting of

active RefSeq Genes (Tss+ RefSeq Genes) was significantly increased ($p = 0.0148$, Fisher's Exact Test for Count Data) compared to umMRCs (Fig. 17A).

We observed the same effect at the level of distance to the TSSs of active RefSeq Genes (Fig. 17A). The distance of umMRCs to TSSs of active RefSeq Genes increased in comparison to distance to TSSs of all RefSeq Genes, whereas proviral integration sites were found to be of similar distances to TSSs of active RefSeq Genes as to TSSs of all RefSeq Genes.

To address the activity of TUs targeted by integrations of GFP+ST, we used publicly available RNAseq data, as described in previous chapters. Briefly, TUs were divided into activity groups according to their mean read per kilobase per million (RPKM) mapped reads of RNA-seq datasets. In correspondence to the previous selection for active RefSeq Genes, RNA-seq data showed that the targeted TUs mostly exhibited transcriptional activity with a RPKM ≥ 1 (Fig. 17C).

These results demonstrated that the proviruses selected for stable transcriptional activity accumulated near the TSSs of active TUs, likely because of their active transcription-associated chromatin environment, which is also permissive for provirus expression.

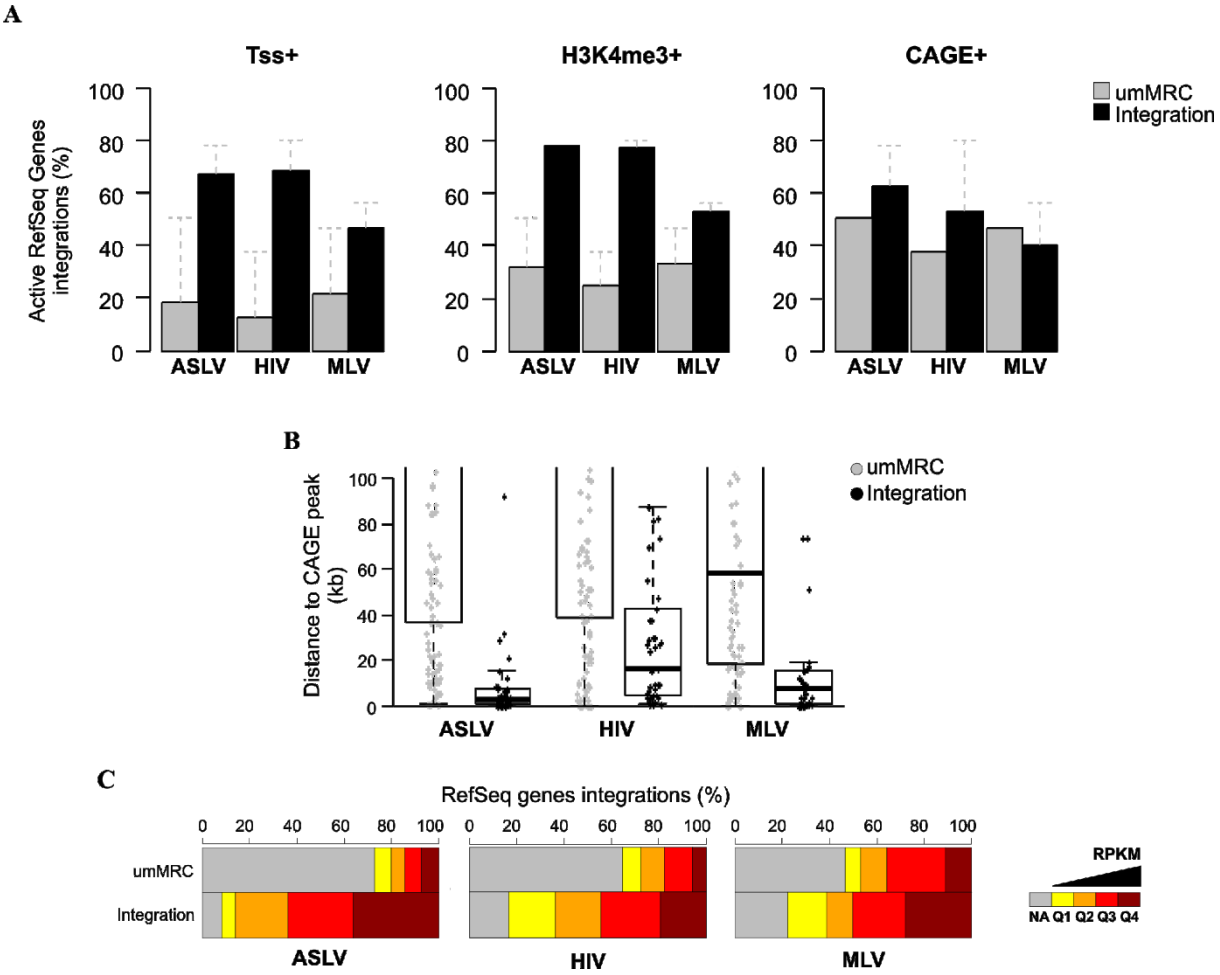

Figure 17. Association of proviral integration sites with active TUs. **A.** Frequency of proviral integration sites in RefSeq Genes whose TSSs were found in the Tss chromatin segment (Tss+), in the peak of H3K4me3 (H3K4me3+), or within the 500 bp from the CAGE peak (CAGE+). Dashed antennas mark original frequency of targeting all RefSeq Genes. **B.** Boxplot representation of the distribution of distances of proviral integration sites to the nearest CAGE peaks marking sites of active transcription initiation. **C.** Transcriptional activity of integration of targeted TUs. RefSeq Genes were divided into five categories according to RPKM, with the NA group containing TUs with mean RPKM lower than 1 and TUs with no match in the database and Q4 group containing TUs with highest mean RPKM. Frequencies of proviral integration sites found in RefSeq Genes belonging to the categories are depicted.

GFP+ST proviruses associate with active chromatin markers

Epigenetic features of the chromatin environment at the site of integration were first described by comparing the distances of the proviruses and umMRC to peaks of eleven histone modifications (Fig. 18A) defined for the K562 cell line by the ENCODE project. HIV-1 and MLV GFP+ST proviruses accumulated in short distances to the peaks of epigenetic modifications that are associated with active chromatin but not with markers of heterochromatin. A common feature of all groups of retroviral vectors was the short median distance to the peaks of H3K4me1/3, H3K9ac and H3K27ac, the trend observed previously with ASLV GFP+ST proviruses.

HIV-1 proviruses displayed a more relaxed pattern showing the median distance to the peaks of histone modifications that was farther away when compared to ASLV and MLV. A similar trend in the median distance of HIV-1 GFP+ST proviruses being more distant to histone modification peak than those of ASLV and MLV was also observed for H2A.Z, H3K79me2, and the peaks of H3K36me3. All of the three marks have distinct distribution with H2A.Z peaking at open chromatin of TSSs and enhancers, while the latter two appear in gene bodies with H3K79me2 peaking close to active TSSs.

An interesting pattern was observed with distances toward the modification of H4K20me1, where GFP+ST proviruses of ASLV showed the median distance below 20 kb, while the HIV-1 and MLV GFP+ST provirus median distance reached 70 kb. This data showed that H4K20me1 might be associated with active TSSs, but might not be a general feature of active TSSs or enhancers.

Together, analysis of the distance to histone modifications data showed that proximity to the sites enriched for marks characteristic of active TSSs and enhancers is a common feature of GFP+ST proviruses of ASLV, HIV-1, and MLV. HIV-1 GFP+ST proviruses, however, showed to be generally more distant to such sites than ASLV and MLV GFP+ST proviruses.

GFP+ST proviruses associate with active regulatory segments

As a next step in defining the chromatin environment at the sites of the GFP+ST integration, we calculated the distances of the GFP+ST proviruses to the chromatin segments that were defined by Ernst and Kellis (2010) for the K562 cell line. We calculated the distances of GFP+ST proviruses to all 25 chromatin states available (Fig. 18B, for explanation of chromatin state mnemonics see Table 2 in the Materials and Methods section). In agreement with previous results, the proviruses of all groups analyzed were found close to the chromatin states associated with active TSSs. Even though previously GFP+ST HIV-1 proviruses showed significantly longer distances to TSSs than GFP+ST proviruses of MLV and ASLV, these differences were lost when distances to the chromatin states flanking active TSSs and promoters (TssF, PromF) were analyzed. More interestingly, GFP+ST proviruses of ASLV, MLV, as well as HIV-1 were found in close proximities to enhancer-associated chromatin states.

Together, the analysis of the epigenomic and functional landscape of GFP+ST showed that regardless of the provirus origin, stable expression of the provirus associates with the proximity to the genomic loci driving genomic transcription. Although GFP+ST proviruses of ASLV and MLV were found closely associated to the features enriched at active TSSs, the active TSS-proximal sites seem not to be general features of GFP+ST proviruses of distinct genera, as HIV-1 GFP+ST proviruses were found further away. However, GFP+ST proviruses of all three groups were found to harbor enhancer-proximal loci, meaning that the enhancer proximity is a general feature of GFP+ST proviruses.

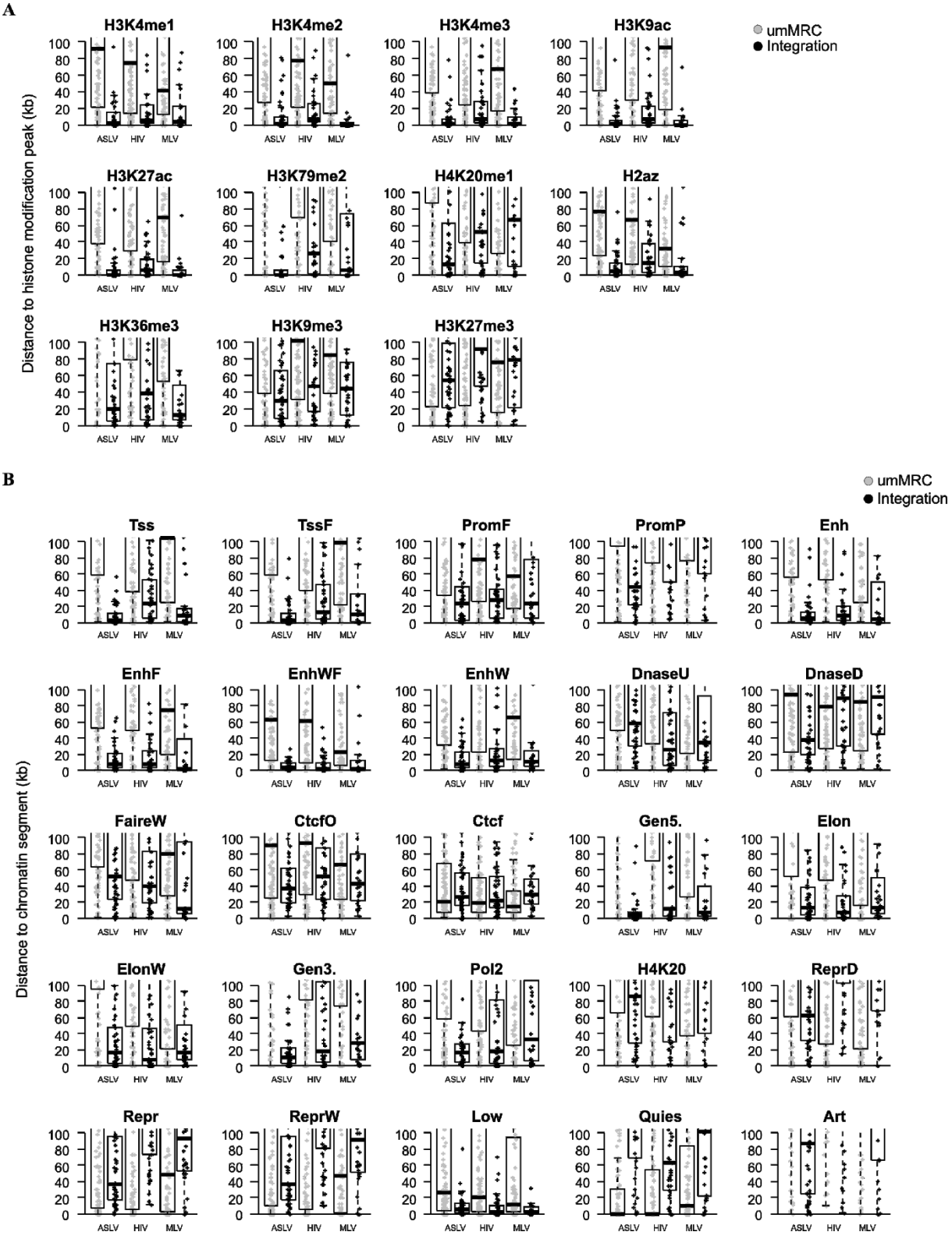

Figure 18. Distribution of proviral integration sites according to histone modification peaks and chromatin segments. **A.** Boxplot representation of the distribution of distances of proviral integration sites to the nearest peaks of histone modifications or histone variants peaks. **B.** Boxplot representation of the distribution of distances of proviral integration sites to the nearest chromatin segment of particular type. For explanation of chromatin segment mnemonics, see the Materials and Methods section.

GFP+ST proviral enhancer proximity is not a function of active gene targeting

As shown in Fig. 17, almost all the GFP+ST proviruses of HIV-1, MLV, and ASLV were found inside or very close to active chromatin segments, active TSSs and enhancers. Furthermore, the majority of GFP+ST ASLV and HIV-1 proviruses were found in genes, which is the result of preference of integration in the case of HIV-1 (Elleder et al. 2002, Schroder et al. 2002, Mitchell et al. 2004) or the result of selection for GFP+ST proviruses in the case of ASLV. Therefore, we sought to investigate whether the proximity to active TSSs and enhancers can be reached by preferential targeting of active genes. For this purpose, another level of matching for umMRC was added by selecting umMRCs that show the same frequency of targeting and a similar distance to the active (Tss+) RefSeq Genes as integration sites of GFP+ST proviruses (Fig. 19A). The new group of matched random controls was then called active gene-matched umMRC (agMRC, see Materials and Methods).

To stress the changes brought about by development of agMRCs, a plot showing the statistical significance of the differences in targeting or distance to genomic or epigenomic features between real integrations and agMRC was constructed (Fig. 19B). As a result of matching, the data showed no significant difference of active RefSeq Genes targeting between GFP+ST proviruses and agMRCs. For ASLV, the distances to most histone modifications differed significantly between agMRC and GFP+ST proviruses, while differences in distances to some active chromatin segments were insignificant. Most importantly, GFP+ST proviruses of ASLV were significantly associated with active TSSs and strong enhancers (Fig. 19C). For the HIV-1 and MLV datasets, the trend depicting the loss of significant differences between GFP+ST proviruses and agMRC in the set of active chromatin segments and the preservation of a significant difference for active TSSs and enhancers was notable. Interestingly, the data for HIV-1 showed that GFP+ST proviruses were significantly associated to active TSSs and enhancers as well as to the histone marks characteristic of those regulatory segments – H3K4 methylation and histone acetylation.

Together, these data show that when integration sites of GFP+ST proviruses are compared to their matched controls that mimic the active gene targeting of the proviruses, the close proximity to the features that are characteristic of active TSSs and enhancers are preserved as a hallmark of GFP+ST proviruses of ASLV, HIV-1, and MLV.

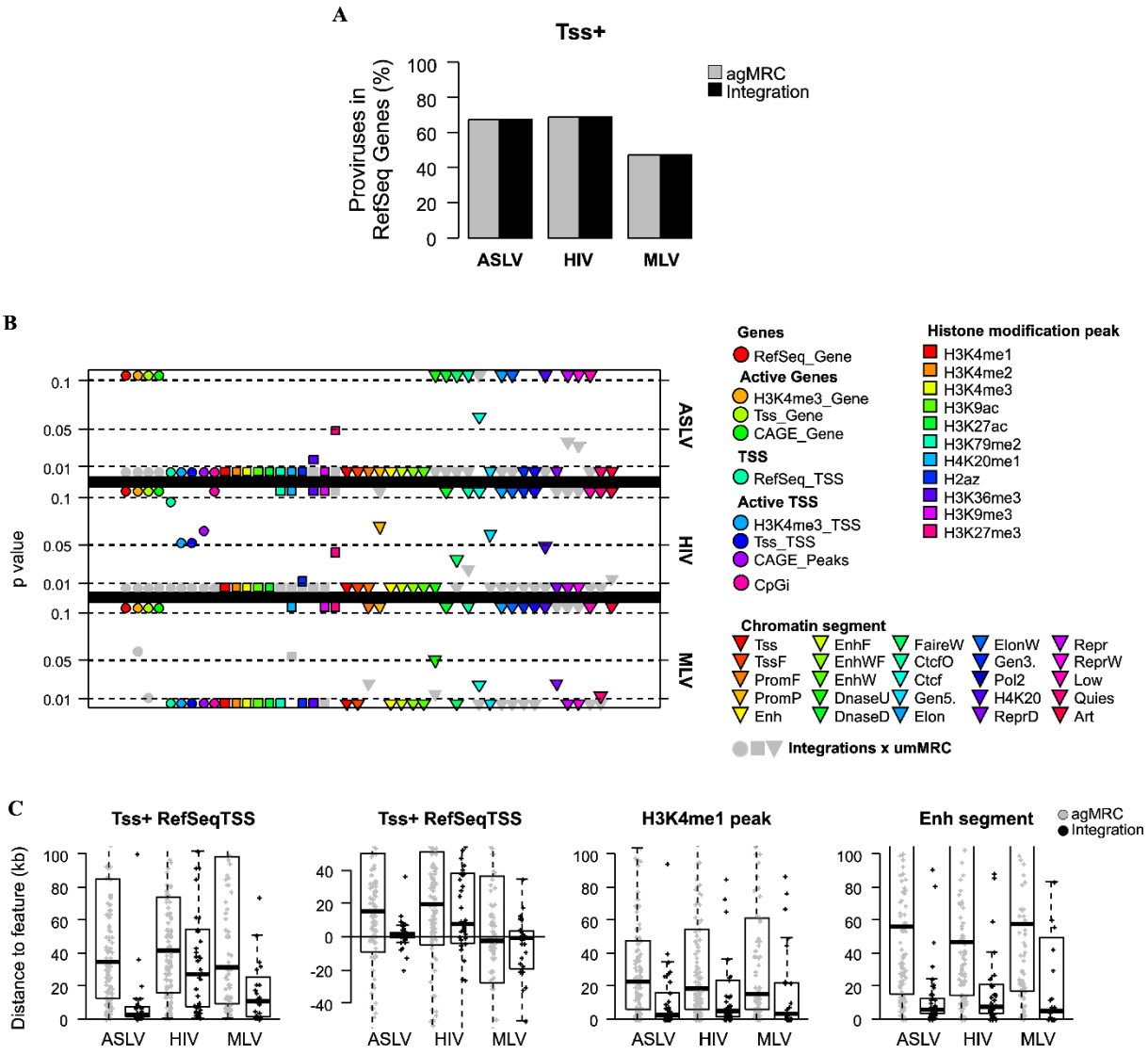

Figure 19. Comparison of proviral integration sites and active gene-targeting matched random controls (agMRCs). Random positions that fulfill the criteria for umMRCs and match proviral integration sites in frequency of targeting and distance to Tss+ RefSeq Genes were generated and called agMRCs. **A.** Presentation of equal frequency of GFP+ST proviral integration sites and agMRCs in Tss+ RefSeq Genes (for comparison with umMRCs, see Fig. 17A). **B.** Statistical p-values of differences between agMRC and integration sites are represented by colored circles (genes), squares (peaks of histone modifications), and triangles (chromatin segments) and aligned with the x-axis. agMRCs are created to match the proviral integration sites with the frequency of targeting RefSeq Genes with the active TSSs part according to the chromatin segment classification. Thin dashed lines mark the p-values of 0.01, 0.05, and 0.1. Values outside the range of 0.01 to 0.1 are located at the lower/upper edge of the chart beyond the dashed lines. Gray symbols represent the significance of the difference between proviral integration sites and umMRCs. **C.** Examples of the charts representing the values of agMRCs and proviral integration sites. From left to right: Tss+RefSeqTSS, absolute distance; Tss+RefSeqTSS, distribution around TSS; peaks of H3K4me1 enrichment; enhancer segments.

Integration site distribution during the selection for HIV-1 GFP+ST proviruses

In the previous chapter we showed that GFP+ST proviruses of HIV-1 origin are enriched in the areas proximal to enhancers and that this distribution is not caused by targeting for active genes, which is a characteristic feature of HIV-1 integration. However, the picture of gradual changes in the distribution of HIV-1 proviruses during the selection for stable expression was missing at the time. The observation of the late silencing of GFP+^{3dpi} proviruses pointed out to the possibility of sequential changes in the distribution of HIV-1 proviruses during the selection for expression, similar to those observed in the case of ASLV. We thus followed the procedure of characterization of HIV-1 proviral integration sites in NS, GFP+^{3dpi}, and GFP+ST populations.

The proximity of HIV-1 GFP+ST proviruses to enhancers is caused by the selection for expressed proviruses

We isolated and characterized new sets of 23 NS and 68 GFP+^{3dpi} proviral integration sites and compared them to previously introduced sets of 135 umMRCs and 45 integration sites of GFP+ST proviruses.

In accordance with previous reports, with the frequency of 80 %, HIV-1 NS proviruses were found to be enriched in RefSeq Genes compared to umMRCs ($p = 4.4 \times 10^{-4}$, Fisher's Exact Test for Count Data, Fig. 20A). Despite the drop of the RefSeq Genes targeting frequency in the population of GFP+^{3dpi} proviruses, no significant differences in RefSeq Genes targeting were found between NS and GFP+^{3dpi}, and only very low significance was found when comparing GFP+^{3dpi} to GFP+ST proviral populations ($p = 0.039$, Fisher's Exact Test for Count Data). Analysis of the distance to TSSs of RefSeq Genes revealed that HIV-1 NS proviruses show the median of distance to nearest TSSs less than 20 kb which is significantly less that the median of distance of umMRCs ($p = 7.7 \times 10^{-4}$, Fisher's Exact Test for Count Data, Fig. 20B). No significant changes to the patterns set in the population of NS proviruses were observed in GFP+^{3dpi} and between GFP+^{3dpi} and GFP+ST. The same is valid for the distribution of proviruses around TSSs (Fig. 20C). These results corroborate the statement that the frequency of gene targeting and distance distribution to TSSs of GFP+ST proviruses is set by the integrational preferences of HIV-1 and is not shaped by the selection for active or stable expression.

In the previous chapter, HIV-1 GFP+ST proviruses were observed to be closer to enhancers compared to both umMRCs and agMRCs. We thus counted the distances of proviral integration sites of NS and GFP+^{3dpi} proviruses to ChromHMM chromatin segments (Fig. 21B). HIV-1 NS proviruses were significantly closer to the chromatin segments associated with active transcription. Selection for GFP+^{3dpi} and GFP+ST proviruses did not cause any shift in the distribution of proviruses toward the most of the "active" segments. For instance, all groups of HIV-1 proviruses displayed similar distribution to the Tss segment, with the median distance being around 20 kb. However, a significant shift closer toward enhancer-associated segments was observed for GFP+^{3dpi} proviruses (NS x GFP+^{3dpi} in Fig. 21B). GFP+ST proviruses displayed the distribution of distances to enhancers similar to that observed with GFP+^{3dpi} proviruses. Analysis of the distances to chromatin segments shows that HIV-1 targets active chromatin and that with selection for active proviruses of HIV-1, the proviruses are found closer to enhancers. The absence of the shift in the association with any of the features analyzed between GFP+^{3dpi} and GFP+ST HIV-1 proviruses indicates that single-step selection for integration sites occurs during the selection for expression early after integration.

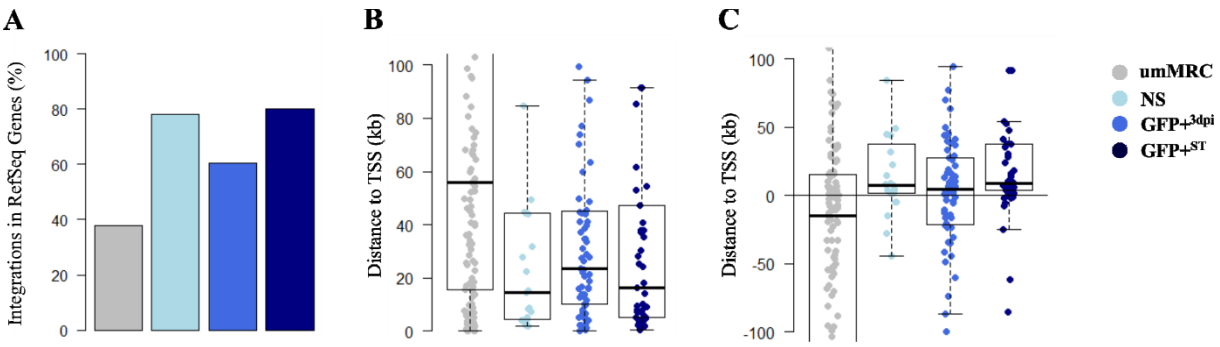

Figure 20. Distribution of HIV-1 proviral integration sites according to genomic features. **A.** Frequency of proviral integration sites found in RefSeq Genes. **B.** Boxplot representation of the distribution of distances of proviral integration sites to the nearest TSS. **C.** Boxplot representation of the distribution of proviral integration sites distances around TSSs. Negative values represent proviral integration sites outside TUs – distances represent the distances to the nearest TSSs. Positive values represent proviral integration sites inside TUs – distances represent the distances to the TSS of the targeted TU.

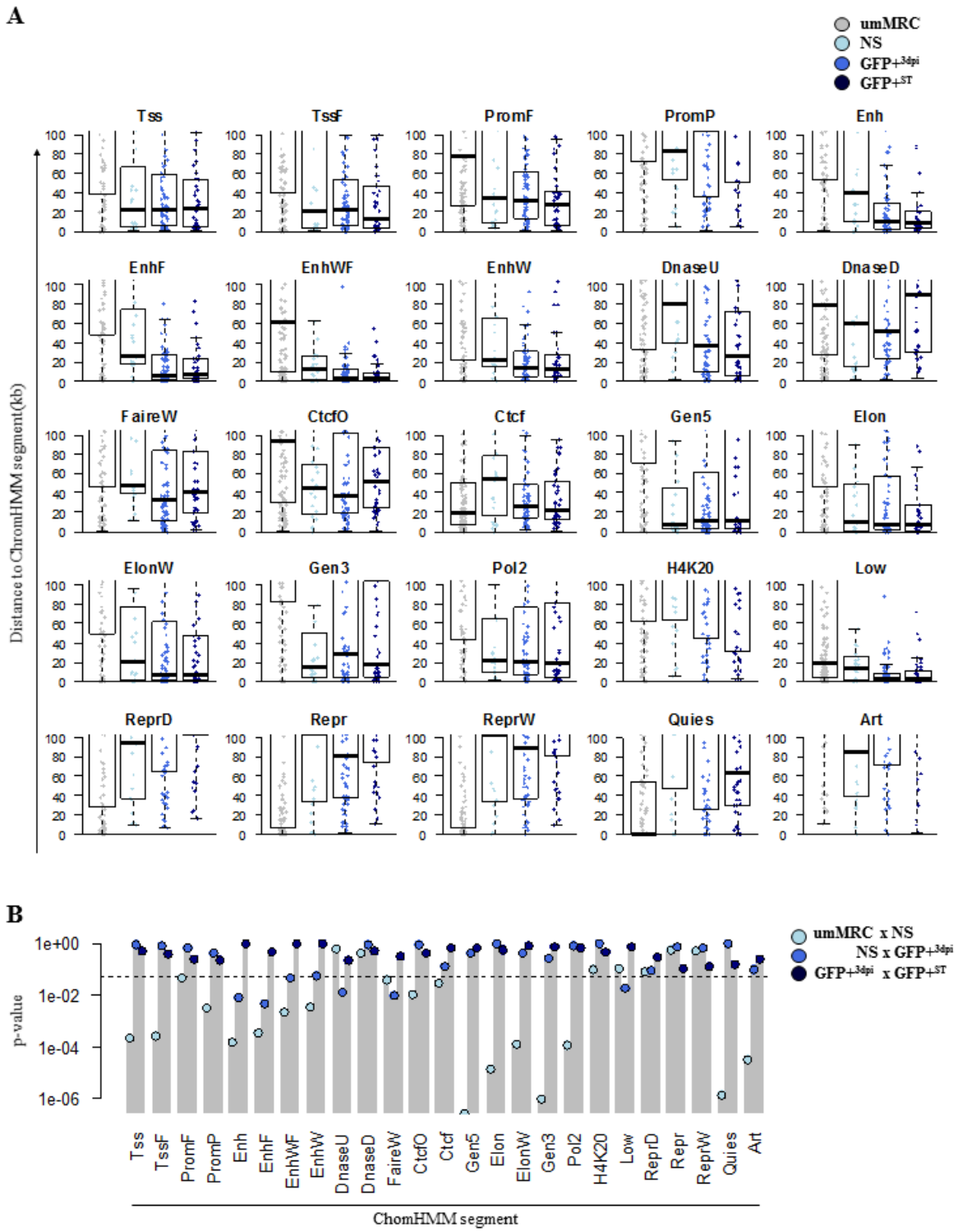

Figure 21. Distribution of HIV-1 proviral integration site distances according to chromatin segments. **A.** Boxplot representation of the distribution of distances of proviral integration sites to the nearest marked chromatin segment. **B.** Graphical representation of Wilcoxon test p-values of differences in the distance distributions of categories of HIV-1 proviral integration sites. P-values are on a logarithmic scale. The dashed line represents the value of 0.05.

Proximity of HIV-1 NS proviruses to TSSs as a function of short-gene targeting

In the previous experiment we showed that the proximity of HIV GFP+ST proviruses to TSSs is not the result of selection for expression as observed for ASLV. The distance distribution observed rather seems to be the result of native chromatin targeting of HIV-1, as NS HIV-1 proviruses displayed similar distance distribution toward TSSs as GFP+^{3dpi} and GFP+ST proviruses. Nonrandom proximity to TSSs of HIV-1 proviruses can be caused by targeting of TSSs by HIV-1 PIC or might be a bystander effect of targeting other features proximal to TSSs. As HIV-1 PIC is known to preferentially target active genes, we hypothesize that the HIV-1 proximity to TSSs is caused by the gene-targeting character of HIV-1 integration.

To challenge the hypothesis, we took advantage of a huge dataset of HIV-1 integration sites produced by Zhyvoloup et al. (2017), which we analyzed *in silico*. Four sets of integration sites, where each set comprised a triplicate experiment, were separated into two groups: cells treated by compound digoxin or a control group treated by DMSO (for the effect of the digoxin treatment see Zhyvoloup et al. (2017)). Two sets in each group were formed by integration sites of HIV-1, one with wild-type Gag (CA^{wt}), and one with N74D mutation in CA (CA^{N74D}) that lacked the ability to interact with PIC-tethering factor CPSF6. The coordinates of integration sites were downloaded from the paper data and remapped to the hg38 version of human genome assembly. To each integration site in a replicate we also created an *in silico* generated random genomic position. The numbers of random positions thus matched the numbers of integration sites in the particular set.

First, targeting of RefSeg Genes in the sets of integration sites was analyzed. Fig. 22A shows the reconstruction of results presented by Zhyvoloup et al. (2017), where both CA^{wt} and CA^{N74D} HIV-1 vectors preferentially target RefSeq Genes, although CA^{N74D} with lower frequency. HIV-1 CA^{wt} proviruses were also preferentially found in 10 kb windows around TSSs, while CA^{N74D} proviruses targeted the window with the frequency comparable to random targetAdditionally, in Fig. 22B we show the distribution of proviral distances to the nearest TSSs. Depiction of the distribution shows that the populations of HIV-1 proviruses tend to accumulate closer to TSSs compared to the random genomic positions regardless of the treatment. While the mean of distance median of CA^{wt} proviruses was under 20 kb, the mean of distance median of CA^{N74D} proviral was shifted further away above 20 kb from TSSs. However, proviral integration sites of both vectors peaked at the distance of 3 – 5 kb downstream to TSSs (Fig. 22C). Hence we corroborated the findings of Zhyvoloup et al. (2017) that CA^{wt} proviruses show higher gene targeting frequencies and tighter association with TSSs than CA^{N74D} proviruses and random genomic positions. We also corroborated results observed by us previously with small counts of NS HIV-1 proviruses representing the population of HIV-1 proviruses.

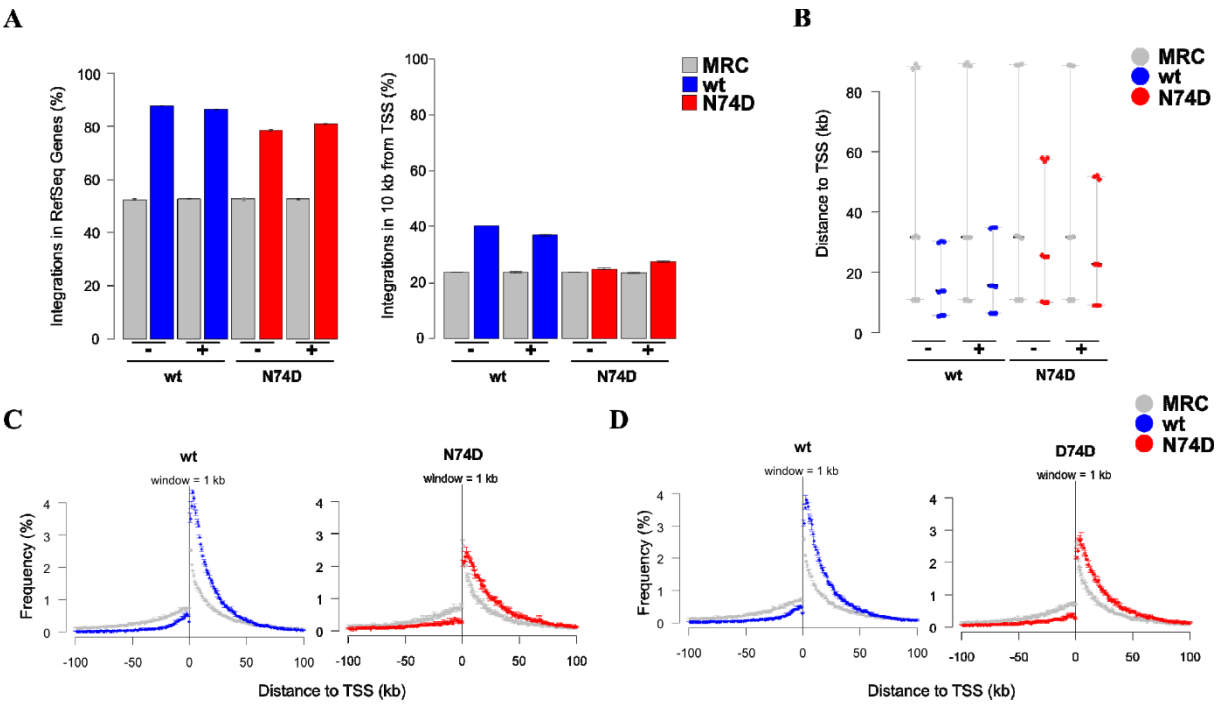

Figure 22. Differential distribution of proviral integration sites of HIV-1 CA variants according to TU-associated features. **A.** Frequency of proviral integration sites in RefSeq Genes and in a 10 kb window around RefSeq Genes TSSs. These charts repeat the results presented by Zhyvoloup et al. (2017). **B.** Distribution of distances of proviral integration sites to the nearest TSS. Colored dots represent the 1st quartile, median, and 3rd quartile values of triplicates. Means of triplicates and the interquartile range are displayed by gray lines. + and – signs in panels **A** and **B** represent the fact that samples were (+) or were not (-) treated by digoxin. **C.** and **D.** Distribution of proviral integration sites around TSS depicted as frequency of proviral integration sites in a 1 kb window with a particular distance to TSS. Negative values represent proviral integration sites outside TUs – distances represent the distances to the nearest TSS. Positive values represent proviral integration sites inside TUs – distances represent the distances to the TSS of the targeted TU. Dots represent means of triplicate targeting frequency. Bars represent standard deviation values of triplicates. **C** represents control DMSO-treated samples. **D** represents digoxin-treated samples.

To test whether gene targeting is the cause for HIV-1 provirus appearance close to TSSs, we generated another set of random positions in which each position matched a particular integration site in the distance to genes. This set of random positions was called gene-matched random controls (GMRC). As a result, GMRCs mimicked the frequencies of RefSeq Gene targeting of experimental integration site sets (Fig. 23A). Intergenic GMRCs, however, mimicked integration sites in the distance to the nearest gene represented by random choice by the distance to TSS or transcriptional end of the gene. GMRCs were observed to display the means of medians of the distances to TSSs slightly above 20 kb, which was not distinguishable from the means of medians of CA^{N74D} proviruses (Fig. 22B). However, unlike the HIV-1 integration sites, GMRCs peaked at the 1 kb window downstream to the TSSs (Fig. 22C). Targeting the RefSeq Genes thus shifts GMRCs closer to the TSSs of RefSeq Genes, resembling the distribution of the HIV-1 CA^{N74D} mutant. However, the proximity to TSSs of HIV-1 still is not caused by simple gene targetPeaking of proviral occurrence at the 3 – 5 kb window is also a nonrandom effect caused by the HIV-1 integration preference.

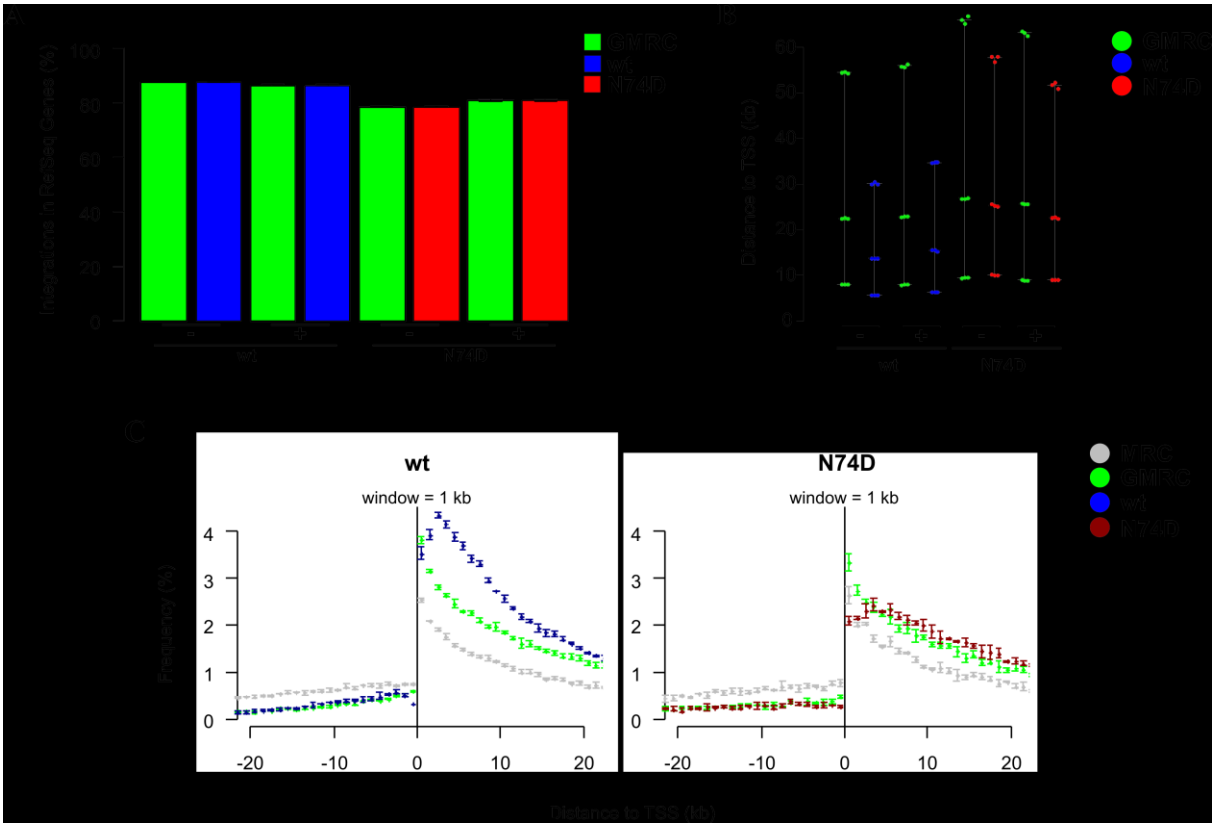

Figure 23. Comparison of proviral integration sites with gene-matched random controls (GMRCs). **A.** Random positions matching proviral integration sites in RefSeq Gene targeting frequency were generated. **B.** Distribution of distances of proviral integration sites to the nearest TSS. Colored dots represent the 1st quartile, median, and 3rd quartile values of triplicates. Means of triplicates and the interquartile range are displayed by gray lines. + and – signs in panels **A.** and **B.** represent the fact that samples were (+) or were not (-) treated by digoxin. **C.** Distribution of proviral integration sites around TSS depicted as frequency of proviral integration sites in a 1 kb window at a particular distance to TSS. Negative values represent proviral integration sites outside TUs – distances represent the distances to the nearest TSS. Positive values represent proviral integration sites inside TUs – distances represent the distances to the TSS of the targeted TU. Dots represent means of triplicate targeting frequency. Bars represent standard deviation values of triplicates.

Limited approximation of the GMRC distance distribution to that observed with HIV-1 CA^{wt} may point to the fact that CA^{wt} HIV-1 might preferentially target shorter TUs than would be expected by random targetTo test this, we used RNA-seq prepared by Zhyvoloup et al. (2017) from the control DMSO-treated Jurkat cell line. To allow counting for the target gene length, the transcriptome obtained from RNA-seq was simplified to "dominant transcripts". A dominant transcript is here a single transcript reconstructed from RNA-seq, which selected by the highest RPKM among other transcripts of the same gene represents a particular gene model. Plotting dominant transcript lengths against the RPKM of dominant transcripts showed that there is no correlation between the two (Fig. 24F). For further analysis, dominant transcripts with RPKM ≥ 0.5 were selected. Random genomic positions matching the integration sites by the distance to dominant transcripts were generated (active gene-matched random controls, AGMRCs). As expected, the gene-targeting frequency of GMRCs dominant transcripts targeting dropped to about half of values observed when all RefSeq Genes were used (compare Fig. 23A and Fig. 24A), while HIV-1 integration site frequencies decreased only slightly, meaning that most of HIV-1 proviruses of both CA variants integrate into transcribed genes (Fig. 24A). Analysis of the distribution of distances to TSSs showed that AGMRCs much closely resemble HIV-1 integration sites than GMRCs (Fig. 24B and C). In the case of CA^{N74D}, no significant difference could be found between GMRCs and proviral integration sites, meaning that AGMRCs display similar distribution of

distances to TSSs as HIV-1 proviral integration sites. On the other hand, significant differences were still found between CA^{wt} integration sites and the respective AGMRCs ($p = 3.5 \times 10^{-4}$, t-test). Analysis of the distribution of integration sites along gene bodies (Fig. 24D) showed that AGMRCs and CA^{N74D} integration sites are evenly distributed along the targeted dominant transcripts. By contrast, HIV-1 CA^{wt} integration sites were distributed unequally along dominant transcripts, peaking between 20 % – 60 % of the dominant transcript length with underrepresented areas in the first 10 % and last 20 % of targeted dominant transcripts. Finally, analysis of targeted dominant transcript lengths showed that the median of targeted transcript length in CA^{wt} samples is lower than in the respective samples of GMRCs and AGMRCs (Fig. 24E). On the other hand, distribution of target transcript lengths in CA^{N74D} samples was indistinguishable from the respective AGMRCs. From this analysis we conclude that the proximity of HIV-1 proviruses correlates with short transcript targeting and is disrupted in the CA^{N74D} mutant that displays the distribution observed by random active gene targeting.

To test whether TSS-proximity of wt HIV-1 proviruses is caused by the qualities of targeted genes, the last matched random controls set, exact-gene matched random controls (EGMRCs), was generated. EGMRCs are not only placed in an identical distance to genes as matched integration sites, but the distance is associated with exactly the same dominant transcript as integration sites. EGMRCs showed identical distribution of distances to the nearest TSSs as integration sites of both CA variants of HIV-1 (Fig. 25A) as well as similar distribution in targeted transcripts (Fig. 25B). Interestingly, the differences between EGMRCs and HIV-1 integration sites may be observed in short distances downstream to TSSs where the percentage of integration sites drops, which may mean that despite nonrandom proximity to TSSs of HIV-1 integration, close TSS-downstream loci are disfavored by HIV-1 integration irrespective of the CA variant present.

In this chapter we showed that HIV-1 proviruses are found in close proximities downstream to TSSs of targeted genes and that this distribution is most likely caused by preferred genic integration of HIV-1. Short genes are most likely preferred by HIV-1 equipped with wt CA, which might at least in part cause the close presence of HIV-1 to TSSs. Shown by comparison of CA^{wt} and CA^{N74D} vectors, the interaction of HIV-1 capsid protein with CPSF6 is most likely behind the short gene targeting of HIV-1 and shifting of integration closer toward the TSS-proximal areas.

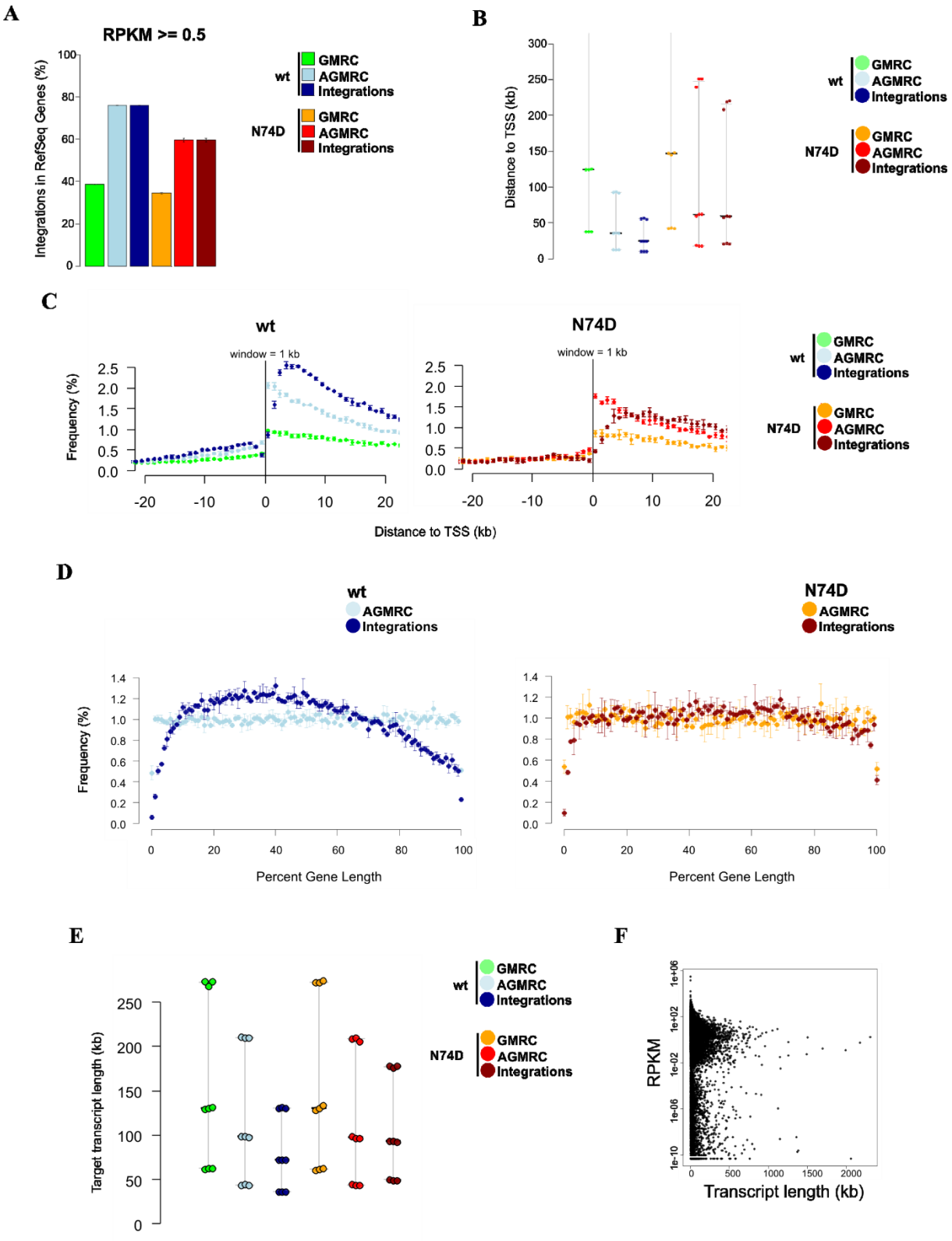

Figure 24. Distribution and gene length of proviral integration sites and active gene-matched random controls (AGMRCs). **A.** Frequency of proviral integration sites in RefSeq Genes with RPKM ≥ 0.5. AGMRCs that target transcribed RefSeq Genes with the same frequency as proviral integration sites were generated. Bars represent triplicate means. Antennas represent standard deviations. **B.** Distribution of distances of proviral integration sites to the nearest TSS. Colored dots represent the 1st quartile, medianm and 3rd quartile values of triplicates. Means of triplicates and the interquartile range are displayed by gray lines. **C.** Distribution of proviral integration sites around TSS depicted as frequency of proviral integration sites in a 1 kb window at a particular distance to TSS. Negative values represent proviral integration sites outside TUs – distances represent the distances to the nearest

TSS. Positive values represent proviral integration sites inside TUs – distances represent the distances to the TSS of the targeted TU. Dots represent means of triplicate targeting frequency. Bars represent standard deviation values of the triplicates. **D.** Frequency targeting of normalized gene. Dots represent the mean triplicate frequency in 1 % window along the normalized gene. Bars represent standard deviation values of triplicates. **E.** Distribution of dominant transcript length targeted by proviral integration. Colored dots represent the 1st quartile, median, and 3rd quartile values of triplicates. Means of triplicates and the interquartile range are displayed by gray lines. **F.** RPKM and length of all dominant transcripts obtained from RNA-seq data. All panels contain dominant transcripts with RPKM ≥ 0.5. wt – HIV-1 CA^{wt}, N74D – HIV-1 CA^{N74D}, GMRC – gene-matched random control, AGMRC – active gene matched random control, TSS – transcriptional start site, RPKM – reads per kilobase per million.

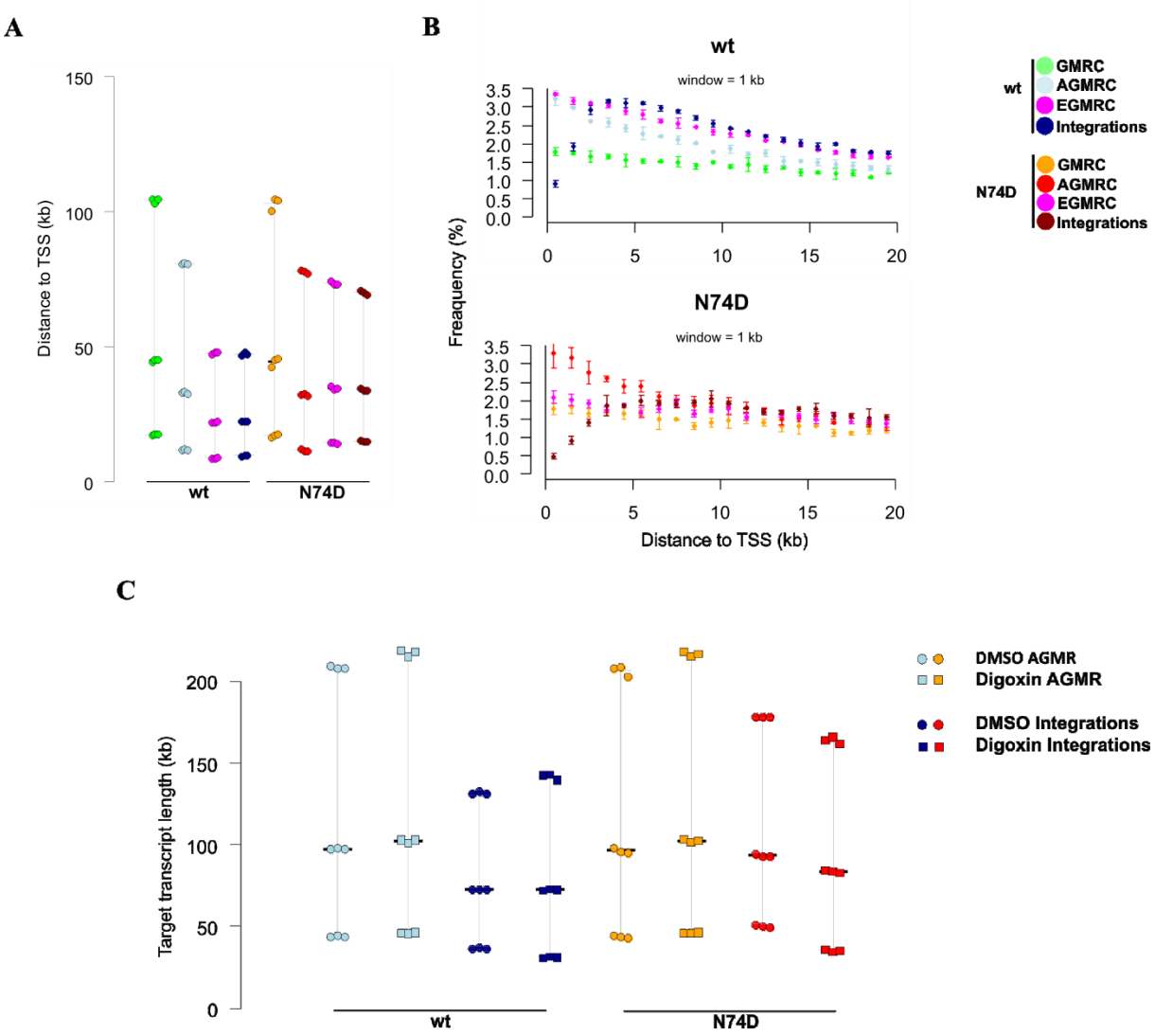

Figure 25. Distribution and gene length of proviral integration sites and exact gene-matched random controls (EGMRCs). **A.** Distribution of distances of proviral integration sites to the nearest TSS. Colored dots represent the 1st quartile, median, and 3rd quartile values of triplicates. Means of triplicates and the interquartile range are displayed by gray lines. **B.** Distribution of proviral integration sites around TSS depicted as frequency of proviral integration sites in a 1 kb window at a particular distance to TSS. Negative values represent proviral integration sites outside TUs – distances represent the distances to the nearest TSS. Positive values represent proviral integration sites inside TUs – distances represent the distances to the TSS of the targeted TU. Dots represent means of triplicate targeting frequency. Bars represent standard deviation values of triplicates. **C.** Distribution of dominant transcript length targeted by proviral integration. Colored dots represent the 1st quartile, median, and 3rd quartile values of triplicates. Means of triplicates and the interquartile range are displayed by gray lines. All panels contain dominant transcripts with RPKM ≥ 0.5. wt – HIV-1 CA^{wt}, N74D – HIV-1 CA^{N74D}, GMRC – gene-matched

random control, AGMRC – active gene matched random control, EGMRC – exact gene-matching random control, TSS – transcriptional start site, RPKM – reads per kilobase per million.

Expression of HIV-1 is long-term stable but toxic in T-cell lines

When the long-term expressional stability of the HIV-1-derived vector was examined in the K562 cell line, a decrease in the count of GFP+ST clones between 30 dpi and 60 dpi was observed. To test whether this late silencing is a general feature of HIV-1 expression, we analyzed i) expression of HIV-1 in the Jurkat cell line, ii) expression of a replication deficient full-length HIV-1 vector in Jurkat cells and other human T-cell-derived cell lines.

K562 and Jurkat cells were transduced by a minimal HIV-1-derived vector used in previous chapters (herein called HIV^{min}). The number GFP+ST clones, established from GFP+ cells sorted at 3 dpi, was counted at 30 and 60 dpi. With accordance to previous observations (Fig. 15A), the number of GFP+ST clones decreased between 30 and 60 dpi in K562 clones. On the other hand, the number of HIV^{min} GFP+ST clones in the Jurkat cell line was observed to be stable until 60 dpi (Fig. 26A). To test whether stable expression in the Jurkat cell line is also true for full-length HIV-1, we used two replication-deficient HIV-1-derived vectors encoding most of the HIV-1 proteins (Fig. 26B). Full-length (HIV^{fl}) and red-green (HIV^{RG}) vectors both carry replication-inactivating mutations in Env and PR and express GFP instead of the viral Nef protein. In addition, the HIV^{RG} vector expresses the red mCherry fluorescent protein instead of Gag. Jurkat cells were transduced by the three VSV-G-pseudotyped vectors, and GFP expression was followed over time. Surprisingly, the HIV^{fl}-transduced GFP+ cell count rapidly decreased over time, while HIV^{min} and HIV^{RG} GFP+ cell counts increased and kept stable from 3 to 9 dpi (Fig. 26C). Clonal analysis showed that HIV^{min} and HIV^{RG}-transduced cells efficiently grew into clones that mostly kept the expression of the provirus. On the other hand, HIV^{fl}-transduced cells were mostly unable to grow into the clones (Fig. 26D), which if achieved mostly did not express GFP (Fig. 26E). Finally, we tested the GFP expression of HIV^{fl} and HIV^{RG} in different T-cell-derived cell lines. As shown in Fig. 26F, the expression of GFP in the HIV^{fl}-transduced population dropped over time in each cell line examined, while HIV^{RG}-transduced cells were mostly stably present in the population.

Here we showed that, unlike in the K562 cell line, once established, the expression of HIV-1 is rather stable in T-cells. The stability of the minimal HIV-1-derived vector was also observed to be comparable to the expression of the replication-deficient full-length HIV-1 vector. However, when the *gag* gene was present in the vector (HIV^{fl}), GFP+ cells were rapidly lost from the transduced population and the ability of GFP+ cells to grow into the cellular clones was severely affected. High, long-term expression of HIV-1 can thus be toxic to the infected cells. Our results suggest that the *gag* gene products seem to play a role in the toxicity of HIV-1 expression. However, the source and the mechanism of the loss of HIV-1-expressing cells from the infected population needs to be examined in greater detail.

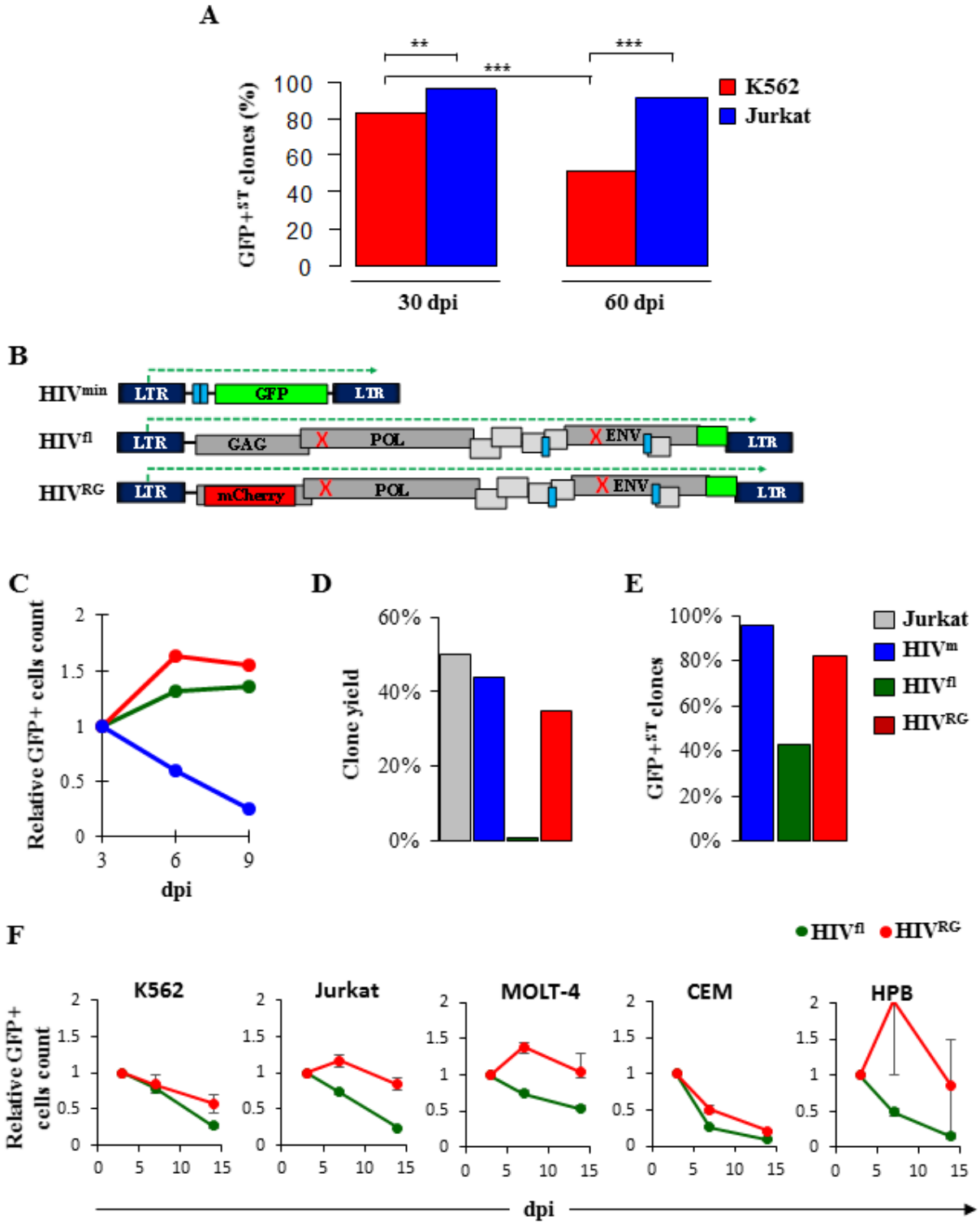

Figure 26. ***Expression stability of HIV-1 vectors in T-cell derived cell lines.*** **A.** Frequency of GFP+ST clones (≥ 90% GFP+ cells) of minimal HIV-1 vector in K562 and Jurkat cell lines. The frequency is counted for all clones received from a single-cell sort. Statistical significance was tested by Fisher's Exact Test for Count Data. * - p value < 0.05, ** - p value < 0.01, *** - p value < 0.001. **B.** Graphical representation of HIV-1-derived vectors. Viral protein-coding genes are in gray, Tat protein-coding exons are in blue, red crosses mark replication-inhibiting mutations present in viral protein-coding genes, EGFP-coding gene is marked in green, mCherry-coding gene is marked in red, LTRs are marked in blue, green arrow marks proviral transcription. **C.** Example of an experiment where the number of GFP+ cells was observed during a 9-day period in Jurkat cells. Frequencies of GFP+ cells are presented as relative numbers to the frequency of GFP+ cells at 3 dpi. **D.** Frequency of the Jurkat cell line

clones gained from all single-cell sorted cells. Hundred % mark the number of all wells of 96-well plate to which the cell was sorted. **E.** Frequency of GFP+ST clones (≥ 90% GFP+ cells) gained in panel **D.**. **F.** Relative frequency of GFP+ cells in T-cell derived cell lines transduced by the full-length HIV-1 variant during the period of 14 days. Frequencies of GFP+ cells are presented as relative numbers to the frequency of GFP+ cells at 3 dpi. Dots represent means of three parallel transductions. Bars represent standard deviations. HIV^{m} – minimal HIV-1 vector, HIV^{fl} – full-length pNL4-derived vector (pNL4-R7), HIV^{RG} – pNL4-RG vector.

The expression and integration sites of retargeted MLV vectors

We previously found MLV GFP+ST proviruses closely associated with TSSs and enhancers. However, γ-retroviral integration is naturally targeted to TSSs and enhancers thanks to the association of IN with the family of BET proteins (De Ravin et al. 2014, El Ashkar et al. 2014, LaFave et al. 2014), making it hard to distinguish whether the integration site distribution pattern observed is the result of selection for expression or is rather established by the natural integration preference of MLV. However, mutations in IN abolishing the IN-BET interactions and disrupting the natural preference of MLV were identified (El Ashkar et al. 2014). A single amino acid substitution of tryptophan at position 390 of unstructured C-terminal tail of IN was referred to be sufficient for disruption of the IN-BET interaction. We thus constructed a W390A mutant of IN (IN^{W390A}) and in combination with previously used minimal MLV-derived vector we compared the expressional stability and integration site distribution to the same MLV-derived vector combined with wt IN (IN^{wt}).

Stable expression of retargeted MLV vector

The K562 cell line was transduced with VSV-pseudotyped MLV vectors carrying IN^{wt} or IN^{W390A}. Numbers of GFP+ cells were analyzed both in the unsorted population and in the polyclonal population of GFP+ cells sorted at 3 dpi (Fig. 27A and B). In both cases, the percentages of GFP+ cells decreased only minimally to about 30 dpi, showing marks of stable expression. We also established cellular clones by sorting GFP+ cells at 3 dpi and examined the intraclonal stability of GFP expression at 30 dpi. In both IN^{wt} and IN^{W390A}, about 90 % of clones preserved at least 90 % of GFP+ cells in the clonal population (Fig. 27C).

Expression analysis shows that the MLV-derived vector utilizing either IN^{wt} or IN^{W390A} is stably expressed and both IN groups are indistinguishable in the stability of expression.

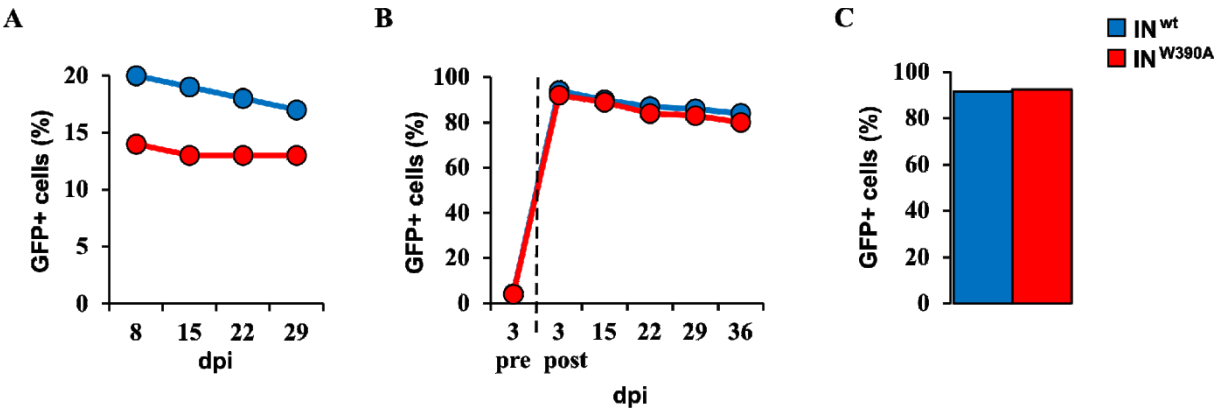

Figure 27. Expression stability of MLV vector IN variants in K562 cells. **A.** Frequency of GFP+ cells in polyclonal population transduced by high MOI during the period of 29 days. **B.** Frequency of GFP+ cell in sorted polyclonal population during the period of 36 days. The dashed line divides the chart to the parts displaying frequency of GFP+ cells before (pre) and after (post) cell sort**C.** Frequency of GFP+ST clones (≥ 90% GFP+ cells) at 30 dpi. Frequency is counted from all clones obtained from a single-cell sort. MOI – multiplicity of infection, IN^{wt} – minimal MLV vector with wild-type IN, IN^{W390A} – minimal MLV vector with IN bearing N74D mutation, dpi – days post infection.

The integration site landscape of retargeted vectors

In the next step we intended to characterize the integration sites of the IN^{W390A} vector to i) verify integration retargeting of the MLV vector, and ii) to see whether there is a shift toward any feature during the selection for the expressing proviruses. We thus transduced the K562 cell line with MLV-derived vector equipped with IN^{wt} or IN^{W390A}, and at 3 dpi we sorted the population of GFP+ cells ($GFP+^{3dpi}$ population). As the expression from both vectors was shown to be stable from 3 dpi, the integration sites from $GFP+^{3dpi}$ populations were isolated and compared to the integration sites from NS populations.

In total, 2,426 integration sites were identified, of which 350 and 223 belonged to IN^{wt} and IN^{W390} NS populations and 922 and 931 belonged to IN^{wt} and IN^{W390} $GFP+^{3dpi}$ populations, respectively. Proviral integration sites of all MLV groups were found in the RefSeq Genes with the frequency of 60 % (Fig. 28A). No signifficant differences in gene targeting were thus observed between IN^{wt} and IN^{W390A} MLV vectors. Analysis of $GFP+^{ST}$ MLV proviruses in previous experiments showed that the majority of proviruses were integrated in antisense orientation relative to the transcription of the targeted gene (Fig. 16B). We thus examined the orientation of IN^{wt} and IN^{W390A} proviruses in targeted RefSeq Genes. Except for IN^{wt} NS population, all populations contained equal portions of proviruses in both orientations (Fig. 28B). IN^{wt} NS proviruses displayed about 60 % of sense oriented proviruses, which was enough to achieve significant differences from IN^{W390A} NS ($p = 0.023$, Fisher's Exact Test for Count Data) and IN^{wt} $GFP+^{3dpi}$ ($p = 1.5 \times 10^{-4}$, Fisher's Exact Test for Count Data) populations.

Although no differences in RefSeq Genes targeting were observed between IN^{wt} and IN^{W390} vectors, differences were found in the distances of proviruses to TSSs. While the median of IN^{wt} NS proviruses was found around 3 kb, IN^{W390A} NS proviruses were found in the median distance of about 6 kb to TSSs ($p = 4.7 \times 10^{-4}$, Wilcoxon rank sum test with continuity correction, Fig. 28C). The difference observed between median distances of the two vectors was even more pronounced in $GFP+^{3dpi}$ populations, where the medians were 2 and 14 kb for IN^{wt} and IN^{W390A} MLV vectors, respectively ($p < 2.2e-16$, Wilcoxon rank sum test with continuity correction). Fig. 28D shows that he integration sites of IN^{W390A} NS and $GFP+^{3dpi}$ proviruses display wider distribution around TSSs than those of IN^{wt} vectors. We thus corroborate that IN^{W390A} is targeted further away from TSSs and we made the observation that with the selection of proviral expression, IN^{W390A} proviruses keep more distant from TSSs compared to IN^{wt} proviruses.

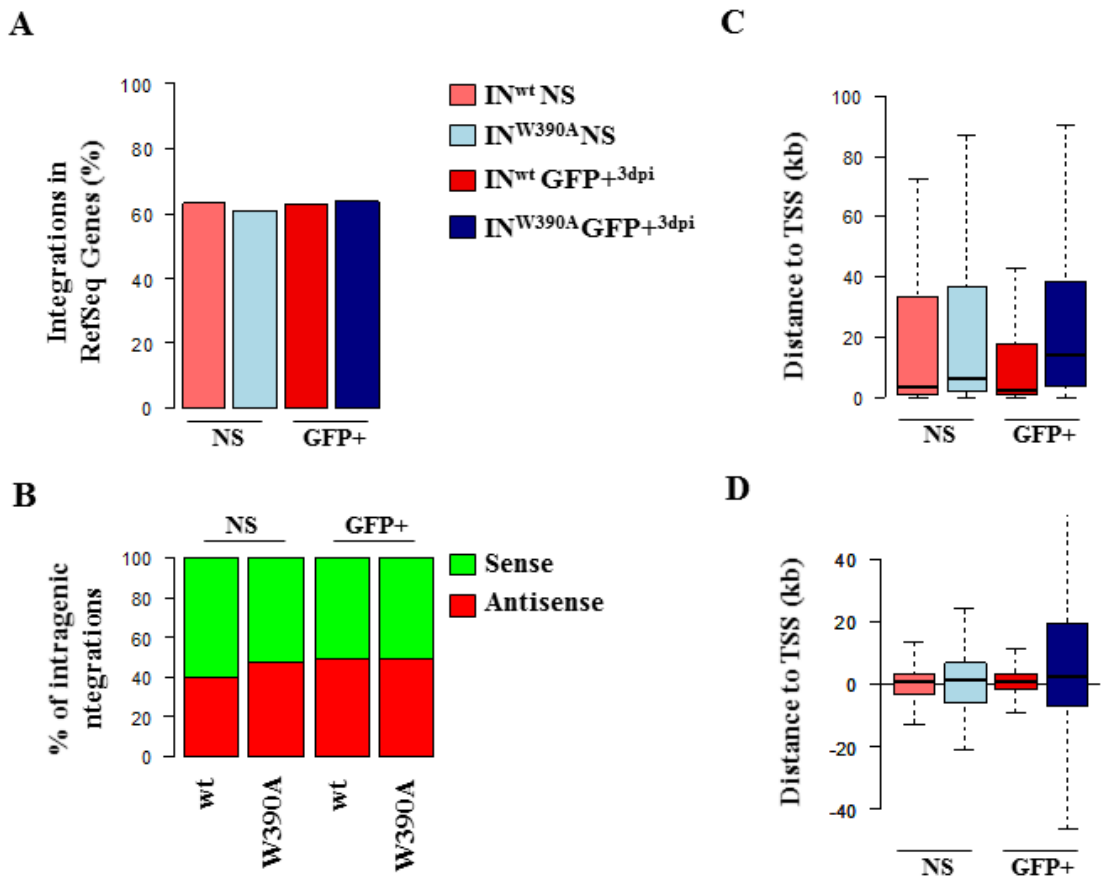

Figure 28. Distribution of proviral integration sites of MLV IN variants according to genomic features. **A.** Frequency of proviral integration sites identified in RefSeq Genes. **B.** Relative proviral orientation of proviruses integrated in RefSeq Genes. Sense orientation marks frequency of proviruses with the same orientation as transcription of the targeted RefSeq Gene. **C.** Boxplot representation of the distribution of distances of proviral integration sites to the nearest TSS. **D.** Distribution of proviral integration sites around TSS. Negative values represent proviral integration sites outside TUs – distances represent the distances to the nearest TSS. Positive values represent proviral integration sites inside TUs – distances represent the distances to the TSS of the targeted TU. NS – non-selected population, GFP+^{3dpi} GFP+ population sorted at 3dpi, INwt – minimal MLV vector with wild-type IN, INW390A – minimal MLV vector with IN bearing the N74D mutation.

We analyzed the distances of proviruses to the set of chromatin segments (Fig. 29A). As expected, MLV IN^{wt} NS proviruses were observed to be close to the Tss chromatin segment marking active TSSs with the median distance of proviruses being under 1 kb. The same distribution of distances toward the Tss segments was conserved for IN^{wt} GFP+^{3dpi} proviruses. The median distance of IN^{W390A} NS and GFP+^{3dpi} proviruses was at 10 kb, meaning that IN^{W390A} NS proviruses were more distant to Tss segments than their IN^{wt} counterparts ($p = 3.4 \times 10^{-4}$ for NS population and $p = 1.3 \times 10^{-54}$ for GFP+^{3dpi} population, Wilcoxon rank sum test with continuity correction, Fig. 29B), and the distant distribution was conserved even in the GFP+^{3dpi} population of IN^{W390A} proviruses, with median of distance at 12 kb. The same distribution was observed for Enh segments, where IN^{wt} proviruses were found closer than IN^{W390A}. Similar results were observed with the distances to enhancer flanking segments (EnhF) or 5' parts of genes (Gen5). Distances to other segments including weak enhancers (EnhW) did not show any significant changes between the groups of integration sites. IN^{W390A} was thus observed to integrate further away from active TSSs and enhancers than IN^{wt}. Selection for active proviruses did not cause any shift of IN^{W390A} proviruses closer to any feature observed. We thus conclude that the integration preference of the MLV IN^{W390A} vector is shifted and that MLV expression is stable despite the integration retargeting caused by disruption of the IN-BET interaction.

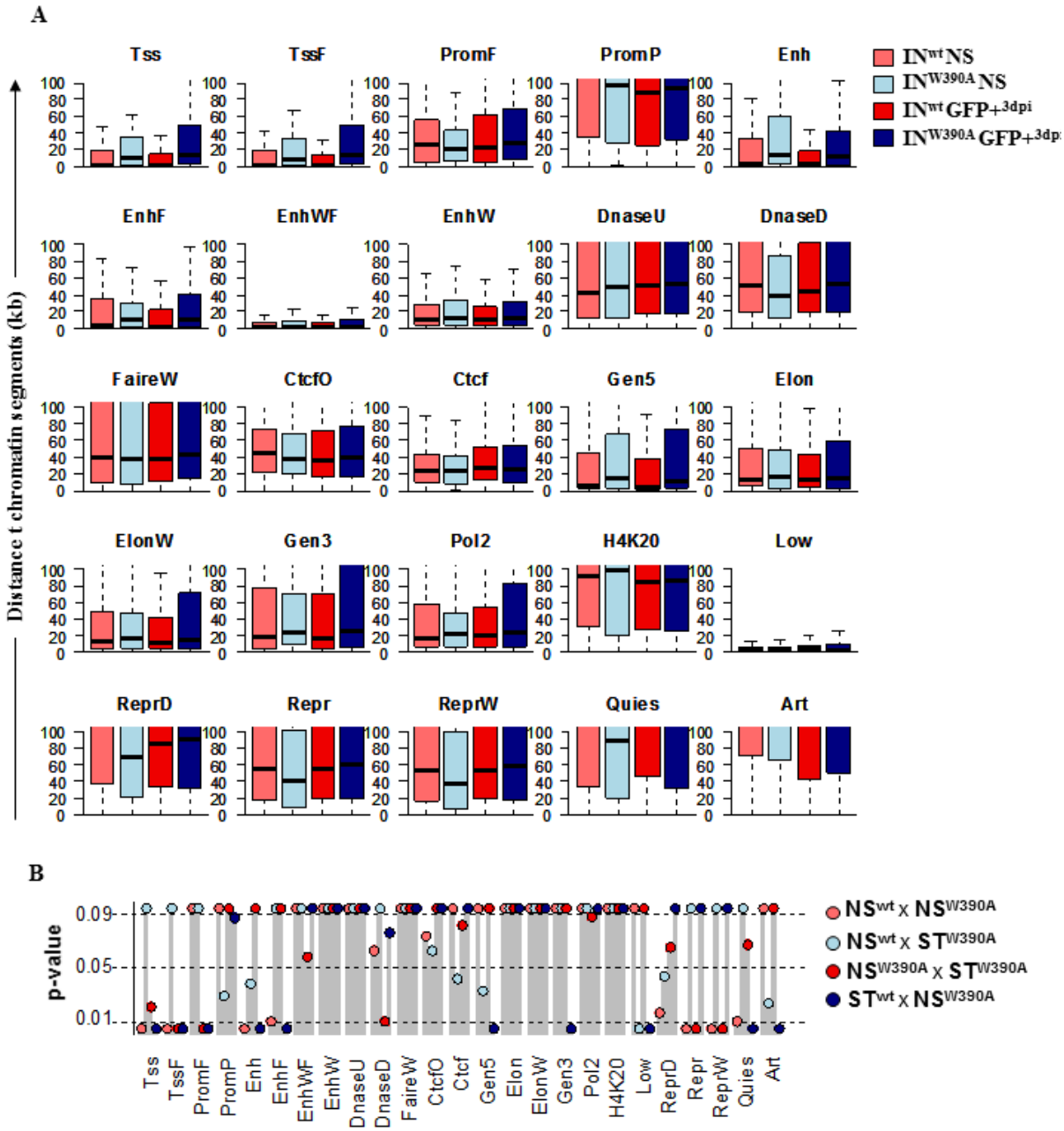

Figure 29. Distribution of MLV variant distances to chromatin segments. **A.** Boxplot representation of the distribution of distances of proviral integration sites to the nearest marked chromatin segment. For explanation of chromatin segment mnemonics, see Materials and Methods. **B.** Graphical representation of Wilcoxon test p-values between the groups of integration sites. Values > 0.09 or < 0.01 are positioned over or under respective dashed lines. NS – non-selected population, GFP+^{3dpi} / ST - GFP+ population sorted at 3dpi, wt – minimal MLV vector with wild-type IN, W390A - minimal MLV vector with IN bearing the N74D mutation.

Conclusions

The aim of the thesis was to define the host genomic and epigenomic features associated with stable proviral expression. We examined the long-term stability of expression of ASLV-, MLV-, and HIV-1-derived vectors. Changes in the proviral integration site distribution during the selection for active proviruses were observed. Our observations have led us to the conclusion that there are certain chromatin markers defining the environment supporting stable expression of proviruses and that despite different integration preferences of particular retroviruses, the features of the expression-supportive proviral integration sites are shared across retroviral genera.

We thus conclude that:

- Expression of CpG island-modified ASLV proviruses is about an order of magnitude more stable than the expression from ASLV with wt LTRs.
- MLV- and HIV-1-derived vectors examined in our work exhibit stable expression in human cells.
- The bodies of transcribed TUs form permissive environment for MLV, HIV-1 and CpG island-modified ASLV but not for ASLV with wt LTRs.
- GFP+ST proviruses of ASLV-derived vector are found in transcribed TUs but in close proximity to active TSSs.
- GFP+ST proviruses of MLV, HIV-1 and CpG island-modified ASLV are found to be associated with enhancers.
- Observed significant proximity of HIV-1 GFP+ST proviruses is caused by the selection for the expressed proviruses while significant proximity of HIV-1 GFP+ST proviruses to active TSSs is caused by the nature of HIV-1 integration, probably resulting from CA-directed preference for shorter TUs.
- The stability of MLV-derived vector expression is not dependent on IN-BET interaction-directed targeting of enhancers and active TSSs.

Our results thus suggest that, with the exception of MLV-derived vector, the proviral expression is mostly integration site dependent. Our results may affect the design and evaluation of retroviral vectors used for transgene expression. We stress the importance of proviral distribution testing after the selection for proviral expression stability, since we have observed changes during the selection. Generally, our results presented here should be considered as part of the discussion about retroviral expression and silencing.

Discussion

Here we presented the results of the studies of proviral expression stability and the effect of selection for stably active proviruses on the distribution of proviral integration sites. Methodologically, we comprised a variety of methods used to study the proviral integration sites, from a low-output clonal approach to the high-throughput sequencing methods yielding high numbers of proviral integration sites. In fact, clonal analysis of proviral expression was a prominent methodological approach to observe the stability of proviral expression. The main drawback of the approach is its high labor intensity and low yield of integration sites gained. As an alternative, Chen et al. (2017) recently developed high-throughput sequencing methodology for identifying proviral integration sites of transcribed proviruses. Where meaningful, we left sequencing of clonal populations and moved to sequencing of polyclonal populations. This was, however, subject to invariability in the expressional phenotype observed. The clonal approach can still offer some advantages, allowing single provirus expression trackEven when integration sites were obtained by sequencing the libraries prepared from polyclonal populations, analysis of the stability of proviral expression in clonal populations was performed. The clonal approach was also successfully applied in the studies focused on the stochastic nature of gene expression (Skupsky et al. 2010). The method is thus relevant for the studies of gene expression.

In our work, we evaluated the time course of GFP expression and collected and categorized the phenotype profiles of proviruses. Enhanced GFP has a half-life of about 26 h and less stable variants are available (Corish and Tyler-Smith 1999). GFP stability can thus mask quick stochastic changes in the expression of the gene. This, however, does not matter because our goal was to observe the long-term stability of proviral expression. By day 30 after viral transduction, which was the minimal time-point at which we examined the proviral expression stability, there should be no effect of GFP stability on the results obtained. The last time-point at which we examined the proviral expression stability was set to 60 dpi. As we did not follow the expression any longer, it may be a subject of debate whether the period of two months is sufficient to pronounce proviral expression to be stable. However, for most proviruses, only mild changes in the expression stability and subsequently in proviral integration site distribution was observed between 30 and 60 dpi. An exceptional situation was observed in ASLV-transduced cells, where strong signs of late silencing events were observed. In the case of ASLV such observation could be meaningful. Based on our results, we, however, conclude that most silencing events, if occurring, happen early after proviral integration. It is also important to mention that our definition of stable expression allows infrequent but noticeable silencing of proviral expression. Even in the most stably active vectors, some GFP-negative cells were observed during the experiments. Whether this was an artefact of the measurements or the result of stochastic silencing of gene expression is not clear. It is thus questionable whether something like absolutely stable expression can be observed. Yet we believe that our approach is satisfactory for the identification of low-frequency silenced proviruses.

Bioinformatic pipelines useful for the analysis of integration sites were also developed as a part of the studies. Unlike in the most of the studies concerning proviral integration sites where frequency in distance windows were analyzed, the main metric in our analysis is the distance distribution to the studied feature. This allows us to show the full picture of the distance distribution and its changes during the selection of proviruses. Throughout the whole study, publicly available data about expression of TUs or epigenome for the particular cell line were used. Data thus do not come from the same cells that were cultivated in the lab. Moreover, the K562 cell line used in most of the experiments is hypotriploid, with many aberrant changes compared to the reference genome (Naumann et al. 2001). Since we did not perform sequencing of the K562 genome and relied simply on the reference genome, the last two factors may have introduced inaccuracy into the analysis and the results, and should be taken into account when drawing the conclusions. To assess the effects of retroviral integration and selection on the distribution of proviral integration sites, we applied *in silico* created integration sites. The application of random sites shows the probability of feature targeting without any biological bias to be present. We presented several possible ways of generating random controls. Matching the methodological and biological

features of proviral integration sites became an important part of the work presented. Uniquely mapped random controls (umMRCs), which match the filters applied during the integration site mapping, were generated to avoid genomic loci to which the integration sites can hardly be assigned. Although the effect of umMRC application was not tested, we believe that application of unique-mapping criteria is an important part of the random control generation. Moreover, we generated MRCs matching the integration sites in targeting of the features of interest. These MRCs helped us to define the probability that targeting of one feature affects the targeting of another feature. Finally, sets of proviral integration sites corresponding to proviruses resistant to silencing over the experimental period were obtained and the results bring new facts to the discussion on the importance of integration site in proviral activity.

Retroviral integration was previously shown to be nonrandom and targeted toward the features of transcriptionally active chromatin. HIV-1, MLV, and ASLV became model retroviruses for integration targeting, showing differential distribution of proviral integration sites (Mitchell et al. 2004). HIV-1 was shown to preferentially target the bodies of active TUs (Elleder et al. 2002, Schroder et al. 2002), MLV active promoters and enhancers (Wu et al. 2003, De Ravin et al. 2014, LaFave et al. 2014), and ALSV was shown to display a mild, non-significant preference for the genes otherwise being close to random distribution of integration (Narezkina et al. 2004). Mechanistically, the targeting of certain features in the genome are mostly assigned to interactions of IN and CA of proviral PIC with cellular proteins that preferentially localize to the genomic loci later targeted by retroviral PIC (Ciuffi et al. 2005, El Ashkar et al. 2014, Winans et al. 2017). While the mechanisms behind the nonrandom targeting of retroviral integrations are known, the impact of the targeting preference on the expression of proviral genome is still unclear. In the work presented here, we utilize replication-deficient retroviral vectors to study the stability of retroviral expression and distribution of integration sites of proviruses selected for stable expression.

Our main experimental model was an ASLV-derived vector transducing human cells. Previously it was shown that ASLV is extensively silenced when infecting mammalian cells (Wyke and Quade 1980, Chiswell et al. 1982, Hejnar et al. 1994). Our group has shown that the mechanism of ASLV expression silencing is dependent on the environment of proviral integration and that rarely found unsilenced proviruses are found in TSS-proximal areas marked by H3K4me3 (Senigl et al. 2012). Together with more or less random distribution of integration, ASLV represents a good model for searching the features affecting proviral expression. Here we corroborated the previously obtained results with an ASLV-derived vector transducing the human K562 cell line when we showed that only a few percent of ASLV proviruses found active after integration keep stable expression as far as to 60 dpi and that these proviruses are preferentially found inside transcriptionally highly active TUs in close proximity to their TSSs. Furthermore, owing to extensive data about the K562 epigenome available, we were able to define new associations between the expressed proviruses and histone modifications.

For the first time we showed that ASLV integration was enriched in active chromatin compared to random sites. After selection for GFP+^{3dpi} and GFP+ST proviruses, the ratio of proviruses in active chromatin remained the same, but the percentage of proviruses integrated in promoter-associated chromatin was increased, corroborating the importance of active TSS-proximity for ASLV proviral expression in mammalian cells. Selection for expressed proviruses established association with the H2A.Z histone variant in the proviral population. This association remained at the same level throughout all next stages of selection. H2A.Z represents an open chromatin marker that has been found in association with active TSSs and enhancers (Soboleva et al. 2011). Association with H2A.Z thus represents transcribed proviruses but does not distinguish between lately silenced and stably expressed proviruses. Transcribed ASLV proviruses were more closely associated with characteristic markers of active TSSs and enhancers – all three methylation states of H3K4, acetylated H3K9 and H3K27. Remarkable association was observed between ASLV proviruses from the GFP+ST population and H3K79me2 and H4K20me1. The histones methylated at H3K79 and H4K20 were described to be associated with several chromatin states, including either active or transcriptionally silent chromatin.

Our results, however, show the association of H3K79me2 and H4K20me1 with transcriptionally active chromatin forming a transcription-permissive environment for proviral expression.

When LTR of ASLV was equipped with the CpG island core element, striking changes of GFP expression in proviral populations were observed. First, the expression of the modified vector was stabilized without any increase in expression intensity being observed. Second, proviral populations of GFP+^{3dpi} and GFP+ST proviruses were released from TSS-proximal areas toward the distal parts of the bodies of active TUs. Notably, the frequency with which proviruses found in active TUs were expressed was comparable with the frequency observed in the unmodified ASLV populations. The bodies of active TUs are the genomic regions where ASLV is silenced by the mechanism dependent on DNA methylation (Senigl et al. 2012). The CpG island was shown to protect the ASLV proviruses from DNA methylation and subsequent silencing of its expression (Hejnar et al. 2001, Senigl et al. 2008). The DNA methylation-protecting attribute was assigned to Sp1 binding sites of the CpG island (Machon et al. 1998). Based on the results presented here and elsewhere, we hypothesize that the Sp1 binding site-containing CpG island stabilizes the expression of ASLV-derived vector in gene bodies through protection from DNA methylation. The mechanism of this attribute of Sp1 binding is unknown. Although GFP+ST proviruses of the CpG-modified ASLV vector were found further away from TSS-proximal areas, they were found to be associated with the H3K4me1 peaks and enhancer chromatin segments. As discussed later in this section, the proximity to enhancers seems to be a general feature of silencing-resistant long-term expressed proviruses.

Active TUs seem to be the genomic features associated with expressed proviruses. About 80 % of GFP+ST proviruses of both ASLV vectors and of all experimental groups of HIV-1 vectors presented in this work, were observed to be integrated in active TUs. As epigenetic environment, the bodies of TUs are generally considered to be repressive for the initiation of transcription thanks to the action of H3K36me3-driven DNA methylation of gene bodies (Neri et al. 2017). According to the data obtained from the experiments with CpG island-modified ASLV, we suggest that retroviruses can defend themselves against the intragenic DNA methylation. It was also reported that mammalian retroviruses contain CpG-rich Sp1 binding sites (Harrich et al. 1989, Wahlers et al. 2002). Finding that DNA methylation acts on HIV-1 LTR as a locker of the already silenced state of promoter induced by other epigenetic mechanisms (Pion et al. 2003, Blazkova et al. 2009, Trejbalova et al. 2016) supports the theory that DNA methylation does not act at Sp1-equipped provirus as primary silencing factor. For the transcription of mammalian retroviruses, bodies of active TUs thus seem to be an epigenetically permissive environment. However, the distribution of MLV and HIV-1 expressed proviruses in the bodies of active TUs in respect to the TSSs were observed to be different from random. MLV GFP+ST proviruses were concentrated around TSSs. Targeting of areas proximal to TSSs is, however, a general feature of MLV integration (Wu et al. 2003, LaFave et al. 2014), and we did not observe any shift of proviral integration sites toward the TSSs during the course of selection for the expressed proviruses. On the other hand, HIV-1 distribution of proviral integration sites of GFP+ST proviruses was observed to be wider, reaching more distal parts of TUs. However, HIV-1+ST proviruses were observed to be nonrandomly close to active TSSs and enhancers. Unlike the proximity of HIV-1 proviruses to TSSs, which we assign to the natural preference of HIV-1 for short TUs driven by HIV-1 CA interacting partner CPSF6, no preferential integration to the proximity of enhancers was observed in the NS population of HIV-1-transduced cells, and it appeared only after selection for the population of expressed proviruses. Moreover, *in silico* simulation of TU-preferred integration revealed no shift of integration sites toward enhancer regions. HIV-1 and CpG-modified ASLV vectors were thus preferentially observed to be expressed in the bodies of active TUs and at the same time close to the enhancers. Notably, none of the two vectors showed any preference for enhancers.

Another feature of active TUs acting repressively on the establishment and maintenance of *de novo* transcription in gene bodies is transcriptional interference. The transcriptional interference driven by convergent endogenous transcription was proposed to be one of the mechanisms silencing the proviral

expression (Han et al. 2008, Lenasi et al. 2008, Shan et al. 2011). Our data, however, suggest limited action of transcriptional interference on proviral expression. Senigl et al. (2012) described a mild correlation between the distance to TSSs and orientation of the expressed provirus. In the data presented here, we also see a slight increase in the ratio of proviruses integrated in sense orientation to targeted TU transcription during the selection of expressed ASLV proviruses. However, no conclusion about the role of proviral orientation in the stability of expression can be made from the data, as the differences between the selection steps are not significant. MLV and HIV-1 show no preference for proviral intragenic orientation in the populations selected for expression. Our data thus show that proviruses integrated in TUs are capable of stable transcription in either orientation relative to the transcription of the targeted TU. Kaczmarek Michaels et al. (2015) also reported the cases where silent HIV-1 proviruses integrated in active TUs were capable of reactivation without affecting the activity of targeted TU. We thus conclude that in our data, no relation between stable expression and proviral orientation in the targeted TU was observed.

Enhancers came out of our study as an important feature for proviral expression. Except for ASLV, GFP+^{3dpi} and GFP+ST proviruses of CpG-modified ASLV, HIV-1, and MLV vectors were enriched in the proximity to enhancers and enhancer-associated features. Chen et al. (2017) also reported that HIV-1 transcriptionally active and reactivation-prone proviruses are integrated closer to enhancers than their silenced and unreactivable counterparts. Enhancers are defined as regions capable of activation or enhancement of transcription from promoters with which they interact. At the level of functional genomics and epigenomics, enhancers are characterized by the presence of bound transcription factors, acetylated histones (especially H3K27ac), H3K4me1, and bidirectional transcription giving rise to enhancer RNAs (Kim and Shiekhattar 2015). The importance of enhancers in the biology of retroviruses is stressed by the fact that BET protein family interaction-directed targeting of enhancers is conserved among γ-retroviruses and is also the feature of *piggyBac* transposon integration (Gogol-Doring et al. 2016). It is, however, unclear what role endogenous enhancers play in proviral transcription. A general feature of enhancers is their ability to activate transcription of TUs close to which they are positioned. This positive position effect may also apply to closely integrated proviruses. Interestingly, while positioned close to enhancers, the majority of CpG-modified ASLV and HIV-1 proviruses were also found in active TUs, meaning that intragenic enhancers may play a role in establishment of a transcription-permissive intragenic environment. Indeed, about half of identified enhancers can be found inside active TUs (Andersson et al. 2014, Cinghu et al. 2017). Interestingly, beside their enhancing effect in transcription of the TU in which they are positioned, intragenic enhancers were also found to attenuate transcription from the TU, most likely by the mechanism of transcriptional interference (Cinghu et al. 2017). Intragenic enhancers thus offer an epigenomically transcription-permissive environment and also act as endogenous transcription attenuators. The effect of endogenous enhancer activity may be an important topic for studies of the relation between proviral integration and proviral expression activity.

Of all the retroviruses studied, MLV displays the sharpest integration preference, with proviruses being concentrated at active TSSs and enhancers. Since TSSs and enhancers were identified as features associated with proviral expression, we questioned whether integration preference is the cause of the very stable expression of MLV that we also observed. A BET-independent IN mutant of MLV showed stability of expression similar to a BET-utilizing MLV vector despite the shifted integration toward TSS- and enhancer-distant positions. No shift toward any active-chromatin feature was observed for either vector after the selection for stable proviral expression. The results suggest that the expressional stability of MLV lies in the strength of MLV LTR-encoded promoter. Results proposing stable expression of BET-independent MLV-derived retroviral vector were also presented by El Ashkar et al. (2017). In the study, the authors, however, used a self-inactivated vector utilizing an internal promoter different from MLV LTR to drive the transgene expression. High multiplicity used by the authors may also mask the transgene silencHowever, the results presented so far signal that BET-independent MLV-derived vectors may still be targeted to the loci permissive for proviral expression. One of the factors that may shape the distribution of MLV integration is p12 – the product of the γ-retroviral *gag* gene.

p12 was identified as a proviral chromatin-tethering factor responsible for the attachment of MLV PIC to chromosomes during cellular division (Elis et al. 2012, Wanaguru et al. 2018). The next generation of BET-independent vectors utilizing heterochromatin-binding peptide-fused IN (El Ashkar et al. 2017) may help to answer the question whether γ-retroviral vectors are stably expressed in heterochromatin-proximal regions.

From the experiments conducted with HIV-1-derived vectors, we conclude that once proviral expression is established, very weak silencing of HIV-1 expression is observed in human T-cell-derived lines. The late silencing events observed in the K562 cell line were probably a cell-line specific phenomenon. Despite the stable expression of HIV-1 vectors we observed a dramatic decrease of GFP+ cells in the population of cells transduced by the replication-deficient NL4-derived HIV-1 vector. This lack of transduced cells seems to be dependent on the expression of Gag, since a Gag-deficient NL4 vector (Nl4-RG) demonstrates stable expression in the population. So far, HIV-1-induced death of CD4+ T-cells was attributed to reverse transcription products and unintegrated DNA detection-triggered programed cell death (Doitsh et al. 2010, Muñoz-Arias et al. 2015). To our best knowledge, no toxic effect of HIV-1 genome expression on infected cells has been reported. Therefore, whether the Gag expression causes cell death, arrest in cell cycle, or there is another yet unknown effect in play, we cannot say. Our results give the impression that *in vivo*, HIV-1 expression may be toxic to infected cells and thus may hardly be observed. The integration site experiments suggest that the main wave of silencing is present early after integration. This observation is in line with the observations of others suggesting that the latent reservoir of HIV-1 is established rapidly and early after infection (Chavez et al. 2015). The main feature of HIV-1 active proviruses is the shift toward the enhancers. Interestingly, we also observed nonrandom accumulation of HIV-1 proviruses near to TSSs, yet this is, according to our observations, rather the result of CPSF6-directed short gene targetResults leading to a similar conclusion can be found in the study conducted by Achuthan et al. (2018). These authors, however, do not discuss shorter gene targeting, and the mechanistic explanation of the phenomenon is thus still missZhyvoloup et al. (2017), authors of the study of which we used data for *in silico* gene-targeting experiments, reported that CPSF6-utilizing HIV-1 PIC preferentially target tissue-specific genes. If tissue-specific genes were found to be shorter in length, then targeting tissue-specific genes would explain the observation. Since CPSF6 is a component of the complex acting on polyadenylation of mRNA, it is unclear how CPSF6 could conduce to PIC-targeting of short genes.

Retroviral integration site selection studies were fueled by the concerns over genotoxic stress caused by retroviral vectors used for gene therapy. Indeed, dysregulation of endogenous gene expression upon retrovirus integration was observed in the gene therapy trials using both γ-retroviral (Hacein-Bey-Abina et al. 2003) and lentiviral vectors (Cavazzana-Calvo et al. 2010). The most often observed perturbation induced by proviral integration is clonal expansion of provirus-containing cells. Unlike in γ-retroviral trials, no pathological events raising from clonal expansion was observed in lentivirus-utilizing trials. Interestingly, clonal expansion is also observed in the CD4+ cell population of HIV-1-infected patients (Maldarelli et al. 2014). It is, however, still subject of discussion whether the clonality is caused by provirus-induced genotoxic stress or by homeostatic proliferation of infected cells (Chomont et al. 2009). As reported by Suerth et al. (2012), α-retroviral vectors provide an alternative to the γ-retroviral and lentiviral vectors, offering stable expression together with less genotoxic stress, supporting genome targetGenomic positions supporting sufficient transgene expression while not affecting genomic integrity were called “safe harbors”. From the definition, such safe harbor should meet several criteria covering the minimal distance to TSS (50 kb), distance to the cancer-related gene and miRNA (300 kb), and location outside TUs and ultraconserved regions (Papapetrou et al. 2011, Sadelain et al. 2011). Integrating vectors, including those derived from retroviruses or transposons, are challenged by the criteria for targeting the safe harbors. In the study concerning integration sites vectors derived from retroviruses and transposons, Gogol-Doring et al. (2016) stated that retroviral vectors have a very low frequency of integration into the loci meeting criteria for safe harbors (about 3 % for HIV-1) and that the most random-integrating of vectors – the one derived from *piggyBac* transposon – has a frequency

of integration into safe harbor about 20 %. The data were, however, obtained using cells that were not selected for expression, and it is thus unclear whether the expressed vectors displayed the same distribution of integration sites as unselected vectors. Another study concerning safe harbors conducted by Papapetrou et al. (2011) reported that about 17 % of lentiviral vector integration sites meet the criteria and that the chance of retrieving the expressed provirus in safe harbor loci is about 7 %. Our data show that the xpressed proviruses of α-, γ-retroviral, and lentiviral origin expressed from the viral LTR promoter would hardly meet the criteria of stable expression in safe harbors. Although we did not directly test the targeting of the safe harbor criteria-meeting loci, it is quite clear that stably expressed proviruses are frequently found in active TUs and in the proximity to regulatory elements. However, the chances of such proviruses, e.g., in an intron of active TU close to an active enhancer, to dysregulate the gene expression and cause changes in the cellular physiology are unknown. Based on our results we propose that the integration sites and target site-affection risks should be examined after the selection for proviruses meeting the criteria for stable expression.

References

Achuthan, V., J. M. Perreira, G. A. Sowd, M. Puray-Chavez, W. M. McDougall, A. Paulucci-Holthauzen, X. Wu, H. J. Fadel, E. M. Poeschla, A. S. Multani, S. H. Hughes, S. G. Sarafianos, A. L. Brass and A. N. Engelman (2018). "Capsid-CPSF6 Interaction Licenses Nuclear HIV-1 Trafficking to Sites of Viral DNA Integration." Cell Host Microbe **24**(3): 392-404.e398.

Akroyd, J., V. J. Fincham, A. R. Green, P. Levantis, S. Searle and J. A. Wyke (1987). "Transcription of Rous sarcoma proviruses in rat cells is determined by chromosomal position effects that fluctuate and can operate over long distances." Oncogene **1**(4): 347-354.

Albanese, A., D. Arosio, M. Terreni and A. Cereseto (2008). "HIV-1 pre-integration complexes selectively target decondensed chromatin in the nuclear periphery." PLoS One **3**(6): e2413.

Andersson, R., C. Gebhard, I. Miguel-Escalada, I. Hoof, J. Bornholdt, M. Boyd, Y. Chen, X. Zhao, C. Schmidl, T. Suzuki, E. Ntini, E. Arner, E. Valen, K. Li, L. Schwarzfischer, D. Glatz, J. Raithel, B. Lilje, N. Rapin, F. O. Bagger, M. Jorgensen, P. R. Andersen, N. Bertin, O. Rackham, A. M. Burroughs, J. K. Baillie, Y. Ishizu, Y. Shimizu, E. Furuhata, S. Maeda, Y. Negishi, C. J. Mungall, T. F. Meehan, T. Lassmann, M. Itoh, H. Kawaji, N. Kondo, J. Kawai, A. Lennartsson, C. O. Daub, P. Heutink, D. A. Hume, T. H. Jensen, H. Suzuki, Y. Hayashizaki, F. Muller, A. R. R. Forrest, P. Carninci, M. Rehli and A. Sandelin (2014). "An atlas of active enhancers across human cell types and tissues." Nature **507**(7493): 455-461.

Andrake, M. D., M. M. Sauter, K. Boland, A. D. Goldstein, M. Hussein and A. M. Skalka (2008). "Nuclear import of Avian Sarcoma Virus integrase is facilitated by host cell factors." Retrovirology **5**: 73.

Andrake, M. D. and A. M. Skalka (1995). "Multimerization determinants reside in both the catalytic core and C terminus of avian sarcoma virus integrase." J Biol Chem **270**(49): 29299-29306.

Angelov, D., M. Charra, M. Seve, J. Cote, S. Khochbin and S. Dimitrov (2000). "Differential remodeling of the HIV-1 nucleosome upon transcription activators and SWI/SNF complex binding." J Mol Biol **302**(2): 315-326.

Arroyo, J., E. Winchester, B. S. McLellan and B. T. Huber (1997). "Shared promoter elements between a viral superantigen and the major histocompatibility complex class II-associated invariant chain." J Virol **71**(2): 1237-1245.

Ballandras-Colas, A., M. Brown, N. J. Cook, T. G. Dewdney, B. Demeler, P. Cherepanov, D. Lyumkis and A. N. Engelman (2016). "Cryo-EM reveals a novel octameric integrase structure for betaretroviral intasome function." Nature **530**(7590): 358-361.

Ballandras-Colas, A., D. P. Maskell, E. Serrao, J. Locke, P. Swuec, S. R. Jonsson, A. Kotecha, N. J. Cook, V. E. Pye, I. A. Taylor, V. Andresdottir, A. N. Engelman, A. Costa and P. Cherepanov (2017). "A supramolecular assembly mediates lentiviral DNA integration." Science **355**(6320): 93-95.

Barr, S. D., A. Ciuffi, J. Leipzig, P. Shinn, J. R. Ecker and F. D. Bushman (2006). "HIV integration site selection: targeting in macrophages and the effects of different routes of viral entry." Mol Ther **14**(2): 218-225.

Barr, S. D., J. Leipzig, P. Shinn, J. R. Ecker and F. D. Bushman (2005). "Integration targeting by avian sarcoma-leukosis virus and human immunodeficiency virus in the chicken genome." J Virol **79**(18): 12035-12044.

Battivelli, E., M. S. Dahabieh, M. Abdel-Mohsen, J. P. Svensson, I. Tojal Da Silva, L. B. Cohn, A. Gramatica, S. Deeks, W. C. Greene, S. K. Pillai and E. Verdin (2018). "Distinct chromatin functional states correlate with HIV latency reactivation in infected primary CD4(+) T cells." Elife **7**.

Baum, C., K. Itoh, J. Meyer, C. Laker, Y. Ito and W. Ostertag (1997). "The potent enhancer activity of the polycythemic strain of spleen focus-forming virus in hematopoietic cells is governed by a binding site for Sp1 in the upstream control region and by a unique enhancer core motif, creating an exclusive target for PEBP/CBF." J Virol **71**(9): 6323-6331.

Benleulmi, M. S., J. Matysiak, X. Robert, C. Miskey, E. Mauro, D. Lapaillerie, P. Lesbats, S. Chaignepain, D. R. Henriquez, C. Calmels, O. Oladosu, E. Thierry, O. Leon, M. Lavigne, M. L. Andreola, O. Delelis, Z. Ivics, M. Ruff, P. Gouet and V. Parissi (2017). "Modulation of the functional association between the HIV-1 intasome and the nucleosome by histone amino-terminal tails." Retrovirology **14**(1): 54.

Berry, C., S. Hannenhalli, J. Leipzig and F. D. Bushman (2006). "Selection of target sites for mobile DNA integration in the human genome." PLoS Comput Biol **2**(11): e157.

Berry, C. C. (2017). "restrSiteUtils: Restriction Site Distances and Matched Samples. R package version 1.2.8.".

Berry, C. C., C. Nobles, E. Six, Y. Wu, N. Malani, E. Sherman, A. Dryga, J. K. Everett, F. Male, A. Bailey, K. Bittinger, M. J. Drake, L. Caccavelli, P. Bates, S. Hacein-Bey-Abina, M. Cavazzana and F. D. Bushman (2017). "INSPIIRED: Quantification and Visualization Tools for Analyzing Integration Site Distributions." Mol Ther Methods Clin Dev **4**: 17-26.

Blazkova, J., K. Trejbalova, F. Gondois-Rey, P. Halfon, P. Philibert, A. Guiguen, E. Verdin, D. Olive, C. Van Lint, J. Hejnar and I. Hirsch (2009). "CpG methylation controls reactivation of HIV from latency." PLoS Pathog **5**(8): e1000554.

Bonczkowski, P., M. A. De Scheerder, E. Malatinkova, A. Borch, Z. Melkova, R. Koenig, W. De Spiegelaere and L. Vandekerckhove (2016). "Protein expression from unintegrated HIV-1 DNA introduces bias in primary in vitro post-integration latency models." Sci Rep **6**: 38329.

Borrenberghs, D., I. Zurnic, F. De Wit, A. Acke, L. Dirix, A. Cereseto, Z. Debyser and J. Hendrix (2018). "Post-mitotic BET-induced reshaping of integrase quaternary structure supports wild-type MLV integration." Nucleic Acids Res.

Bouyac-Bertoia, M., J. D. Dvorin, R. A. Fouchier, Y. Jenkins, B. E. Meyer, L. I. Wu, M. Emerman and M. H. Malim (2001). "HIV-1 infection requires a functional integrase NLS." Mol Cell **7**(5): 1025-1035.

Bruce, J. W., R. Reddington, E. Mathieu, M. Bracken, J. A. Young and P. Ahlquist (2013). "ZASC1 stimulates HIV-1 transcription elongation by recruiting P-TEFb and TAT to the LTR promoter." PLoS Pathog **9**(10): e1003712.

Bruner, K. M., Z. Wang, F. R. Simonetti, A. M. Bender, K. J. Kwon, S. Sengupta, E. J. Fray, S. A. Beg, A. A. R. Antar, K. M. Jenike, L. N. Bertagnolli, A. A. Capoferri, J. T. Kufera, A. Timmons, C. Nobles, J. Gregg, N. Wada, Y. C. Ho, H. Zhang, J. B. Margolick, J. N. Blankson, S. G. Deeks, F. D. Bushman, J. D. Siliciano, G. M. Laird and R. F. Siliciano (2019). "A quantitative approach for measuring the reservoir of latent HIV-1 proviruses." Nature **566**(7742): 120-125.

Brussel, A. and P. Sonigo (2003). "Analysis of early human immunodeficiency virus type 1 DNA synthesis by use of a new sensitive assay for quantifying integrated provirus." J Virol **77**(18): 10119-10124.

Brzezinski, J. D., R. Felkner, A. Modi, M. Liu and M. J. Roth (2016). "Phosphorylation Requirement of Murine Leukemia Virus p12." J Virol **90**(24): 11208-11219.

Brzezinski, J. D., A. Modi, M. Liu and M. J. Roth (2016). "Repression of the Chromatin-Tethering Domain of Murine Leukemia Virus p12." J Virol **90**(24): 11197-11207.

Buffone, C., A. Martinez-Lopez, T. Fricke, S. Opp, M. Severgnini, I. Cifola, L. Petiti, S. Frabetti, K. Skorupka, K. K. Zadrozny, B. K. Ganser-Pornillos, O. Pornillos, F. Di Nunzio and F. Diaz-Griffero (2018). "Nup153 Unlocks the Nuclear Pore Complex for HIV-1 Nuclear Translocation in Nondividing Cells." J Virol **92**(19).

Bukrinsky, M. I., S. Haggerty, M. P. Dempsey, N. Sharova, A. Adzhubel, L. Spitz, P. Lewis, D. Goldfarb, M. Emerman and M. Stevenson (1993). "A nuclear localization signal within HIV-1 matrix protein that governs infection of non-dividing cells." Nature **365**(6447): 666-669.

Bukrinsky, M. I., N. Sharova, M. P. Dempsey, T. L. Stanwick, A. G. Bukrinskaya, S. Haggerty and M. Stevenson (1992). "Active nuclear import of human immunodeficiency virus type 1 preintegration complexes." Proc Natl Acad Sci U S A **89**(14): 6580-6584.

Burke, C. J., G. Sanyal, M. W. Bruner, J. A. Ryan, R. L. LaFemina, H. L. Robbins, A. S. Zeft, C. R. Middaugh and M. G. Cordingley (1992). "Structural implications of spectroscopic characterization of a putative zinc finger peptide from HIV-1 integrase." J Biol Chem **267**(14): 9639-9644.

Bushman, F. D., A. Engelman, I. Palmer, P. Wingfield and R. Craigie (1993). "Domains of the integrase protein of human immunodeficiency virus type 1 responsible for polynucleotidyl transfer and zinc binding." Proc Natl Acad Sci U S A **90**(8): 3428-3432.

Busschots, K., A. Voet, M. De Maeyer, J. C. Rain, S. Emiliani, R. Benarous, L. Desender, Z. Debyser and F. Christ (2007). "Identification of the LEDGF/p75 binding site in HIV-1 integrase." J Mol Biol **365**(5): 1480-1492.

Butler, S. L., M. S. Hansen and F. D. Bushman (2001). "A quantitative assay for HIV DNA integration in vivo." Nat Med **7**(5): 631-634.

Canonne-Hergaux, F., D. Aunis and E. Schaeffer (1995). "Interactions of the transcription factor AP-1 with the long terminal repeat of different human immunodeficiency virus type 1 strains in Jurkat, glial, and neuronal cells." J Virol **69**(11): 6634-6642.

Cara, A., A. Cereseto, F. Lori and M. S. Reitz, Jr. (1996). "HIV-1 protein expression from synthetic circles of DNA mimicking the extrachromosomal forms of viral DNA." J Biol Chem **271**(10): 5393-5397.

Cavazzana-Calvo, M., E. Payen, O. Negre, G. Wang, K. Hehir, F. Fusil, J. Down, M. Denaro, T. Brady, K. Westerman, R. Cavallesco, B. Gillet-Legrand, L. Caccavelli, R. Sgarra, L. Maouche-Chretien, F. Bernaudin, R. Girot, R. Dorazio, G. J. Mulder, A. Polack, A. Bank, J. Soulier, J. Larghero, N. Kabbara, B. Dalle, B. Gourmel, G. Socie, S. Chretien, N. Cartier, P. Aubourg, A. Fischer, K. Cornetta, F. Galacteros, Y. Beuzard, E. Gluckman, F. Bushman, S. Hacein-Bey-Abina and P. Leboulch (2010). "Transfusion independence and HMGA2 activation after gene therapy of human beta-thalassaemia." Nature **467**(7313): 318-322.

Cinghu, S., P. Yang, J. P. Kosak, A. E. Conway, D. Kumar, A. J. Oldfield, K. Adelman and R. Jothi (2017). "Intragenic Enhancers Attenuate Host Gene Expression." Molecular cell **68**(1): 104-117.e106.

Ciuffi, A., M. Llano, E. Poeschla, C. Hoffmann, J. Leipzig, P. Shinn, J. R. Ecker and F. Bushman (2005). "A role for LEDGF/p75 in targeting HIV DNA integration." Nat Med **11**(12): 1287-1289.

Ciuffi, A., R. S. Mitchell, C. Hoffmann, J. Leipzig, P. Shinn, J. R. Ecker and F. D. Bushman (2006). "Integration site selection by HIV-based vectors in dividing and growth-arrested IMR-90 lung fibroblasts." Mol Ther **13**(2): 366-373.

Cooney, A. J., S. Y. Tsai, B. W. O'Malley and M. J. Tsai (1991). "Chicken ovalbumin upstream promoter transcription factor binds to a negative regulatory region in the human immunodeficiency virus type 1 long terminal repeat." J Virol **65**(6): 2853-2860.

Corish, P. and C. Tyler-Smith (1999). "Attenuation of green fluorescent protein half-life in mammalian cells." Protein Eng **12**(12): 1035-1040.

Coull, J. J., F. Romerio, J. M. Sun, J. L. Volker, K. M. Galvin, J. R. Davie, Y. Shi, U. Hansen and D. M. Margolis (2000). "The human factors YY1 and LSF repress the human immunodeficiency virus type 1 long terminal repeat via recruitment of histone deacetylase 1." J Virol **74**(15): 6790-6799.

Crise, B., Y. Li, C. Yuan, D. R. Morcock, D. Whitby, D. J. Munroe, L. O. Arthur and X. Wu (2005). "Simian immunodeficiency virus integration preference is similar to that of human immunodeficiency virus type 1." J Virol **79**(19): 12199-12204.

Cron, R. Q., S. R. Bartz, A. Clausell, S. J. Bort, S. J. Klebanoff and D. B. Lewis (2000). "NFAT1 enhances HIV-1 gene expression in primary human CD4 T cells." Clin Immunol **94**(3): 179-191.

Curristin, S. M., K. J. Bird, R. J. Tubbs and A. Ruddell (1997). "VBP and RelA regulate avian leukosis virus long terminal repeat-enhanced transcription in B cells." J Virol **71**(8): 5972-5981.

Dahabieh, M. S., M. Ooms, C. Brumme, J. Taylor, P. R. Harrigan, V. Simon and I. Sadowski (2014). "Direct non-productive HIV-1 infection in a T-cell line is driven by cellular activation state and NFkappaB." Retrovirology **11**: 17.

De Iaco, A., F. Santoni, A. Vannier, M. Guipponi, S. Antonarakis and J. Luban (2013). "TNPO3 protects HIV-1 replication from CPSF6-mediated capsid stabilization in the host cell cytoplasm." Retrovirology **10**: 20.

De Ravin, S. S., L. Su, N. Theobald, U. Choi, J. L. Macpherson, M. Poidinger, G. Symonds, S. M. Pond, A. L. Ferris, S. H. Hughes, H. L. Malech and X. Wu (2014). "Enhancers are major targets for murine leukemia virus vector integration." J Virol **88**(8): 4504-4513.

De Rijck, J., K. Bartholomeeusen, H. Ceulemans, Z. Debyser and R. Gijsbers (2010). "High-resolution profiling of the LEDGF/p75 chromatin interaction in the ENCODE region." Nucleic Acids Res **38**(18): 6135-6147.

De Rijck, J., C. de Kogel, J. Demeulemeester, S. Vets, S. El Ashkar, N. Malani, F. D. Bushman, B. Landuyt, S. J. Husson, K. Busschots, R. Gijsbers and Z. Debyser (2013). "The BET family of proteins targets moloney murine leukemia virus integration near transcription start sites." Cell Rep **5**(4): 886-894.

Deaton, A. M. and A. Bird (2011). "CpG islands and the regulation of transcription." Genes Dev **25**(10): 1010-1022.

Delelis, O., V. Parissi, H. Leh, G. Mbemba, C. Petit, P. Sonigo, E. Deprez and J. F. Mouscadet (2007). "Efficient and specific internal cleavage of a retroviral palindromic DNA sequence by tetrameric HIV-1 integrase." PLoS One **2**(7): e608.

Delelis, O., C. Petit, H. Leh, G. Mbemba, J. F. Mouscadet and P. Sonigo (2005). "A novel function for spumaretrovirus integrase: an early requirement for integrase-mediated cleavage of 2 LTR circles." Retrovirology **2**: 31.

Derse, D., B. Crise, Y. Li, G. Princler, N. Lum, C. Stewart, C. F. McGrath, S. H. Hughes, D. J. Munroe and X. Wu (2007). "Human T-cell leukemia virus type 1 integration target sites in the human genome: comparison with those of other retroviruses." J Virol **81**(12): 6731-6741.

Di Primio, C., V. Quercioli, A. Allouch, R. Gijsbers, F. Christ, Z. Debyser, D. Arosio and A. Cereseto (2013). "Single-cell imaging of HIV-1 provirus (SCIP)." Proc Natl Acad Sci U S A **110**(14): 5636-5641.

Dobard, C. W., M. S. Briones and S. A. Chow (2007). "Molecular mechanisms by which human immunodeficiency virus type 1 integrase stimulates the early steps of reverse transcription." J Virol **81**(18): 10037-10046.

Doitsh, G., M. Cavrois, K. G. Lassen, O. Zepeda, Z. Yang, M. L. Santiago, A. M. Hebbeler and W. C. Greene (2010). "Abortive HIV Infection Mediates CD4 T Cell Depletion and Inflammation in Human Lymphoid Tissue." Cell **143**(5): 789-801.

Drelich, M., R. Wilhelm and J. Mous (1992). "Identification of amino acid residues critical for endonuclease and integration activities of HIV-1 IN protein in vitro." Virology **188**(2): 459-468.

du Chene, I., E. Basyuk, Y. L. Lin, R. Triboulet, A. Knezevich, C. Chable-Bessia, C. Mettling, V. Baillat, J. Reynes, P. Corbeau, E. Bertrand, A. Marcello, S. Emiliani, R. Kiernan and M. Benkirane (2007). "Suv39H1 and HP1gamma are responsible for chromatin-mediated HIV-1 transcriptional silencing and post-integration latency." Embo j **26**(2): 424-435.

Eidahl, J. O., B. L. Crowe, J. A. North, C. J. McKee, N. Shkriabai, L. Feng, M. Plumb, R. L. Graham, R. J. Gorelick, S. Hess, M. G. Poirier, M. P. Foster and M. Kvaratskhelia (2013). "Structural basis for high-affinity binding of LEDGF PWWP to mononucleosomes." Nucleic Acids Res **41**(6): 3924-3936.

El Ashkar, S., J. De Rijck, J. Demeulemeester, S. Vets, P. Madlala, K. Cermakova, Z. Debyser and R. Gijsbers (2014). "BET-independent MLV-based Vectors Target Away From Promoters and Regulatory Elements." Mol Ther Nucleic Acids **3**: e179.

El Ashkar, S., D. Van Looveren, F. Schenk, L. S. Vranckx, J. Demeulemeester, J. De Rijck, Z. Debyser, U. Modlich and R. Gijsbers (2017). "Engineering Next-Generation BET-Independent MLV Vectors for Safer Gene Therapy." Mol Ther Nucleic Acids **7**: 231-245.

Elis, E., M. Ehrlich, A. Prizan-Ravid, N. Laham-Karam and E. Bacharach (2012). "p12 tethers the murine leukemia virus pre-integration complex to mitotic chromosomes." PLoS Pathog **8**(12): e1003103.

Elleder, D., A. Pavlicek, J. Paces and J. Hejnar (2002). "Preferential integration of human immunodeficiency virus type 1 into genes, cytogenetic R bands and GC-rich DNA regions: insight from the human genome sequence." FEBS Lett **517**(1-3): 285-286.

Emery, A., S. Zhou, E. Pollom and R. Swanstrom (2017). "Characterizing HIV-1 Splicing by Using Next-Generation Sequencing." J Virol **91**(6).

Emiliani, S., A. Mousnier, K. Busschots, M. Maroun, B. Van Maele, D. Tempe, L. Vandekerckhove, F. Moisant, L. Ben-Slama, M. Witvrouw, F. Christ, J. C. Rain, C. Dargemont, Z. Debyser and R. Benarous (2005). "Integrase mutants defective for interaction with LEDGF/p75 are impaired in chromosome tethering and HIV-1 replication." J Biol Chem **280**(27): 25517-25523.

Engelman, A. and R. Craigie (1992). "Identification of conserved amino acid residues critical for human immunodeficiency virus type 1 integrase function in vitro." J Virol **66**(11): 6361-6369.

Engelman, A., A. B. Hickman and R. Craigie (1994). "The core and carboxyl-terminal domains of the integrase protein of human immunodeficiency virus type 1 each contribute to nonspecific DNA binding." J Virol **68**(9): 5911-5917.

Engelman, A. N. and P. Cherepanov (2017). "Retroviral intasomes arising." Curr Opin Struct Biol **47**: 23-29.

Ernst, J. and M. Kellis (2010). "Discovery and characterization of chromatin states for systematic annotation of the human genome." Nat Biotechnol **28**(8): 817-825.

Ernst, J. and M. Kellis (2012). "ChromHMM: automating chromatin-state discovery and characterization." Nat Methods **9**(3): 215-216.

Felber, B. K., M. Hadzopoulou-Cladaras, C. Cladaras, T. Copeland and G. N. Pavlakis (1989). "rev protein of human immunodeficiency virus type 1 affects the stability and transport of the viral mRNA." Proc Natl Acad Sci U S A **86**(5): 1495-1499.

Feng, L., V. Dharmarajan, E. Serrao, A. Hoyte, R. C. Larue, A. Slaughter, A. Sharma, M. R. Plumb, J. J. Kessl, J. R. Fuchs, F. D. Bushman, A. N. Engelman, P. R. Griffin and M. Kvaratskhelia (2016). "The Competitive Interplay between Allosteric HIV-1 Integrase Inhibitor BI/D and LEDGF/p75 during the Early Stage of HIV-1 Replication Adversely Affects Inhibitor Potency." ACS Chem Biol **11**(5): 1313-1321.

Ferris, A. L., X. Wu, C. M. Hughes, C. Stewart, S. J. Smith, T. A. Milne, G. G. Wang, M. C. Shun, C. D. Allis, A. Engelman and S. H. Hughes (2010). "Lens epithelium-derived growth factor fusion proteins redirect HIV-1 DNA integration." Proc Natl Acad Sci U S A **107**(7): 3135-3140.

Fincham, V. J. and J. A. Wyke (1991). "Differences between cellular integration sites of transcribed and nontranscribed Rous sarcoma proviruses." J Virol **65**(1): 461-463.

Fischinger, P. F., N. Tuttle-Fuller, G. Huper and D. P. Bolognesi (1975). "Mitosis is required for production of murine leukemia virus and structural proteins during de novo infection." J Virol **16**(2): 267-274.

Fuhrman, S. A., C. Van Beveren and I. M. Verma (1981). "Identification of a RNA polymerase II initiation site in the long terminal repeat of Moloney murine leukemia viral DNA." Proc Natl Acad Sci U S A **78**(9): 5411-5415.

Gallay, P., T. Hope, D. Chin and D. Trono (1997). "HIV-1 infection of nondividing cells through the recognition of integrase by the importin/karyopherin pathway." Proc Natl Acad Sci U S A **94**(18): 9825-9830.

Galy, A. (2017). "Major Advances in the Development of Vectors for Clinical Gene Therapy of Hematopoietic Stem Cells from European Groups over the Last 25 Years." Hum Gene Ther **28**(11): 964-971.

Gessain, A. and O. Cassar (2012). "Epidemiological Aspects and World Distribution of HTLV-1 Infection." Front Microbiol **3**: 388.

Gijsbers, R., K. Ronen, S. Vets, N. Malani, J. De Rijck, M. McNeely, F. D. Bushman and Z. Debyser (2010). "LEDGF hybrids efficiently retarget lentiviral integration into heterochromatin." Mol Ther **18**(3): 552-560.

Gogol-Doring, A., I. Ammar, S. Gupta, M. Bunse, C. Miskey, W. Chen, W. Uckert, T. F. Schulz, Z. Izsvak and Z. Ivics (2016). "Genome-wide Profiling Reveals Remarkable Parallels Between Insertion Site Selection Properties of the MLV Retrovirus and the piggyBac Transposon in Primary Human CD4(+) T Cells." Mol Ther **24**(3): 592-606.

Gorman, C. M., G. T. Merlino, M. C. Willingham, I. Pastan and B. H. Howard (1982). "The Rous sarcoma virus long terminal repeat is a strong promoter when introduced into a variety of eukaryotic cells by DNA-mediated transfection." Proc Natl Acad Sci U S A **79**(22): 6777-6781.

Grawenhoff, J. and A. N. Engelman (2017). "Retroviral integrase protein and intasome nucleoprotein complex structures." World J Biol Chem **8**(1): 32-44.

Green, A. R., C. J. Poole, S. M. Povey, D. Rowe, S. Searle and J. A. Wyke (1990). "Fusion of Rous-sarcoma-virus-transformed rat cells to morphologically normal human or rat cells results in transcriptional suppression of the provirus that depends on its chromosomal integration site." Int J Cancer **46**(2): 220-227.

Greuel, B. T., L. Sealy and J. E. Majors (1990). "Transcriptional activity of the Rous sarcoma virus long terminal repeat correlates with binding of a factor to an upstream CCAAT box in vitro." Virology **177**(1): 33-43.

Guntaka, R. V., O. C. Richards, P. R. Shank, H. J. Kung and N. Davidson (1976). "Covalently closed circular DNA of avian sarcoma virus: purification from nuclei of infected quail tumor cells and measurement by electron microscopy and gel electrophoresis." J Mol Biol **106**(2): 337-357.

Gupta, S. S., T. Maetzig, G. N. Maertens, A. Sharif, M. Rothe, M. Weidner-Glunde, M. Galla, A. Schambach, P. Cherepanov and T. F. Schulz (2013). "Bromo- and extraterminal domain chromatin regulators serve as cofactors for murine leukemia virus integration." J Virol **87**(23): 12721-12736.

Hacein-Bey-Abina, S., A. Fischer and M. Cavazzana-Calvo (2002). "Gene therapy of X-linked severe combined immunodeficiency." Int J Hematol **76**(4): 295-298.

Hacein-Bey-Abina, S., C. Von Kalle, M. Schmidt, M. P. McCormack, N. Wulffraat, P. Leboulch, A. Lim, C. S. Osborne, R. Pawliuk, E. Morillon, R. Sorensen, A. Forster, P. Fraser, J. I. Cohen, G. de Saint Basile, I. Alexander, U. Wintergerst, T. Frebourg, A. Aurias, D. Stoppa-Lyonnet, S. Romana, I. Radford-Weiss, F. Gross, F. Valensi, E. Delabesse, E. Macintyre, F. Sigaux, J. Soulier, L. E. Leiva, M. Wissler, C. Prinz, T. H. Rabbitts, F. Le Deist, A. Fischer and M. Cavazzana-Calvo (2003). "LMO2-associated clonal T cell proliferation in two patients after gene therapy for SCID-X1." Science **302**(5644): 415-419.

Han, Y., Y. B. Lin, W. An, J. Xu, H. C. Yang, K. O'Connell, D. Dordai, J. D. Boeke, J. D. Siliciano and R. F. Siliciano (2008). "Orientation-dependent regulation of integrated HIV-1 expression by host gene transcriptional readthrough." Cell Host Microbe **4**(2): 134-146.

Hare, S., S. S. Gupta, E. Valkov, A. Engelman and P. Cherepanov (2010). "Retroviral intasome assembly and inhibition of DNA strand transfer." Nature **464**(7286): 232-236.

Hare, S., G. N. Maertens and P. Cherepanov (2012). "3'-processing and strand transfer catalysed by retroviral integrase in crystallo." The EMBO journal **31**(13): 3020-3028.

Harrich, D., J. Garcia, F. Wu, R. Mitsuyasu, J. Gonazalez and R. Gaynor (1989). "Role of SP1-binding domains in in vivo transcriptional regulation of the human immunodeficiency virus type 1 long terminal repeat." J Virol **63**(6): 2585-2591.

Hatziioannou, T. and S. P. Goff (2001). "Infection of nondividing cells by Rous sarcoma virus." J Virol **75**(19): 9526-9531.

He, G. and D. M. Margolis (2002). "Counterregulation of chromatin deacetylation and histone deacetylase occupancy at the integrated promoter of human immunodeficiency virus type 1 (HIV-1) by the HIV-1 repressor YY1 and HIV-1 activator Tat." Mol Cell Biol **22**(9): 2965-2973.

Heinzinger, N. K., M. I. Bukrinsky, S. A. Haggerty, A. M. Ragland, V. Kewalramani, M. A. Lee, H. E. Gendelman, L. Ratner, M. Stevenson and M. Emerman (1994). "The Vpr protein of human immunodeficiency virus type 1 influences nuclear localization of viral nucleic acids in nondividing host cells." Proc Natl Acad Sci U S A **91**(15): 7311-7315.

Hejnar, J., P. Hajkova, J. Plachy, D. Elleder, V. Stepanets and J. Svoboda (2001). "CpG island protects Rous sarcoma virus-derived vectors integrated into nonpermissive cells from DNA methylation and transcriptional suppression." Proc Natl Acad Sci U S A **98**(2): 565-569.

Hejnar, J., J. Plachy, J. Geryk, O. Machon, K. Trejbalova, R. V. Guntaka and J. Svoboda (1999). "Inhibition of the rous sarcoma virus long terminal repeat-driven transcription by in vitro methylation: different sensitivity in permissive chicken cells versus mammalian cells." Virology **255**(1): 171-181.

Hejnar, J., J. Svoboda, J. Geryk, V. J. Fincham and R. Hak (1994). "High rate of morphological reversion in tumor cell line H-19 associated with permanent transcriptional suppression of the LTR, v-src, LTR provirus." Cell Growth Differ **5**(3): 277-285.

Hematti, P., B. K. Hong, C. Ferguson, R. Adler, H. Hanawa, S. Sellers, I. E. Holt, C. E. Eckfeldt, Y. Sharma, M. Schmidt, C. von Kalle, D. A. Persons, E. M. Billings, C. M. Verfaillie, A. W. Nienhuis, T. G. Wolfsberg, C. E. Dunbar and B. Calmels (2004). "Distinct genomic integration of MLV and SIV vectors in primate hematopoietic stem and progenitor cells." PLoS Biol **2**(12): e423.

Henderson, A., A. Holloway, R. Reeves and D. J. Tremethick (2004). "Recruitment of SWI/SNF to the human immunodeficiency virus type 1 promoter." Mol Cell Biol **24**(1): 389-397.

Henderson, A. J., X. Zou and K. L. Calame (1995). "C/EBP proteins activate transcription from the human immunodeficiency virus type 1 long terminal repeat in macrophages/monocytes." J Virol **69**(9): 5337-5344.

Ho, Y.-C., L. Shan, Nina N. Hosmane, J. Wang, Sarah B. Laskey, Daniel I. S. Rosenbloom, J. Lai, Joel N. Blankson, Janet D. Siliciano and Robert F. Siliciano (2013). "Replication-Competent Noninduced Proviruses in the Latent Reservoir Increase Barrier to HIV-1 Cure." Cell **155**(3): 540-551.

Hoffman, M. M., O. J. Buske, J. Wang, Z. Weng, J. A. Bilmes and W. S. Noble (2012). "Unsupervised pattern discovery in human chromatin structure through genomic segmentation." Nat Methods **9**(5): 473-476.

Hoffman, M. M., J. Ernst, S. P. Wilder, A. Kundaje, R. S. Harris, M. Libbrecht, B. Giardine, P. M. Ellenbogen, J. A. Bilmes, E. Birney, R. C. Hardison, I. Dunham, M. Kellis and W. S. Noble (2013). "Integrative annotation of chromatin elements from ENCODE data." Nucleic Acids Res **41**(2): 827-841.

Hombrouck, A., J. De Rijck, J. Hendrix, L. Vandekerckhove, A. Voet, M. De Maeyer, M. Witvrouw, Y. Engelborghs, F. Christ, R. Gijsbers and Z. Debyser (2007). "Virus evolution reveals an exclusive role for LEDGF/p75 in chromosomal tethering of HIV." PLoS Pathog **3**(3): e47.

Horiba, M., L. B. Martinez, J. L. Buescher, S. Sato, J. Limoges, Y. Jiang, C. Jones and T. Ikezu (2007). "OTK18, a zinc-finger protein, regulates human immunodeficiency virus type 1 long terminal repeat through two distinct regulatory regions." J Gen Virol **88**(Pt 1): 236-241.

Hossain, A., K. Ali and C. G. Shin (2014). "Nuclear localization signals in prototype foamy viral integrase for successive infection and replication in dividing cells." Mol Cells **37**(2): 140-148.

Houtz, E. K. and K. F. Conklin (1996). "Identification of EFIV, a stable factor present in many avian cell types that transactivates sequences in the 5' portion of the Rous sarcoma virus long terminal repeat enhancer." J Virol **70**(1): 393-401.

Hronek, B. W., A. Meagher, J. Rovnak and S. L. Quackenbush (2004). "Identification and characterization of cis-acting elements residing in the walleye dermal sarcoma virus promoter." J Virol **78**(14): 7590-7601.

Humphries, E. H., C. Glover and M. E. Reichmann (1981). "Rous sarcoma virus infection of synchronized cells establishes provirus integration during S-phase DNA synthesis prior to cellular division." Proc Natl Acad Sci U S A **78**(4): 2601-2605.

Humphries, E. H. and H. M. Temin (1972). "Cell cycle-dependent activation of rous sarcoma virus-infected stationary chicken cells: avian leukosis virus group-specific antigens and ribonucleic acid." J Virol **10**(1): 82-87.

Humphries, E. H. and H. M. Temin (1974). "Requirement for cell division for initiation of transcription of Rous sarcoma virus RNA." J Virol **14**(3): 531-546.

Chan, C. N., B. Trinite, C. S. Lee, S. Mahajan, A. Anand, D. Wodarz, S. Sabbaj, A. Bansal, P. A. Goepfert and D. N. Levy (2016). "HIV-1 latency and virus production from unintegrated genomes following direct infection of resting CD4 T cells." Retrovirology **13**: 1.

Chatis, P. A., C. A. Holland, J. W. Hartley, W. P. Rowe and N. Hopkins (1983). "Role for the 3' end of the genome in determining disease specificity of Friend and Moloney murine leukemia viruses." Proc Natl Acad Sci U S A **80**(14): 4408-4411.

Chaudhary, P., S. Z. Khan, P. Rawat, T. Augustine, D. A. Raynes, V. Guerriero and D. Mitra (2016). "HSP70 binding protein 1 (HspBP1) suppresses HIV-1 replication by inhibiting NF-kappaB mediated activation of viral gene expression." Nucleic Acids Res **44**(4): 1613-1629.

Chavez, L., V. Calvanese and E. Verdin (2015). "HIV Latency Is Established Directly and Early in Both Resting and Activated Primary CD4 T Cells." PLoS Pathog **11**(6): e1004955.

Chen, H. C., J. P. Martinez, E. Zorita, A. Meyerhans and G. J. Filion (2017). "Position effects influence HIV latency reversal." Nat Struct Mol Biol **24**(1): 47-54.

Cherepanov, P. (2007). "LEDGF/p75 interacts with divergent lentiviral integrases and modulates their enzymatic activity in vitro." Nucleic Acids Res **35**(1): 113-124.

Cherepanov, P., A. L. Ambrosio, S. Rahman, T. Ellenberger and A. Engelman (2005). "Structural basis for the recognition between HIV-1 integrase and transcriptional coactivator p75." Proc Natl Acad Sci U S A **102**(48): 17308-17313.

Cherepanov, P., E. Devroe, P. A. Silver and A. Engelman (2004). "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 **279**(47): 48883-48892.

Cherepanov, P., G. Maertens, P. Proost, B. Devreese, J. Van Beeumen, Y. Engelborghs, E. De Clercq and Z. Debyser (2003). "HIV-1 integrase forms stable tetramers and associates with LEDGF/p75 protein in human cells." J Biol Chem **278**(1): 372-381.

Chin, C. R., J. M. Perreira, G. Savidis, J. M. Portmann, A. M. Aker, E. M. Feeley, M. C. Smith and A. L. Brass (2015). "Direct Visualization of HIV-1 Replication Intermediates Shows that Capsid and CPSF6 Modulate HIV-1 Intra-nuclear Invasion and Integration." Cell Rep **13**(8): 1717-1731.

Chiswell, D. J., D. A. Gillespie and J. A. Wyke (1982). "The changes in proviral chromatin that accompany morphological variation in avian sarcoma virus-infected rat cells." Nucleic Acids Res **10**(13): 3967-3980.

Chomont, N., M. El-Far, P. Ancuta, L. Trautmann, F. A. Procopio, B. Yassine-Diab, G. Boucher, M. R. Boulassel, G. Ghattas, J. M. Brenchley, T. W. Schacker, B. J. Hill, D. C. Douek, J. P. Routy, E. K. Haddad and R. P. Sekaly (2009). "HIV reservoir size and persistence are driven by T cell survival and homeostatic proliferation." Nat Med **15**(8): 893-900.

Christ, F., A. Voet, A. Marchand, S. Nicolet, B. A. Desimmie, D. Marchand, D. Bardiot, N. J. Van der Veken, B. Van Remoortel, S. V. Strelkov, M. De Maeyer, P. Chaltin and Z. Debyser (2010). "Rational design of small-molecule inhibitors of the LEDGF/p75-integrase interaction and HIV replication." Nat Chem Biol **6**(6): 442-448.

Indik, S., W. H. Gunzburg, B. Salmons and F. Rouault (2005). "A novel, mouse mammary tumor virus encoded protein with Rev-like properties." Virology **337**(1): 1-6.

Jenkins, T. M., A. Engelman, R. Ghirlando and R. Craigie (1996). "A soluble active mutant of HIV-1 integrase: involvement of both the core and carboxyl-terminal domains in multimerization." J Biol Chem **271**(13): 7712-7718.

Jones, K. A., J. T. Kadonaga, P. A. Luciw and R. Tjian (1986). "Activation of the AIDS retrovirus promoter by the cellular transcription factor, Sp1." Science **232**(4751): 755-759.

Jordan, A., D. Bisgrove and E. Verdin (2003). "HIV reproducibly establishes a latent infection after acute infection of T cells in vitro." Embo j **22**(8): 1868-1877.

Jordan, A., P. Defechereux and E. Verdin (2001). "The site of HIV-1 integration in the human genome determines basal transcriptional activity and response to Tat transactivation." Embo j **20**(7): 1726-1738.

Jurado, K. A., H. Wang, A. Slaughter, L. Feng, J. J. Kessl, Y. Koh, W. Wang, A. Ballandras-Colas, P. A. Patel, J. R. Fuchs, M. Kvaratskhelia and A. Engelman (2013). "Allosteric integrase inhibitor potency is determined through the inhibition of HIV-1 particle maturation." Proc Natl Acad Sci U S A **110**(21): 8690-8695.

Kaczmarek Michaels, K., F. Wolschendorf, G. M. Schiralli Lester, M. Natarajan, O. Kutsch and A. J. Henderson (2015). "RNAP II processivity is a limiting step for HIV-1 transcription independent of orientation to and activity of endogenous neighboring promoters." Virology **486**: 7-14.

Kalina, J., F. Senigl, A. Micakova, J. Mucksova, J. Blazkova, H. Yan, M. Poplstein, J. Hejnar and P. Trefil (2007). "Retrovirus-mediated in vitro gene transfer into chicken male germ line cells." Reproduction **134**(3): 445-453.

Kanamori-Katayama, M., M. Itoh, H. Kawaji, T. Lassmann, S. Katayama, M. Kojima, N. Bertin, A. Kaiho, N. Ninomiya, C. O. Daub, P. Carninci, A. R. Forrest and Y. Hayashizaki (2011). "Unamplified cap analysis of gene expression on a single-molecule sequencer." Genome Res **21**(7): 1150-1159.

Kang, Y., C. J. Moressi, T. E. Scheetz, L. Xie, D. T. Tran, T. L. Casavant, P. Ak, C. J. Benham, B. L. Davidson and P. B. McCray, Jr. (2006). "Integration site choice of a feline immunodeficiency virus vector." J Virol **80**(17): 8820-8823.

Kao, S. Y., A. F. Calman, P. A. Luciw and B. M. Peterlin (1987). "Anti-termination of transcription within the long terminal repeat of HIV-1 by tat gene product." Nature **330**(6147): 489-493.

Katz, R. A., J. G. Greger, K. Darby, P. Boimel, G. F. Rall and A. M. Skalka (2002). "Transduction of interphase cells by avian sarcoma virus." J Virol **76**(11): 5422-5434.

Katz, R. A., S. A. Mitsialis and R. V. Guntaka (1983). "Studies on the methylation of avian sarcoma proviruses in permissive and non-permissive cells." J Gen Virol **64 (Pt 2)**: 429-435.

Keller, A., K. M. Partin, M. Löchelt, H. Bannert, R. M. Flügel and B. R. Cullen (1991). "Characterization of the transcriptional trans activator of human foamy retrovirus." Journal of virology **65**(5): 2589-2594.

Kessl, J. J., S. B. Kutluay, D. Townsend, S. Rebensburg, A. Slaughter, R. C. Larue, N. Shkriabai, N. Bakouche, J. R. Fuchs, P. D. Bieniasz and M. Kvaratskhelia (2016). "HIV-1 Integrase Binds the Viral RNA Genome and Is Essential during Virion Morphogenesis." Cell **166**(5): 1257-1268.e1212.

Khan, E., J. P. Mack, R. A. Katz, J. Kulkosky and A. M. Skalka (1991). "Retroviral integrase domains: DNA binding and the recognition of LTR sequences." Nucleic Acids Res **19**(4): 851-860.

Kilzer, J. M., T. Stracker, B. Beitzel, K. Meek, M. Weitzman and F. D. Bushman (2003). "Roles of host cell factors in circularization of retroviral dna." Virology **314**(1): 460-467.

Kim, T. K. and R. Shiekhattar (2015). "Architectural and Functional Commonalities between Enhancers and Promoters." Cell **162**(5): 948-959.

Konstantoulas, C. J. and S. Indik (2014). "Mouse mammary tumor virus-based vector transduces non-dividing cells, enters the nucleus via a TNPO3-independent pathway and integrates in a less biased fashion than other retroviruses." Retrovirology **11**: 34.

Kukolj, G., K. S. Jones and A. M. Skalka (1997). "Subcellular localization of avian sarcoma virus and human immunodeficiency virus type 1 integrases." J Virol **71**(1): 843-847.

Kukolj, G., R. A. Katz and A. M. Skalka (1998). "Characterization of the nuclear localization signal in the avian sarcoma virus integrase." Gene **223**(1-2): 157-163.

Kulkosky, J., K. S. Jones, R. A. Katz, J. P. Mack and A. M. Skalka (1992). "Residues critical for retroviral integrative recombination in a region that is highly conserved among retroviral/retrotransposon integrases and bacterial insertion sequence transposases." Mol Cell Biol **12**(5): 2331-2338.

LaFave, M. C., G. K. Varshney, D. E. Gildea, T. G. Wolfsberg, A. D. Baxevanis and S. M. Burgess (2014). "MLV integration site selection is driven by strong enhancers and active promoters." Nucleic Acids Res **42**(7): 4257-4269.

Lai, L., H. Liu, X. Wu and J. C. Kappes (2001). "Moloney murine leukemia virus integrase protein augments viral DNA synthesis in infected cells." Journal of virology **75**(23): 11365-11372.

Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, Y. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J. C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R. H. Waterston, R. K. Wilson, L. W. Hillier, J. D. McPherson, M. A. Marra, E. R. Mardis, L. A. Fulton, A. T. Chinwalla, K. H. Pepin, W. R. Gish, S. L. Chissoe, M. C. Wendl, K. D. Delehaunty, T. L. Miner, A. Delehaunty, J. B. Kramer, L. L. Cook, R. S. Fulton, D. L. Johnson, P. J. Minx, S. W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J. F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, R. A. Gibbs, D. M. Muzny, S. E. Scherer, J. B. Bouck, E. J. Sodergren, K. C. Worley, C. M. Rives, J. H. Gorrell, M. L. Metzker, S. L. Naylor, R. S.

Kucherlapati, D. L. Nelson, G. M. Weinstock, Y. Sakaki, A. Fujiyama, M. Hattori, T. Yada, A. Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissenbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, D. R. Smith, L. Doucette-Stamm, M. Rubenfield, K. Weinstock, H. M. Lee, J. Dubois, A. Rosenthal, M. Platzer, G. Nyakatura, S. Taudien, A. Rump, H. Yang, J. Yu, J. Wang, G. Huang, J. Gu, L. Hood, L. Rowen, A. Madan, S. Qin, R. W. Davis, N. A. Federspiel, A. P. Abola, M. J. Proctor, R. M. Myers, J. Schmutz, M. Dickson, J. Grimwood, D. R. Cox, M. V. Olson, R. Kaul, C. Raymond, N. Shimizu, K. Kawasaki, S. Minoshima, G. A. Evans, M. Athanasiou, R. Schultz, B. A. Roe, F. Chen, H. Pan, J. Ramser, H. Lehrach, R. Reinhardt, W. R. McCombie, M. de la Bastide, N. Dedhia, H. Blocker, K. Hornischer, G. Nordsiek, R. Agarwala, L. Aravind, J. A. Bailey, A. Bateman, S. Batzoglou, E. Birney, P. Bork, D. G. Brown, C. B. Burge, L. Cerutti, H. C. Chen, D. Church, M. Clamp, R. R. Copley, T. Doerks, S. R. Eddy, E. E. Eichler, T. S. Furey, J. Galagan, J. G. Gilbert, C. Harmon, Y. Hayashizaki, D. Haussler, H. Hermjakob, K. Hokamp, W. Jang, L. S. Johnson, T. A. Jones, S. Kasif, A. Kaspryzk, S. Kennedy, W. J. Kent, P. Kitts, E. V. Koonin, I. Korf, D. Kulp, D. Lancet, T. M. Lowe, A. McLysaght, T. Mikkelsen, J. V. Moran, N. Mulder, V. J. Pollara, C. P. Ponting, G. Schuler, J. Schultz, G. Slater, A. F. Smit, E. Stupka, J. Szustakowki, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, A. Williams, Y. I. Wolf, K. H. Wolfe, S. P. Yang, R. F. Yeh, F. Collins, M. S. Guyer, J. Peterson, A. Felsenfeld, K. A. Wetterstrand, A. Patrinos, M. J. Morgan, P. de Jong, J. J. Catanese, K. Osoegawa, H. Shizuya, S. Choi, Y. J. Chen and J. Szustakowki (2001). "Initial sequencing and analysis of the human genome." Nature **409**(6822): 860-921.

Lang, A., V. J. Fincham and J. A. Wyke (1993). "Factors influencing physiological variations in the activity of the Rous sarcoma virus long terminal repeat." Virology **196**(2): 564-575.

Lawrence, M., W. Huber, H. Pages, P. Aboyoun, M. Carlson, R. Gentleman, M. T. Morgan and V. J. Carey (2013). "Software for computing and annotating genomic ranges." PLoS Comput Biol **9**(8): e1003118.

Lee, K., Z. Ambrose, T. D. Martin, I. Oztop, A. Mulky, J. G. Julias, N. Vandegraaff, J. G. Baumann, R. Wang, W. Yuen, T. Takemura, K. Shelton, I. Taniuchi, Y. Li, J. Sodroski, D. R. Littman, J. M. Coffin, S. H. Hughes, D. Unutmaz, A. Engelman and V. N. KewalRamani (2010). "Flexible use of nuclear import pathways by HIV-1." Cell Host Microbe **7**(3): 221-233.

Lee, K., A. Mulky, W. Yuen, T. D. Martin, N. R. Meyerson, L. Choi, H. Yu, S. L. Sawyer and V. N. Kewalramani (2012). "HIV-1 capsid-targeting domain of cleavage and polyadenylation specificity factor 6." J Virol **86**(7): 3851-3860.

Lenasi, T., X. Contreras and B. M. Peterlin (2008). "Transcriptional interference antagonizes proviral gene expression to promote HIV latency." Cell Host Microbe **4**(2): 123-133.

Lenz, J., D. Celander, R. L. Crowther, R. Patarca, D. W. Perkins and W. A. Haseltine (1984). "Determination of the leukaemogenicity of a murine retrovirus by sequences within the long terminal repeat." Nature **308**(5958): 467-470.

Lesbats, P., E. Serrao, D. P. Maskell, V. E. Pye, N. O'Reilly, D. Lindemann, A. N. Engelman and P. Cherepanov (2017). "Structural basis for spumavirus GAG tethering to chromatin." Proceedings of the National Academy of Sciences of the United States of America **114**(21): 5509-5514.

Leung, D. C., K. B. Dong, I. A. Maksakova, P. Goyal, R. Appanah, S. Lee, M. Tachibana, Y. Shinkai, B. Lehnertz, D. L. Mager, F. Rossi and M. C. Lorincz (2011). "Lysine methyltransferase G9a is required

for de novo DNA methylation and the establishment, but not the maintenance, of proviral silencing." Proc Natl Acad Sci U S A **108**(14): 5718-5723.

Lewinski, M. K., D. Bisgrove, P. Shinn, H. Chen, C. Hoffmann, S. Hannenhalli, E. Verdin, C. C. Berry, J. R. Ecker and F. D. Bushman (2005). "Genome-wide analysis of chromosomal features repressing human immunodeficiency virus transcription." J Virol **79**(11): 6610-6619.

Lewis, P., M. Hensel and M. Emerman (1992). "Human immunodeficiency virus infection of cells arrested in the cell cycle." Embo j **11**(8): 3053-3058.

Lewis, P. F. and M. Emerman (1994). "Passage through mitosis is required for oncoretroviruses but not for the human immunodeficiency virus." J Virol **68**(1): 510-516.

Li, G., M. Simm, M. J. Potash and D. J. Volsky (1993). "Human immunodeficiency virus type 1 DNA synthesis, integration, and efficient viral replication in growth-arrested T cells." J Virol **67**(7): 3969-3977.

Li, Y., E. Golemis, J. W. Hartley and N. Hopkins (1987). "Disease specificity of nondefective Friend and Moloney murine leukemia viruses is controlled by a small number of nucleotides." J Virol **61**(3): 693-700.

Liu, H., E. C. Dow, R. Arora, J. T. Kimata, L. M. Bull, R. C. Arduino and A. P. Rice (2006). "Integration of human immunodeficiency virus type 1 in untreated infection occurs preferentially within genes." J Virol **80**(15): 7765-7768.

Llano, M., M. Vanegas, O. Fregoso, D. Saenz, S. Chung, M. Peretz and E. M. Poeschla (2004). "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 **78**(17): 9524-9537.

Llano, M., M. Vanegas, N. Hutchins, D. Thompson, S. Delgado and E. M. Poeschla (2006). "Identification and characterization of the chromatin-binding domains of the HIV-1 integrase interactor LEDGF/p75." J Mol Biol **360**(4): 760-773.

Lochelt, M., W. Muranyi and R. M. Flugel (1993). "Human foamy virus genome possesses an internal, Bel-1-dependent and functional promoter." Proc Natl Acad Sci U S A **90**(15): 7317-7321.

Lounkova, A., E. Draberova, F. Senigl, K. Trejbalova, J. Geryk, J. Hejnar and J. Svoboda (2014). "Molecular events accompanying rous sarcoma virus rescue from rodent cells and the role of viral gene complementation." J Virol **88**(6): 3505-3515.

Lusic, M., A. Marcello, A. Cereseto and M. Giacca (2003). "Regulation of HIV-1 gene expression by histone acetylation and factor recruitment at the LTR promoter." Embo j **22**(24): 6550-6561.

MacNeil, A., J. L. Sankale, S. T. Meloni, A. D. Sarr, S. Mboup and P. Kanki (2006). "Genomic sites of human immunodeficiency virus type 2 (HIV-2) integration: similarities to HIV-1 in vitro and possible differences in vivo." J Virol **80**(15): 7316-7321.

Maertens, G., P. Cherepanov, W. Pluymers, K. Busschots, E. De Clercq, Z. Debyser and Y. Engelborghs (2003). "LEDGF/p75 is essential for nuclear and chromosomal targeting of HIV-1 integrase in human cells." J Biol Chem **278**(35): 33528-33539.

Maertens, G. N., S. Hare and P. Cherepanov (2010). "The mechanism of retroviral integration from X-ray structures of its key intermediates." Nature **468**(7321): 326-329.

Machon, O., V. Strmen, J. Hejnar, J. Geryk and J. Svoboda (1998). "Sp1 binding sites inserted into the rous sarcoma virus long terminal repeat enhance LTR-driven gene expression." Gene **208**(1): 73-82.

Maldarelli, F., X. Wu, L. Su, F. R. Simonetti, W. Shao, S. Hill, J. Spindler, A. L. Ferris, J. W. Mellors, M. F. Kearney, J. M. Coffin and S. H. Hughes (2014). "HIV latency. Specific HIV integration sites are linked to clonal expansion and persistence of infected cells." Science **345**(6193): 179-183.

Malim, M. H., J. Hauber, S. Y. Le, J. V. Maizel and B. R. Cullen (1989). "The HIV-1 rev trans-activator acts through a structured target sequence to activate nuclear export of unspliced viral mRNA." Nature **338**(6212): 254-257.

Mansky, L. M. and H. M. Temin (1995). "Lower in vivo mutation rate of human immunodeficiency virus type 1 than that predicted from the fidelity of purified reverse transcriptase." J Virol **69**(8): 5087-5094.

Marban, C., S. Suzanne, F. Dequiedt, S. de Walque, L. Redel, C. Van Lint, D. Aunis and O. Rohr (2007). "Recruitment of chromatin-modifying enzymes by CTIP2 promotes HIV-1 transcriptional silencing." Embo j **26**(2): 412-423.

Marini, B., A. Kertesz-Farkas, H. Ali, B. Lucic, K. Lisek, L. Manganaro, S. Pongor, R. Luzzati, A. Recchia, F. Mavilio, M. Giacca and M. Lusic (2015). "Nuclear architecture dictates HIV-1 integration site selection." Nature **521**(7551): 227-231.

Marshall, H. M., K. Ronen, C. Berry, M. Llano, H. Sutherland, D. Saenz, W. Bickmore, E. Poeschla and F. D. Bushman (2007). "Role of PSIP1/LEDGF/p75 in lentiviral infectivity and integration targeting." PLoS One **2**(12): e1340.

Matsui, T., D. Leung, H. Miyashita, I. A. Maksakova, H. Miyachi, H. Kimura, M. Tachibana, M. C. Lorincz and Y. Shinkai (2010). "Proviral silencing in embryonic stem cells requires the histone methyltransferase ESET." Nature **464**(7290): 927-931.

Matysiak, J., P. Lesbats, E. Mauro, D. Lapaillerie, J. W. Dupuy, A. P. Lopez, M. S. Benleulmi, C. Calmels, M. L. Andreola, M. Ruff, M. Llano, O. Delelis, M. Lavigne and V. Parissi (2017). "Modulation of chromatin structure by the FACT histone chaperone complex regulates HIV-1 integration." Retrovirology **14**(1): 39.

Mbondji-Wonje, C., M. Dong, X. Wang, J. Zhao, V. Ragupathy, A. M. Sanchez, T. N. Denny and I. Hewlett (2018). "Distinctive variation in the U3R region of the 5' Long Terminal Repeat from diverse HIV-1 strains." PLoS One **13**(4): e0195661.

Meltzer, B., D. Dabbagh, J. Guo, F. Kashanchi, M. Tyagi and Y. Wu (2018). "Tat controls transcriptional persistence of unintegrated HIV genome in primary human macrophages." Virology **518**: 241-252.

Mertz, J. A., M. S. Simper, M. M. Lozano, S. M. Payne and J. P. Dudley (2005). "Mouse mammary tumor virus encodes a self-regulatory RNA export protein and is a complex retrovirus." J Virol **79**(23): 14737-14747.

Miklik, D., F. Senigl and J. Hejnar (2018). "Proviruses with Long-Term Stable Expression Accumulate in Transcriptionally Active Chromatin Close to the Gene Regulatory Elements: Comparison of ASLV-, HIV- and MLV-Derived Vectors." Viruses **10**(3).

Miller-Jensen, K., R. Skupsky, P. S. Shah, A. P. Arkin and D. V. Schaffer (2013). "Genetic selection for context-dependent stochastic phenotypes: Sp1 and TATA mutations increase phenotypic noise in HIV-1 gene expression." PLoS Comput Biol **9**(7): e1003135.

Mitchell, R. S., B. F. Beitzel, A. R. Schroder, P. Shinn, H. Chen, C. C. Berry, J. R. Ecker and F. D. Bushman (2004). "Retroviral DNA integration: ASLV, HIV, and MLV show distinct target site preferences." PLoS Biol **2**(8): E234.

Mizutani, T., A. Ishizaka, M. Tomizawa, T. Okazaki, N. Yamamichi, A. Kawana-Tachikawa, A. Iwamoto and H. Iba (2009). "Loss of the Brm-type SWI/SNF chromatin remodeling complex is a strong barrier to the Tat-independent transcriptional elongation of human immunodeficiency virus type 1 transcripts." J Virol **83**(22): 11569-11580.

Morabito, J. E., J. F. Trott, D. M. Korz, H. E. Fairfield, S. H. Buck and R. C. Hovey (2008). "A 5' distal palindrome within the mouse mammary tumor virus-long terminal repeat recruits a mammary gland-specific complex and is required for a synergistic response to progesterone plus prolactin." J Mol Endocrinol **41**(2): 75-90.

Mullers, E., K. Stirnnagel, S. Kaulfuss and D. Lindemann (2011). "Prototype foamy virus gag nuclear localization: a novel pathway among retroviruses." J Virol **85**(18): 9276-9285.

Munir, S., S. Thierry, F. Subra, E. Deprez and O. Delelis (2013). "Quantitative analysis of the time-course of viral DNA forms during the HIV-1 life cycle." Retrovirology **10**: 87-87.

Muñoz-Arias, I., G. Doitsh, Z. Yang, S. Sowinski, D. Ruelas and Warner C. Greene (2015). "Blood-Derived CD4 T Cells Naturally Resist Pyroptosis during Abortive HIV-1 Infection." Cell Host & Microbe **18**(4): 463-470.

Nam, J. S., J. E. Lee, K. H. Lee, Y. Yang, S. H. Kim, G. U. Bae, H. Noh and K. I. Lim (2019). "Shifting Retroviral Vector Integrations Away from Transcriptional Start Sites via DNA-Binding Protein Domain Insertion into Integrase." Mol Ther Methods Clin Dev **12**: 58-70.

Narezkina, A., K. D. Taganov, S. Litwin, R. Stoyanova, J. Hayashi, C. Seeger, A. M. Skalka and R. A. Katz (2004). "Genome-wide analyses of avian sarcoma virus integration sites." J Virol **78**(21): 11656-11663.

Naumann, S., D. Reutzel, M. Speicher and H. J. Decker (2001). "Complete karyotype characterization of the K562 cell line by combined application of G-banding, multiplex-fluorescence in situ hybridization, fluorescence in situ hybridization, and comparative genomic hybridization." Leuk Res **25**(4): 313-322.

Ne, E., R. J. Palstra and T. Mahmoudi (2018). "Transcription: Insights From the HIV-1 Promoter." Int Rev Cell Mol Biol **335**: 191-243.

Neri, F., S. Rapelli, A. Krepelova, D. Incarnato, C. Parlato, G. Basile, M. Maldotti, F. Anselmi and S. Oliviero (2017). "Intragenic DNA methylation prevents spurious transcription initiation." Nature **543**(7643): 72-77.

Niwa, O., Y. Yokota, H. Ishida and T. Sugahara (1983). "Independent mechanisms involved in suppression of the Moloney leukemia virus genome during differentiation of murine teratocarcinoma cells." Cell **32**(4): 1105-1113.

Nonnemacher, M. R., V. Pirrone, R. Feng, B. Moldover, S. Passic, B. Aiamkitsumrit, W. Dampier, A. Wojno, E. Kilareski, B. Blakey, T. S. Ku, S. Shah, N. T. Sullivan, J. M. Jacobson and B. Wigdahl (2016). "HIV-1 Promoter Single Nucleotide Polymorphisms Are Associated with Clinical Disease Severity." PLoS One **11**(4): e0150835.

Pagès, H., P. Aboyoun, R. Gentleman and S. DebRoy (2017). "Biostrings: Efficient manipulation of biological strings." R package version 2.46.0.

Pandey, K. K., S. Bera, K. Shi, H. Aihara and D. P. Grandgenett (2017). "A C-terminal "Tail" Region in the Rous Sarcoma Virus Integrase Provides High Plasticity of Functional Integrase Oligomerization during Intasome Assembly." J Biol Chem **292**(12): 5018-5030.

Pannell, D., C. S. Osborne, S. Yao, T. Sukonnik, P. Pasceri, A. Karaiskakis, M. Okano, E. Li, H. D. Lipshitz and J. Ellis (2000). "Retrovirus vector silencing is de novo methylase independent and marked by a repressive histone code." Embo j **19**(21): 5884-5894.

Papapetrou, E. P., G. Lee, N. Malani, M. Setty, I. Riviere, L. M. Tirunagari, K. Kadota, S. L. Roth, P. Giardina, A. Viale, C. Leslie, F. D. Bushman, L. Studer and M. Sadelain (2011). "Genomic safe harbors permit high beta-globin transgene expression in thalassemia induced pluripotent stem cells." Nat Biotechnol **29**(1): 73-78.

Passos, D. O., M. Li, R. Yang, S. V. Rebensburg, R. Ghirlando, Y. Jeon, N. Shkriabai, M. Kvaratskhelia, R. Craigie and D. Lyumkis (2017). "Cryo-EM structures and atomic model of the HIV-1 strand transfer complex intasome." Science **355**(6320): 89-92.

Patton, G. S., O. Erlwein and M. O. McClure (2004). "Cell-cycle dependence of foamy virus vectors." J Gen Virol **85**(Pt 10): 2925-2930.

Pelascini, L. P., J. M. Janssen and M. A. Goncalves (2013). "Histone deacetylase inhibition activates transgene expression from integration-defective lentiviral vectors in dividing and non-dividing cells." Hum Gene Ther **24**(1): 78-96.

Pion, M., A. Jordan, A. Biancotto, F. Dequiedt, F. Gondois-Rey, S. Rondeau, R. Vigne, J. Hejnar, E. Verdin and I. Hirsch (2003). "Transcriptional suppression of in vitro-integrated human immunodeficiency virus type 1 does not correlate with proviral DNA methylation." J Virol **77**(7): 4025-4032.

Plachy, J., J. Kotab, P. Divina, M. Reinisova, F. Senigl and J. Hejnar (2010). "Proviruses selected for high and stable expression of transduced genes accumulate in broadly transcribed genome areas." J Virol **84**(9): 4204-4211.

Prizan-Ravid, A., E. Elis, N. Laham-Karam, S. Selig, M. Ehrlich and E. Bacharach (2010). "The Gag cleavage product, p12, is a functional constituent of the murine leukemia virus pre-integration complex." PLoS Pathog **6**(11): e1001183.

Qu, D., C. Li, F. Sang, Q. Li, Z. Q. Jiang, L. R. Xu, H. J. Guo, C. Zhang and J. H. Wang (2016). "The variances of Sp1 and NF-kappaB elements correlate with the greater capacity of Chinese HIV-1 B'-LTR for driving gene expression." Sci Rep **6**: 34532.

Quercioli, V., C. Di Primio, A. Casini, L. C. F. Mulder, L. S. Vranckx, D. Borrenberghs, R. Gijsbers, Z. Debyser and A. Cereseto (2016). "Comparative Analysis of HIV-1 and Murine Leukemia Virus Three-Dimensional Nuclear Distributions." J Virol **90**(10): 5205-5209.

Rabbi, M. F., M. Saifuddin, D. S. Gu, M. F. Kagnoff and K. A. Roebuck (1997). "U5 region of the human immunodeficiency virus type 1 long terminal repeat contains TRE-like cAMP-responsive elements that bind both AP-1 and CREB/ATF proteins." Virology **233**(1): 235-245.

Rafati, H., M. Parra, S. Hakre, Y. Moshkin, E. Verdin and T. Mahmoudi (2011). "Repressive LTR nucleosome positioning by the BAF complex is required for HIV latency." PLoS Biol **9**(11): e1001206.

Rasheedi, S., M. C. Shun, E. Serrao, G. A. Sowd, J. Qian, C. Hao, T. Dasgupta, A. N. Engelman and J. Skowronski (2016). "The Cleavage and Polyadenylation Specificity Factor 6 (CPSF6) Subunit of the Capsid-recruited Pre-messenger RNA Cleavage Factor I (CFIm) Complex Mediates HIV-1 Integration into Genes." J Biol Chem **291**(22): 11809-11819.

Roberts, J. D., K. Bebenek and T. A. Kunkel (1988). "The accuracy of reverse transcriptase from HIV-1." Science **242**(4882): 1171-1173.

Roe, T., T. C. Reynolds, G. Yu and P. O. Brown (1993). "Integration of murine leukemia virus DNA depends on mitosis." Embo j **12**(5): 2099-2108.

Roguel, N., H. Moskowitz, H. Relevy, J. Hamburger and M. Kotler (1987). "The methylation state of the proviruses in avian sarcoma virus transformed chick and rat cells." Biochim Biophys Acta **910**(2): 116-122.

Rosen, C. A., J. G. Sodroski and W. A. Haseltine (1985). "The location of cis-acting regulatory sequences in the human T cell lymphotropic virus type III (HTLV-III/LAV) long terminal repeat." Cell **41**(3): 813-823.

Rosen, C. A., J. G. Sodroski, R. Kettman and W. A. Haseltine (1986). "Activation of enhancer sequences in type II human T-cell leukemia virus and bovine leukemia virus long terminal repeats by virus-associated trans-acting regulatory factors." J Virol **57**(3): 738-744.

Sadelain, M., E. P. Papapetrou and F. D. Bushman (2011). "Safe harbours for the integration of new DNA in the human genome." Nat Rev Cancer **12**(1): 51-58.

Santoni, F. A., O. Hartley and J. Luban (2010). "Deciphering the code for retroviral integration target site selection." PLoS Comput Biol **6**(11): e1001008.

Searle, S., D. A. Gillespie, D. J. Chiswell and J. A. Wyke (1984). "Analysis of the variations in proviral cytosine methylation that accompany transformation and morphological reversion in a line of Rous sarcoma virus-infected Rat-1 cells." Nucleic Acids Res **12**(13): 5193-5210.

Senigl, F., M. Auxt and J. Hejnar (2012). "Transcriptional provirus silencing as a crosstalk of de novo DNA methylation and epigenomic features at the integration site." Nucleic Acids Res **40**(12): 5298-5312.

Senigl, F., D. Miklik, M. Auxt and J. Hejnar (2017). "Accumulation of long-term transcriptionally active integrated retroviral vectors in active promoters and enhancers." Nucleic Acids Res **45**(22): 12752-12765.

Senigl, F., J. Plachy and J. Hejnar (2008). "The core element of a CpG island protects avian sarcoma and leukosis virus-derived vectors from transcriptional silencing." J Virol **82**(16): 7818-7827.

Shan, L., H. C. Yang, S. A. Rabi, H. C. Bravo, N. S. Shroff, R. A. Irizarry, H. Zhang, J. B. Margolick, J. D. Siliciano and R. F. Siliciano (2011). "Influence of host gene transcription level and orientation on HIV-1 latency in a primary-cell model." J Virol **85**(11): 5384-5393.

Sharma, A., R. C. Larue, M. R. Plumb, N. Malani, F. Male, A. Slaughter, J. J. Kessl, N. Shkriabai, E. Coward, S. S. Aiyer, P. L. Green, L. Wu, M. J. Roth, F. D. Bushman and M. Kvaratskhelia (2013). "BET proteins promote efficient murine leukemia virus integration at transcription start sites." Proc Natl Acad Sci U S A **110**(29): 12036-12041.

Shaw, A. and K. Cornetta (2014). "Design and Potential of Non-Integrating Lentiviral Vectors." Biomedicines **2**(1): 14-35.

Sheridan, P. L., T. P. Mayall, E. Verdin and K. A. Jones (1997). "Histone acetyltransferases regulate HIV-1 enhancer activity in vitro." Genes Dev **11**(24): 3327-3340.

Sherrill-Mix, S., M. K. Lewinski, M. Famiglietti, A. Bosque, N. Malani, K. E. Ocwieja, C. C. Berry, D. Looney, L. Shan, L. M. Agosto, M. J. Pace, R. F. Siliciano, U. O'Doherty, J. Guatelli, V. Planelles and F. D. Bushman (2013). "HIV latency and integration site placement in five cell-based models." Retrovirology **10**: 90.

Shoemaker, C., S. Goff, E. Gilboa, M. Paskind, S. W. Mitra and D. Baltimore (1980). "Structure of a cloned circular Moloney murine leukemia virus DNA molecule containing an inverted segment: implications for retrovirus integration." Proc Natl Acad Sci U S A **77**(7): 3932-3936.

Shun, M. C., N. K. Raghavendra, N. Vandegraaff, J. E. Daigle, S. Hughes, P. Kellam, P. Cherepanov and A. Engelman (2007). "LEDGF/p75 functions downstream from preintegration complex formation to effect gene-specific HIV-1 integration." Genes Dev **21**(14): 1767-1778.

Schneider, W. M., D. T. Wu, V. Amin, S. Aiyer and M. J. Roth (2012). "MuLV IN mutants responsive to HDAC inhibitors enhance transcription from unintegrated retroviral DNA." Virology **426**(2): 188-196.

Schrijvers, R., J. De Rijck, J. Demeulemeester, N. Adachi, S. Vets, K. Ronen, F. Christ, F. D. Bushman, Z. Debyser and R. Gijsbers (2012). "LEDGF/p75-independent HIV-1 replication demonstrates a role for HRP-2 and remains sensitive to inhibition by LEDGINs." PLoS Pathog **8**(3): e1002558.

Schroder, A. R., P. Shinn, H. Chen, C. Berry, J. R. Ecker and F. Bushman (2002). "HIV-1 integration in the human genome favors active genes and local hotspots." Cell **110**(4): 521-529.

Sieweke, M. H., H. Tekotte, U. Jarosch and T. Graf (1998). "Cooperative interaction of ets-1 with USF-1 required for HIV-1 enhancer activity in T cells." Embo j **17**(6): 1728-1739.

Singh, A., B. Razooky, C. D. Cox, M. L. Simpson and L. S. Weinberger (2010). "Transcriptional bursting from the HIV-1 promoter is a significant source of stochastic noise in HIV-1 gene expression." Biophys J **98**(8): L32-34.

Skupsky, R., J. C. Burnett, J. E. Foley, D. V. Schaffer and A. P. Arkin (2010). "HIV promoter integration site primarily modulates transcriptional burst size rather than frequency." PLoS Comput Biol **6**(9).

Smith, M. R. and W. C. Greene (1989). "The same 50-kDa cellular protein binds to the negative regulatory elements of the interleukin 2 receptor alpha-chain gene and the human immunodeficiency virus type 1 long terminal repeat." Proc Natl Acad Sci U S A **86**(21): 8526-8530.

Soboleva, T. A., M. Nekrasov, A. Pahwa, R. Williams, G. A. Huttley and D. J. Tremethick (2011). "A unique H2A histone variant occupies the transcriptional start site of active genes." Nat Struct Mol Biol **19**(1): 25-30.

Sodroski, J., C. Rosen, F. Wong-Staal, S. Z. Salahuddin, M. Popovic, S. Arya, R. C. Gallo and W. A. Haseltine (1985). "Trans-acting transcriptional regulation of human T-cell leukemia virus type III long terminal repeat." Science **227**(4683): 171-173.

Sowd, G. A., E. Serrao, H. Wang, W. Wang, H. J. Fadel, E. M. Poeschla and A. N. Engelman (2016). "A critical role for alternative polyadenylation factor CPSF6 in targeting HIV-1 integration to transcriptionally active chromatin." Proc Natl Acad Sci U S A **113**(8): E1054-1063.

Suerth, J. D., T. Maetzig, M. H. Brugman, N. Heinz, J. U. Appelt, K. B. Kaufmann, M. Schmidt, M. Grez, U. Modlich, C. Baum and A. Schambach (2012). "Alpharetroviral self-inactivating vectors: long-

term transgene expression in murine hematopoietic cells and low genotoxicity." Mol Ther **20**(5): 1022-1032.

Swamynathan, S. K., A. Nambiar and R. V. Guntaka (1997). "Chicken YB-2, a Y-box protein, is a potent activator of Rous sarcoma virus long terminal repeat-driven transcription in avian fibroblasts." J Virol **71**(4): 2873-2880.

Tagaya, Y. and R. C. Gallo (2017). "The Exceptional Oncogenicity of HTLV-1." Front Microbiol **8**: 1425.

Tekeste, S. S., T. A. Wilkinson, E. M. Weiner, X. Xu, J. T. Miller, S. F. J. Le Grice, R. T. Clubb and S. A. Chow (2015). "Interaction between Reverse Transcriptase and Integrase Is Required for Reverse Transcription during HIV-1 Replication." Journal of virology **89**(23): 12058-12069.

Thierry, S., S. Munir, E. Thierry, F. Subra, H. Leh, A. Zamborlini, D. Saenz, D. N. Levy, P. Lesbats, A. Saib, V. Parissi, E. Poeschla, E. Deprez and O. Delelis (2015). "Integrase inhibitor reversal dynamics indicate unintegrated HIV-1 dna initiate de novo integration." Retrovirology **12**: 24.

Tobaly-Tapiero, J., P. Bittoun, J. Lehmann-Che, O. Delelis, M. L. Giron, H. de The and A. Saib (2008). "Chromatin tethering of incoming foamy virus by the structural Gag protein." Traffic **9**(10): 1717-1727.

Toyoshima, H., M. Itoh, J. Inoue, M. Seiki, F. Takaku and M. Yoshida (1990). "Secondary structure of the human T-cell leukemia virus type 1 rex-responsive element is essential for rex regulation of RNA processing and transport of unspliced RNAs." J Virol **64**(6): 2825-2832.

Treand, C., I. du Chene, V. Bres, R. Kiernan, R. Benarous, M. Benkirane and S. Emiliani (2006). "Requirement for SWI/SNF chromatin-remodeling complex in Tat-mediated activation of the HIV-1 promoter." Embo j **25**(8): 1690-1699.

Trejbalova, K., D. Kovarova, J. Blazkova, L. Machala, D. Jilich, J. Weber, D. Kucerova, O. Vencalek, I. Hirsch and J. Hejnar (2016). "Development of 5' LTR DNA methylation of latent HIV-1 provirus in cell line models and in long-term-infected individuals." Clin Epigenetics **8**: 19.

Trinite, B., E. C. Ohlson, I. Voznesensky, S. P. Rana, C. N. Chan, S. Mahajan, J. Alster, S. A. Burke, D. Wodarz and D. N. Levy (2013). "An HIV-1 replication pathway utilizing reverse transcription products that fail to integrate." J Virol **87**(23): 12701-12720.

Trobridge, G. and D. W. Russell (2004). "Cell cycle requirements for transduction by foamy virus vectors compared to those of oncovirus and lentivirus vectors." J Virol **78**(5): 2327-2335.

Trobridge, G. D., D. G. Miller, M. A. Jacobs, J. M. Allen, H. P. Kiem, R. Kaul and D. W. Russell (2006). "Foamy virus vector integration sites in normal human cells." Proc Natl Acad Sci U S A **103**(5): 1498-1503.

Uren, A. G., H. Mikkers, J. Kool, L. van der Weyden, A. H. Lund, C. H. Wilson, R. Rance, J. Jonkers, M. van Lohuizen, A. Berns and D. J. Adams (2009). "A high-throughput splinkerette-PCR method for the isolation and sequencing of retroviral insertion sites." Nat Protoc **4**(5): 789-798.

Van Driessche, B., A. Rodari, N. Delacourt, S. Fauquenoy, C. Vanhulle, A. Burny, O. Rohr and C. Van Lint (2016). "Characterization of new RNA polymerase III and RNA polymerase II transcriptional promoters in the Bovine Leukemia Virus genome." Sci Rep **6**: 31125.

Van Lint, C., S. Emiliani, M. Ott and E. Verdin (1996). "Transcriptional activation and chromatin remodeling of the HIV-1 promoter in response to histone acetylation." Embo j **15**(5): 1112-1120.

van Nuland, R., F. M. A. van Schaik, M. Simonis, S. van Heesch, E. Cuppen, R. Boelens, H. T. M. Timmers and H. van Ingen (2013). "Nucleosomal DNA binding drives the recognition of H3K36-methylated nucleosomes by the PSIP1-PWWP domain." Epigenetics & Chromatin **6**(1): 12.

Vanegas, M., M. Llano, S. Delgado, D. Thompson, M. Peretz and E. Poeschla (2005). "Identification of the LEDGF/p75 HIV-1 integrase-interaction domain and NLS reveals NLS-independent chromatin tethering." J Cell Sci **118**(Pt 8): 1733-1743.

Varmus, H. E., N. Quintrell, E. Medeiros, J. M. Bishop, R. C. Nowinski and N. H. Sarkar (1973). "Transcription of mouse mammary tumor virus genes in tissues from high and low tumor incidence mouse strains." J Mol Biol **79**(4): 663-679.

Vatakis, D. N., S. Kim, N. Kim, S. A. Chow and J. A. Zack (2009). "Human immunodeficiency virus integration efficiency and site selection in quiescent CD4+ T cells." J Virol **83**(12): 6222-6233.

Vemula, S. V., R. Veerasamy, V. Ragupathy, S. Biswas, K. Devadas and I. Hewlett (2015). "HIV-1 induced nuclear factor I-B (NF-IB) expression negatively regulates HIV-1 replication through interaction with the long terminal repeat region." Viruses **7**(2): 543-558.

Venter, J. C., M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, J. D. Gocayne, P. Amanatides, R. M. Ballew, D. H. Huson, J. R. Wortman, Q. Zhang, C. D. Kodira, X. H. Zheng, L. Chen, M. Skupski, G. Subramanian, P. D. Thomas, J. Zhang, G. L. Gabor Miklos, C. Nelson, S. Broder, A. G. Clark, J. Nadeau, V. A. McKusick, N. Zinder, A. J. Levine, R. J. Roberts, M. Simon, C. Slayman, M. Hunkapiller, R. Bolanos, A. Delcher, I. Dew, D. Fasulo, M. Flanigan, L. Florea, A. Halpern, S. Hannenhalli, S. Kravitz, S. Levy, C. Mobarry, K. Reinert, K. Remington, J. Abu-Threideh, E. Beasley, K. Biddick, V. Bonazzi, R. Brandon, M. Cargill, I. Chandramouliswaran, R. Charlab, K. Chaturvedi, Z. Deng, V. D. Francesco, P. Dunn, K. Eilbeck, C. Evangelista, A. E. Gabrielian, W. Gan, W. Ge, F. Gong, Z. Gu, P. Guan, T. J. Heiman, M. E. Higgins, R.-R. Ji, Z. Ke, K. A. Ketchum, Z. Lai, Y. Lei, Z. Li, J. Li, Y. Liang, X. Lin, F. Lu, G. V. Merkulov, N. Milshina, H. M. Moore, A. K. Naik, V. A. Narayan, B. Neelam, D. Nusskern, D. B. Rusch, S. Salzberg, W. Shao, B. Shue, J. Sun, Z. Y. Wang, A. Wang, X. Wang, J. Wang, M.-H. Wei, R. Wides, C. Xiao, C. Yan, A. Yao, J. Ye, M. Zhan, W. Zhang, H. Zhang, Q. Zhao, L. Zheng, F. Zhong, W. Zhong, S. C. Zhu, S. Zhao, D. Gilbert, S. Baumhueter, G. Spier, C. Carter, A. Cravchik, T. Woodage, F. Ali, H. An, A. Awe, D. Baldwin, H. Baden, M. Barnstead, I. Barrow, K. Beeson, D. Busam, A. Carver, A. Center, M. L. Cheng, L. Curry, S. Danaher, L. Davenport, R. Desilets, S. Dietz, K. Dodson, L. Doup, S. Ferriera, N. Garg, A. Gluecksmann, B. Hart, J. Haynes, C. Haynes, C. Heiner, S. Hladun, D. Hostin, J. Houck, T. Howland, C. Ibegwam, J. Johnson, F. Kalush, L. Kline, S. Koduru, A. Love, F. Mann, D. May, S. McCawley, T. McIntosh, I. McMullen, M. Moy, L. Moy, B. Murphy, K. Nelson, C. Pfannkoch, E. Pratts, V. Puri, H. Qureshi, M. Reardon, R. Rodriguez, Y.-H. Rogers, D. Romblad, B. Ruhfel, R. Scott, C. Sitter, M. Smallwood, E. Stewart, R. Strong, E. Suh, R. Thomas, N. N. Tint, S. Tse, C. Vech, G. Wang, J. Wetter, S. Williams, M. Williams, S. Windsor, E. Winn-Deen, K. Wolfe, J. Zaveri, K. Zaveri, J. F. Abril, R. Guigó, M. J. Campbell, K. V. Sjolander, B. Karlak, A. Kejariwal, H. Mi, B. Lazareva, T.

Hatton, A. Narechania, K. Diemer, A. Muruganujan, N. Guo, S. Sato, V. Bafna, S. Istrail, R. Lippert, R. Schwartz, B. Walenz, S. Yooseph, D. Allen, A. Basu, J. Baxendale, L. Blick, M. Caminha, J. Carnes-Stine, P. Caulk, Y.-H. Chiang, M. Coyne, C. Dahlke, A. D. Mays, M. Dombroski, M. Donnelly, D. Ely, S. Esparham, C. Fosler, H. Gire, S. Glanowski, K. Glasser, A. Glodek, M. Gorokhov, K. Graham, B. Gropman, M. Harris, J. Heil, S. Henderson, J. Hoover, D. Jennings, C. Jordan, J. Jordan, J. Kasha, L. Kagan, C. Kraft, A. Levitsky, M. Lewis, X. Liu, J. Lopez, D. Ma, W. Majoros, J. McDaniel, S. Murphy, M. Newman, T. Nguyen, N. Nguyen, M. Nodell, S. Pan, J. Peck, M. Peterson, W. Rowe, R. Sanders, J. Scott, M. Simpson, T. Smith, A. Sprague, T. Stockwell, R. Turner, E. Venter, M. Wang, M. Wen, D. Wu, M. Wu, A. Xia, A. Zandieh and X. Zhu (2001). "The Sequence of the Human Genome." Science **291**(5507): 1304-1351.

Verdin, E., P. Paras, Jr. and C. Van Lint (1993). "Chromatin disruption in the promoter of human immunodeficiency virus type 1 during transcriptional activation." Embo j **12**(8): 3249-3259.

von der Ahe, D., S. Janich, C. Scheidereit, R. Renkawitz, G. Schutz and M. Beato (1985). "Glucocorticoid and progesterone receptors bind to the same sites in two hormonally regulated promoters." Nature **313**(6004): 706-709.

von Schwedler, U., R. S. Kornbluth and D. Trono (1994). "The nuclear localization signal of the matrix protein of human immunodeficiency virus type 1 allows the establishment of infection in macrophages and quiescent T lymphocytes." Proc Natl Acad Sci U S A **91**(15): 6992-6996.

Vranckx, L. S., J. Demeulemeester, S. Saleh, A. Boll, G. Vansant, R. Schrijvers, C. Weydert, E. Battivelli, E. Verdin, A. Cereseto, F. Christ, R. Gijsbers and Z. Debyser (2016). "LEDGIN-mediated Inhibition of Integrase-LEDGF/p75 Interaction Reduces Reactivation of Residual Latent HIV." EBioMedicine **8**: 248-264.

Wahlers, A., P. F. Zipfel, M. Schwieger, W. Ostertag and C. Baum (2002). "In vivo analysis of retroviral enhancer mutations in hematopoietic cells: SP1/EGR1 and ETS/GATA motifs contribute to long terminal repeat specificity." J Virol **76**(1): 303-312.

Wanaguru, M., D. J. Barry, D. J. Benton, N. J. O'Reilly and K. N. Bishop (2018). "Murine leukemia virus p12 tethers the capsid-containing pre-integration complex to chromatin by binding directly to host nucleosomes in mitosis." PLoS Pathog **14**(6): e1007117.

Wang, G. P., A. Ciuffi, J. Leipzig, C. C. Berry and F. D. Bushman (2007). "HIV integration site selection: analysis by massively parallel pyrosequencing reveals association with epigenetic modifications." Genome Res **17**(8): 1186-1194.

Wang, G. Z., Y. Wang and S. P. Goff (2016). "Histones Are Rapidly Loaded onto Unintegrated Retroviral DNAs Soon after Nuclear Entry." Cell Host Microbe **20**(6): 798-809.

Wang, G. Z., D. Wolf and S. P. Goff (2014). "EBP1, a novel host factor involved in primer binding site-dependent restriction of moloney murine leukemia virus in embryonic cells." J Virol **88**(3): 1825-1829.

Wang, Z., C. Zang, J. A. Rosenfeld, D. E. Schones, A. Barski, S. Cuddapah, K. Cui, T. Y. Roh, W. Peng, M. Q. Zhang and K. Zhao (2008). "Combinatorial patterns of histone acetylations and methylations in the human genome." Nat Genet **40**(7): 897-903.

Weinberger, L. S., J. C. Burnett, J. E. Toettcher, A. P. Arkin and D. V. Schaffer (2005). "Stochastic gene expression in a lentiviral positive-feedback loop: HIV-1 Tat fluctuations drive phenotypic diversity." Cell **122**(2): 169-182.

Wight, D. J., V. C. Boucherit, M. Nader, D. J. Allen, I. A. Taylor and K. N. Bishop (2012). "The gammaretroviral p12 protein has multiple domains that function during the early stages of replication." Retrovirology **9**: 83.

Wight, D. J., V. C. Boucherit, M. Wanaguru, E. Elis, E. M. Hirst, W. Li, M. Ehrlich, E. Bacharach and K. N. Bishop (2014). "The N-terminus of murine leukaemia virus p12 protein is required for mature core stability." PLoS Pathog **10**(10): e1004474.

Williams, S. A., L. F. Chen, H. Kwon, C. M. Ruiz-Jarabo, E. Verdin and W. C. Greene (2006). "NF-kappaB p50 promotes HIV latency through HDAC recruitment and repression of transcriptional initiation." Embo j **25**(1): 139-149.

Winans, S., R. C. Larue, C. M. Abraham, N. Shkriabai, A. Skopp, D. Winkler, M. Kvaratskhelia and K. L. Beemon (2017). "The FACT Complex Promotes Avian Leukosis Virus DNA Integration." J Virol **91**(7).

Wolf, D., F. Cammas, R. Losson and S. P. Goff (2008). "Primer binding site-dependent restriction of murine leukemia virus requires HP1 binding by TRIM28." J Virol **82**(9): 4675-4679.

Wolf, D. and S. P. Goff (2007). "TRIM28 mediates primer binding site-targeted silencing of murine leukemia virus in embryonic cells." Cell **131**(1): 46-57.

Wolf, D. and S. P. Goff (2009). "Embryonic stem cells use ZFP809 to silence retroviral DNAs." Nature **458**(7242): 1201-1204.

Wu, X., Y. Li, B. Crise and S. M. Burgess (2003). "Transcription start regions in the human genome are favored targets for MLV integration." Science **300**(5626): 1749-1751.

Wu, X., H. Liu, H. Xiao, J. A. Conway, E. Hehl, G. V. Kalpana, V. Prasad and J. C. Kappes (1999). "Human immunodeficiency virus type 1 integrase protein promotes reverse transcription through specific interactions with the nucleoprotein reverse transcription complex." J Virol **73**(3): 2126-2135.

Wu, Y. and J. W. Marsh (2001). "Selective transcription and modulation of resting T cell activity by preintegrated HIV DNA." Science **293**(5534): 1503-1506.

Wyke, J. A. and K. Quade (1980). "Infection of rat cells by avian sarcoma virus: factors affecting transformation and subsequent reversion." Virology **106**(2): 217-233.

Yamamoto, T., B. de Crombrugghe and I. Pastan (1980). "Identification of a functional promoter in the long terminal repeat of Rous sarcoma virus." Cell **22**(3): 787-797.

Yamashita, M. and M. Emerman (2004). "Capsid is a dominant determinant of retrovirus infectivity in nondividing cells." J Virol **78**(11): 5670-5678.

Yamashita, M., O. Perez, T. J. Hope and M. Emerman (2007). "Evidence for direct involvement of the capsid protein in HIV infection of nondividing cells." PLoS Pathog **3**(10): 1502-1510.

Yasukawa, K., K. Iida, H. Okano, R. Hidese, M. Baba, I. Yanagihara, K. Kojima, T. Takita and S. Fujiwara (2017). "Next-generation sequencing-based analysis of reverse transcriptase fidelity." Biochem Biophys Res Commun **492**(2): 147-153.

Yin, Z., K. Shi, S. Banerjee, K. K. Pandey, S. Bera, D. P. Grandgenett and H. Aihara (2016). "Crystal structure of the Rous sarcoma virus intasome." Nature **530**(7590): 362-366.

Zachow, K. R. and K. F. Conklin (1992). "CArG, CCAAT, and CCAAT-like protein binding sites in avian retrovirus long terminal repeat enhancers." J Virol **66**(4): 1959-1970.

Zennou, V., C. Petit, D. Guetard, U. Nerhbass, L. Montagnier and P. Charneau (2000). "HIV-1 genome nuclear import is mediated by a central DNA flap." Cell **101**(2): 173-185.

Zhang, D. W., H. Q. He and S. X. Guo (2014). "Hairpin DNA probe-based fluorescence assay for detecting palindrome cleavage activity of HIV-1 integrase." Anal Biochem **460**: 36-38.

Zhang, Z., E. Kim and D. Martineau (1999). "Functional characterization of a piscine retroviral promoter." J Gen Virol **80 (Pt 12)**: 3065-3072.

Zheng, R., T. M. Jenkins and R. Craigie (1996). "Zinc folds the N-terminal domain of HIV-1 integrase, promotes multimerization, and enhances catalytic activity." Proc Natl Acad Sci U S A **93**(24): 13659-13664.

Zhu, K., C. Dobard and S. A. Chow (2004). "Requirement for integrase during reverse transcription of human immunodeficiency virus type 1 and the effect of cysteine mutations of integrase on its interactions with reverse transcriptase." J Virol **78**(10): 5045-5055.

Zhu, Y., G. Z. Wang, O. Cingoz and S. P. Goff (2018). "NP220 mediates silencing of unintegrated retroviral DNA." Nature **564**(7735): 278-282.

Zhyvoloup, A., A. Melamed, I. Anderson, D. Planas, C. H. Lee, J. Kriston-Vizi, R. Ketteler, A. Merritt, J. P. Routy, P. Ancuta, C. R. M. Bangham and A. Fassati (2017). "Digoxin reveals a functional connection between HIV-1 integration preference and T-cell activation." PLoS Pathog **13**(7): e1006460.