A novel lentiviral vector model to investigate the mechanism of insertional mutagenesis by aberrant splicing

Kanayo Doi

Thesis submitted to University College London for the degree of Doctor of Philosophy

2017

Division of Infection and Immunity
University College London

Declaration

I, Kanayo Doi, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

Retrovirus vector-mediated insertional mutagenesis (IM) can lead to serious cancerous risks in gene therapy. Although multiple cases of IM-induced malignancies have been reported in clinical trials with gammaretroviral vectors, no evidence for malignancies has been reported with HIV-derived lentiviral vectors (LVs). However, the 2010 clinical trial for ß-thalassaemia had reported the clonal dominance of erythroblasts in one patient. This was caused by the LV integration into the *HMGA2* gene locus and the up-regulation of HMGA2 protein by the formation of host-vector chimeric transcripts (Cavazzana-Calvo et al., 2010). The observed pattern of splicing in the *HMGA2* locus, splice-in, is splicing from a host splice donor to a vector splice acceptor. Considering that this literature is the first and only evidence that implies mutagenicity of LVs, We thought that formation of splice-in fusion transcripts could be a key factor to contribute potential LV-mediated IM. Therefore, we set our aim of investigating the mechanism of splice-in caused by vector integration by designing a novel LV model that can induce splice-in fusion transcripts, and by applying to an *in vitro* IM assay previously established by our group (Bokhoven et al., 2009). The IM assay utilise IL-3 dependent Bcl15 cell transformation into IL-3 independent, which can be analysed for the transformation mechanism via LV integration. Phenotypic analysis of the isolated mutant clone (the IL-3I clone) suggested secretion of autocrine factors and the cause of which was interrogated by molecular analysis. Additionally, the IL-3I clone and two bulk populations from different steps of an IM assay were subjected to RNA sequencing (RNA-Seq) to compare the isolated host genes and investigate chimeric mRNAs. Importantly, RNA-Seq identified the *Angpt1* fusion transcript but in the splice-out form using a cryptic vector SD as a possible cause for cell transformation of the IL-3I clone. This gene locus has Retroviral Tagged Cancer Gene Database (RTCGD) hits and was also detected by LM-PCR. A variety of splice sites were used in the detected possible fusion mRNAs. To confirm, *Angpt1* loci is the major cause of IL-3 independence and the pattern of splice sites use depend integration site; further analysis is required and discussed in the final chapter.

Acknowledgement

Firstly, I would like to sincerely thank Prof. Mary Collins and Dr. Yasuhiro Takeuchi for their supervision throughout my 4-year PhD. I was truly lucky to have Mary and Yasu as my PhD supervisors, which gave me the chance to explore this highly advanced and promising study field. Yasu offered me warm support and advice for the success of my PhD project since I first joined the lab. Mary suggested me critical advice in my experimental design and result time to time, showing me the way to look one thing from different angles. I could not have come this far without their countless encouragement and support.

Thanks to the environment of Mary's and Yasu's lab, I could also expand my science network with other experienced scientists at NIBSC. Dr. Edward T. Mee, Dr. Mark Preston and Dr. Giada Mattiuzzo helped me kindly with a lot of rational and practical advice.

During my PhD life at the Wohl Virion Centre, I was grateful to Khaled Samber, Ilaria Nisoli, Hagen Schwenzer and James Heather for the kind help to develop myself as a scientist. I also would like to thank Prof. Robin Weiss, Prof. Ariberto Fassati and Dr. Clare Jolly for the practical feedback and excellent advice during our joint lab meetings.

To my dear friends and colleagues at the Wohl Virion Centre including Sean Knight, Juan Ribes, Edward Tsao, Alice Len, Aksana Labokha, Sarah Watters, Kasia Karwacz, Shimona Starling, Maitrayi Shivkumar, Mazlina Ismail and Mattia Cinelli, I was so pleased that I could share our time together inside and outside the lab, which was all very precious and priceless memory.

I cannot thank enough to my parents who I love and respect the most. They have been always beside me in time of need. I sincerely appreciate their devoted belief in me and their constant encouragement at any time. My beloved aunt was also supportive and always a good listener to me.

Lastly, I would like to thank all professors for their support and encouragement during this PhD journey including Prof. Masanobu Satake, Prof. Kouki Hikosaka and Prof. Toshiyuki Takai. I was also grateful that Japan Student Services Organization that allowed me to seize this precious opportunity by their financial support through my PhD programme.

Table of Contents

List of Tables

List of figures

List of abbreviations

A2UCOE	HNRPA2B1-CBX3 Ubiquitous Chromatin Opening Element
AADL	Aminoacid Decarboxylase
AAVs	Adeno-Associated Vectors
Ad2	Adenovirus Type 2
ADA	Adenosine Deaminase
AIDS	Acquired Immune Deficiency Syndrome
ALD	Adrenoleukodystrophy
ALIX	ALG2-Interacting Protein X
AML	Acute Myeloid Leukemia
AONs	Antisense Oligonucleotides
APC	Adenomatous Polyposis Coli
APOBEC3G	Apolipoprotein B MRNA Editing Enzyme Catalytic Subunit 3G
ARSA	Arylsulfatase A
ASLV	Avian Sarcoma-Leukosis Virus
AVs	Adenoviral Vectors
b	Base
B-ALL	B-Lymphoblastic Leukemia
BET	Bromodomain And Extra Terminal
BLT Mouse Model	Humanized Bone Marrow, Liver, Thymus Mouse Model
ß-TM	ß-Thalassaemia
bp	Base Pair
CA	Capsid
CARs	Chimeric Antigen Receptors
cART	Combinational Antiviral Therapy
CAT	Chloramphenicol Actyltransferase
CCD	Core Catalytic Domain
CCR5	Cysteine-Cysteine Chemokine Receptor 5
CDK9	Cyclin-Dependent Kinase 9

cDNA	Complementary DNA
Ch	Chromosome
CH1	Cyclohydrolase 1
cHS4	Chicken Hypersensitive Site 4
CIAP	Calf Intestinal Alkaline Phosphate
CIS	Common Integration Sites
CMV	Cytomegalovirus
CNS	Central Nerve System
cPPT	Central Polypurine Tract
CPSF6	Cleavage And Polyadenylation Specific Factor 6
Crm1	Chromosome Region Maintenance 1
CTD	C-Terminal Domain
CXCR4	Chemokine C-X-C motif Receptor 4
CycT1	Cyclin T1 Kinase
CypA	Cyclophilin A
Da	Dalton
DMD	Duchenne Muscular Dystrophy
EDTA	Ethylenediaminetetraacetic Acid
EIAV	Equine Infectious Anemia Virus
EM	Electron Microscopic
Em	Emerald
ES cells	Embryonal Stem Cells
ESCRT	Endosomal Sorting Complexes Required For Transport
ESEs	Exonic Splicing Enhancers
ESS	Exonic Splice Silencer
FACS	Fluorescence Activated Cell Sorting
FBS	Fetal Bovine Serum
g	Gram
gDNA	Genomic DNA
Ghr	Growth Hormone Receptor
GP	Glycosylated Protein
GRVs	Gammaretroviral Vectors

GTP	Guanosine-5'-Triphosphate
GVHD	Graft-Versus Host Diseases
HCC	Hepatocellular Carcinomas
HIV	Human Immunodeficiency Virus
HLA	Human Leukocyte Antigen
HMGA2	High Mobility Group AT-Hook 2
hnRNP	Heterogeneous Nuclear RNP
HR-C	Carboxy-teminal Helical Region
HR-N	Amino-terminal Helical Region
HS	Hypersensitive Sites
HSCs	Hematopoietic Stem Cells
HSCT	Hematopoietic Stem Cells Transplantation
HTLV-1	T-cell Leukemia Virus Type 1
IAP	Iintracistarnal A Particle
IBD	Integrase-Binding Domain
ICTV	International Committee on Taxonomy of Viruses
IL-2	Interleukin-2
IM	Insertional Mutagenesis
IN	Integrase
iPS cells	Induced Pluripotent Stem Cells
IRES	Internal Ribosomal Entry Site
ISE	Intronic Splice Enhancer
ISS	Intronic Splice Silencer
k	Kilo
LAM-PCR	Linear Amplification–Mediated PCR
LB	Luria-Bertani
LCR	Locus Control Region
LDL	Low-Density Lipoprotein
LEDGF	Lens Epithelium-Derived Growth Factor
LI	Late Infantile
Lin-	Lineage-Negative
LM-PCR	Ligation-Mediated PCR
LTR	Long Terminal Repeat

LV	Lentiviral Vector
M	Molar
MA	Matrix
Mb	Mega Base
MECOM	MDS1 And EVI1 Complex Locus Protein EVI1
MFI	Mean Fluorescence Intensity
MHC II	Major Histocompatibility Complex class II
ml	Milliliter
MLD	Metachromatic Leukodystrophy
MLV	Murine Leukemia Virus
M.M	Materials and Methods
MMTV	Mouse Mammary Tumor Virus
MOI	Multiplicity Of Infection
MoMLV	Molony Murine Leukemia Virus
MSD	Major Splice Donor
MVB	Multivesicular Body
MZ	Marginal Zone
n	Nano
NADPH	Nicotinamide Adenine Dinucleotide Phosphate
NC	Nucleocapsid
NGS	Next Generation Sequencing
NIH	National Institute Of Health
NK	Natural Killer
NLSs	Nuclear Localization Signals
nm	Nanometer
NMD	Nonsense Mediated Decay
NPC	Nuclear Pore Complex
nrLAM-PCR	Non-Restrictive Linear Amplification-Mediated PCR
nt	Nucleotide
NTD	N-Terminal Domain
oC	Degree Celsius
ORF	Open Reading Frames
p	Pico

P-TEF-b	Positive Transcription Elongation Factor b
PBS	Primer Binding Site
PCLs	Packaging Cell Lines
PCR	Polymerase Chain Reaction
PD	Parkinson's Disease
PFA	Paraformaldehyde
PGK	Phosphoglycerate Kinase
PIC	Pre-Integration Complex
PID	Primary Immunodeficiencies
pim-1	Proto-Oncogene Serine/Threonine-Protein Kinase Pim-1
PMOs	Phosphorodiamidate Morpholino Oligomers
PNS	Peripheral Nervous Systems
PPT	Polypurine Tract
PR	Protease
Psi, ψ	Packaging Signal
PtdIns(4,5)P2	Phosphatidylinositol 4,5-Bisphosphate
puroR	Puromycin Resistant
qPCR	Quantitative PCR
RACE	Rapid Amplification Of cDNA Ends
RBD	RNA-Binding Domain
RCVs	Replication Competent Viruses
RER	Rough Endoplasmic Reticulum
REV-A	Reticuloendotheliosis Virus Strain A
RRE	Rev-Response Element
RSV	Rous Sarcoma Virus
RT	Reverse Transcriptase
RTC	Reverse Transcription Complex
RTCGD	Retroviral Tagged Cancer Gene Database
RVs	Retroviral Vectors
s	Second
SA	Splice Acceptor
SAMHD1	Sam Domain, And HD Domain-Containing Protein 1

SD	Splice Donor
SEM	Standard Error Of Mean
SERINC3	Serine Incorporator 3
SERINC5	Serine Incorporator 5
SFFV	Spleen Focus-Forming Virus
SIN	Self-Inactivating
SIV	Simian Immunodeficiency Virus
SMA	Spinal Muscular Atrophy
SMN1	Survival Motor Neuron 1
snRNPs	Small Nuclear Ribonucleoproteins
SR	Serine/Arginine
SU	Surface Glycoprotein
TAE Buffer	Tris-Base, Acetic Acid And EDTA Buffer
TAR	Transactivation Response
Tat	Transactivator Protein
TCR	T Cell Receptor
TH	Tyrosine Hydroxylase
tk	Thymidine Kinase gene
Tm	Melting Temperature
TM	Transmembrane Glycoprotein
TNBC	Triple Negative Breast Cancer
TNPO3	Transportin 3
TRIM5a	Tripartile Motif Protein 5 alpha
tRNA	Transfer RNA
TSG101	Tumour Susceptibility Gene 101
u	Unit
U2AF	U2 Auxiliary Factor
UBIQ	Ubiquitin Promoter
UTR	Untranslated Region
VLCFAs	Very Long Chain Fatty Acids
VPS4	Atpase Vacuolar Protein Sorting 4
VSV-G	Vesicular Stomatitis Virus
WAS	Wiskott-Aldrich Syndrome

WPRE	Woodchuck Hepatitis Virus Posttrascriptional Regulatory Elements
WT	Wild-Type
X-CGD	X-Linked Chronic Granulomatous Disease
X-SCID	X-Linked Severe Combined Immunodeficiency
ZFNs	Zinc Finger Nucleases
μ	Micro
%	Percentage
6HB	Six-Helical Bundle

1 Introduction

1.1 Retrovirus biology

1.1.1 Virus discovery

The first retrovirus to be discovered was the Rous sarcoma virus, which was isolated from tumours in chickens that were being processed by the poultry meat industry (Rous, 1910). Retroviruses were then isolated from a range of wild and domesticated animals (Shope and Hurst, 1933) (Bittner, 1942) (Sweet and Hilleman, 1960) (Weiss and Vogt, 2011).

The first human retrovirus was not isolated until 1980 when human T-cell leukaemia virus type 1 (HTLV-1) was found in T cells cultured from leukaemia patients (Poiesz et al., 1980). This development of methods for cultivating T cells, following the discovery of T cell growth factor interleukin-2 (IL-2), also allowed the discovery of human immunodeficiency virus (HIV) in 1984 (Gallo, 2005) (Barre-Sinoussi et al., 1983). HIV is widely known as the cause of acquired immune deficiency syndrome (AIDS), which is characterised by the occurrence of opportunistic infection due to immune suppression. In order to seek strategies to treat AIDS, HIV has been extensively studied and characterised. However, in 2015 36.7 million people were still suffering from AIDS worldwide (http://www.who.int/hiv/data/en/). HIV has two subtypes, 1 and 2. In this thesis, “HIV” denotes HIV-1 because this is the major subtype causing the pandemic and because the lentiviral vectors I studied are based on HIV-1.

After its isolation in 1984, infectious molecular clones of HIV were isolated and the genomic sequence was published in 1985 (Wain-Hobson et al., 1985). Then the molecular clones were engineered to produce replication-defective HIV vectors which could be used for gene delivery to eukaryotic cells (Zufferey et al., 1997). Nowadays, lentiviral vectors based on HIV are widely used in laboratories. Their application is extended in clinical gene therapy; lentiviral vectors can be used in various therapeutic strategies, such as a functional gene or cDNA

delivery, shRNA delivery, and gene-editing. Clinical trials using lentiviral vectors are reported in inheritable disorders such as hematological, storage and neurodegenerative diseases as well as cancer (Naldini, 2015). The details are described in the 1.4 Gene therapy section.

This thesis focuses on the characterisation of a model HIV vector and identifies vector-host fusion transcripts to analyse their potential effect on vector genome expression and transduced cell gene expression. Because HIV vectors integrate into the target cell genome they have the potential to cause insertional mutagenesis (IM), leading to the transformation of the transduced cells. My study requires an understanding of basic retrovirus biology such as genome structure and life cycle. Therefore, the introduction starts with retrovirus biology, followed by vectorology and vector application in clinical trials. In addition, as the splicing pattern in host-vector fusion transcripts is of interest, the final section of this introduction describes the mechanism of splicing and explains splicing-mediated genetic disorders.

1.1.2 Retrovirus taxonomy

Virus taxonomy is defined by a consensus group of properties, agreed by the International Committee on Taxonomy of Viruses (ICTV) and the virus classification is publicly accessible online (http://www.ictvonline.org/).

Retroviridae is one of 82 families of viruses listed and is subdivided into two sub-families, Orthoretrovirinae and Spumaretrovirinae. Orthoretrovirinae has six genera and Spumaretrovirinae has one genus. HIV belongs to Lentiviruses in Orthoretrovirinae (Table 1-1).

Subfamily	Genus	Species
Orthoretrovirinae	*Alpharetrovirus*	*e.g. Rous sarcoma virus*
	Betaretrovirus	*e.g. Mouse mammary tumor virus*
	Deltaretrovirus	*e.g. Bovine leukaemia virus*
	Epsilonretrovirus	*e.g. Walleye dermal sarcoma virus*
	Gammaretrovirus	*e.g. Murine leukaemia virus*
	Lentivirus	*Bovine immunodeficiency virus*
		Caprine arthritis encephalitis virus
		Equine infectious anemia virus
		Feline immunodeficiency virus
		Human immunodeficiency virus 1
		Human immunodeficiency virus 2
		Puma lentivirus
		Simian immunodeficiency virus
		Visna/maedi virus
Spumaretrovirinae	*Spumavirus*	*e.g. Simian foamy virus*

Table 1-1 Retrovirus taxonomy

The virus taxonomy of retroviruses was extracted from http://www. ictvonline.org/. Human immunodeficiency virus 1 is highlighted in bold with an underline.

1.1.3 Retrovirus genome structure

The RNA genome of all retroviruses encodes three open reading frames (ORF) corresponding to the major virus proteins, Gag, Pol and Env; these are flanked by parts of the long terminal repeat (LTR) at both 5' and 3' ends. The complexity of the genome structure differs between retroviruses. For instance, Murine leukaemia virus (MLV) has the simplest genome organisation with only the three major structural genes, *gag, pol,* and *env*, which encode polyproteins cleaved during viral assembly (Cullen, 1992) (Fig.1-1, MLV). The *gag* gene encodes viral core proteins: matrix (MA), p12, capsid (CA), and nucleocapsid (NC). The *pol* gene encodes enzymes necessary for viral replication: protease (PR), reverse

transcriptase (RT), and integrase (IN). The *env* gene encodes the viral transmembrane (TM) and surface (SU) glycoproteins. HIV has a more complicated genomic structure than MLV, with additional six genes encoding regulatory proteins (Tat and Rev) and accessory proteins (Vif, Vpr, Vpu and Nef). In addition, the HIV gag precursor polyprotein has a different structure to MLV as it contains a p6 region that is involved in the budding of virion particles, whereas in MLV this function lies in the p12 protein.

The 5' end of the HIV RNA genome possesses R and U5 sequences followed by a packaging signal (ϕ) required for incorporation of the genomic RNA into viral particles (Hayashi et al., 1992). At the 3' end of the HIV genome R is repeated followed by U3. The repeated R sequence on the both ends plays an important role in reverse transcription (Fig.1-1, HIV).

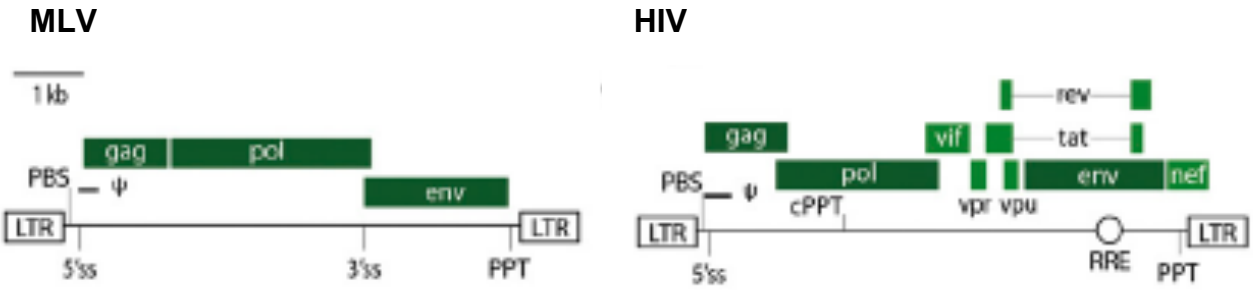


Fig.1-1 Proviral genome structure of MLV and HIV

Structural and regulatory genes (within a green box) composing of each viral genome and their positions are illustrated. LTR: long terminal repeat; PBS: primer binding site; ϕ: packaging signal; cPPT: central polypurine tract; PPT: poluporine tract; RRE: rev response element; ss: splice site. Figure reprinted from (Giacca and Zacchigna, 2012).

1.2 HIV life cycle

1.2.1 Virus entry into a host cell

Virus entry into host cells is the initial step of the life cycle of the retrovirus. Two steps are required for virus entry; engagement of virus env with a surface receptor and co-receptor of a host cell, and the subsequent fusion of virus and host membranes.

1.2.2 Binding of virus envelope to host cell plasma membrane

The initial step of the retrovirus lifecycle is virus entry into a target cell, initiated by binding of the virus envelope to host cell plasma membranes (reviewed in (Overbaugh et al., 2001) (Wilen et al., 2012)). Initial attachment of the retrovirus virion to a host cell surface can occur relatively non-specifically (Sherer et al., 2010). Following this, the envelope must engage with the host receptor proteins involved in virus entry. For HIV this is the CD4 membrane glycoprotein and a G-protein-coupled chemokine receptor. CD4 is a member of the immunoglobulin superfamily and functions to enhance T cell receptor (TCR)-mediated signals by engaging with major histocompatibility complex class II (MHC II) that loads antigen on and leads to subsequent activation and differentiation (Tubo and Jenkins, 2014). CD4 is expressed mainly on helper T cells, but also on monocytes, macrophages, and dendritic cells. The chemokine receptors, also called co-receptors, include cysteine-cysteine chemokine receptor 5 (CCR5) (Choe et al., 1996) or/and chemokine (C-X-C motif) receptor 4 (CXCR4) (Feng et al., 1996).

The viral env is a homotrimer of non-covalently linked heterodimers of glycosylated protein 120 (gp120) and gp41. The gp120 binds to the CD4 primary receptor on the host cell plasma membrane (Kwong et al., 1998). This triggers a conformational change in Env, which is necessary for subsequent co-receptor binding via one of five variable loops within gp120. This co-receptor binding induces exposures of hydrophobic gp41 fusion peptides, resulting in the fusion peptide being inserted into the host cell membrane (Chan et al., 1997).

Membrane fusion occurs when each gp41 in the trimer folds at a hinge region, bringing an amino-terminal helical region (HR-N) and a carboxy-teminal helical region (HR-C) to form a six-helical bundle (6HB) (Weissenhorn et al., 1997). This hydrophobic fusion peptide insertion results in transmembrane anchors in two bilayers in direct apposition, which plays a crucial role in overcoming the hydration force (Weissenhorn et al., 1997). The HIV virion and the host plasma membrane have been reported to fuse in a pH-independent manner (McClure et al., 1990). However, another study showed virus fusion also occurred in an endosomal compartment after virus binding and uptake by endocytosis (Miyauchi et al., 2009).

Following fusion, the HIV virion core is released into the cytoplasm and rearranged to form a complex for conducting reverse transcription, called the reverse transcription complex (RTC). The RTC of Molony MLV (MoMLV) contains the viral RNA genome, RT, IN and CA. The HIV RTC is more complex, containing MA and Vpr together with RT and IN (Hu and Hughes, 2012).

1.2.3 Reverse transcription

Retroviruses package two copies of positive strand RNA. Reverse transcription is one of the defining features of retroviridae discovered in 1970 by two research groups independently (Baltimore, 1970) (Temin and Mizutani, 1970). This changed the accepted concept of the unidirectional flow of genetic information, the so-called “central dogma” which stated that DNA is required to synthesise RNA that is required to synthesise protein. Reverse transcription uses the two copies of the positive strand RNA genome to make viral double-stranded DNA using the viral RT enzyme. HIV RT is a heterodimer comprised of a p66 and a p51 subunit, which is derived from the processed Gag-Pol precursor protein by viral protease into a biologically active form of the enzyme (Goel et al., 1993). Reverse transcription also generates two copies of the long terminal repeat (LTR) on each end of the double stranded DNA genome, the sequence of which is necessary for DNA integration (Yoshinaga and Fujiwara, 1995).

1.2.4 Overview of the mechanism of reverse transcription

The process of reverse transcription is described in Fig.1-2. HIV reverse transcription (reviewed in (Hu and Hughes, 2012)) is initiated by a transfer RNA (tRNA) (Lys3) bound to the primer binding site (PBS) that is located immediately downstream of U5 in the 5' LTR of the HIV RNA genome (Marquet et al., 1995) (Ratner et al., 1985). This primes the synthesis of minus strand cDNA from the 5' end of the RNA genome including U5 and R, termed strong stop. This RNA within this hybrid is degraded by the RNase H function of RT (Molling et al., 1971). The newly synthesised strong stop cDNA then hybrises with the 3' end of the RNA genome because the R sequence is present there. This first strand transfer leads to the extension of minus strand DNA synthesis to 5' end of the template RNA genome. During this step, the template RNA is degraded by RNase H, except the regions highly resistant to RNase H digestion, namely a 3' polypurine tract (PPT) and the central polypurine tract (cPPT). These purine-rich sequences serve as the primers to synthesise the positive strand DNA. Extension of the positive strand DNA is terminated when RT encounters a methylated nucleotide in the tRNA primer. In most retroviruses, RNase H cleavages the entire tRNA primer; however, in HIV RNase H cleaves at one nucleotide from the 3' end of tRNA leaving one additional A at the 5' end of the minus strand DNA. Then, using the complementary PBS sequence appearing on both DNA strands, the second strand transfer occurs. This annealing of PBS sequences leads to a circular DNA intermediate with DNA synthesis proceeding on both strands along each template. In order to copy LTR on both ends of viral DNA, strand displacement synthesis occurs to make the DNA blunt-ended (Fuentes et al., 1996). The completion of reverse transcription results in a longer product than the RNA genome as the provirus DNA genome has the same copy of U3-R-U5 LTR at both ends.

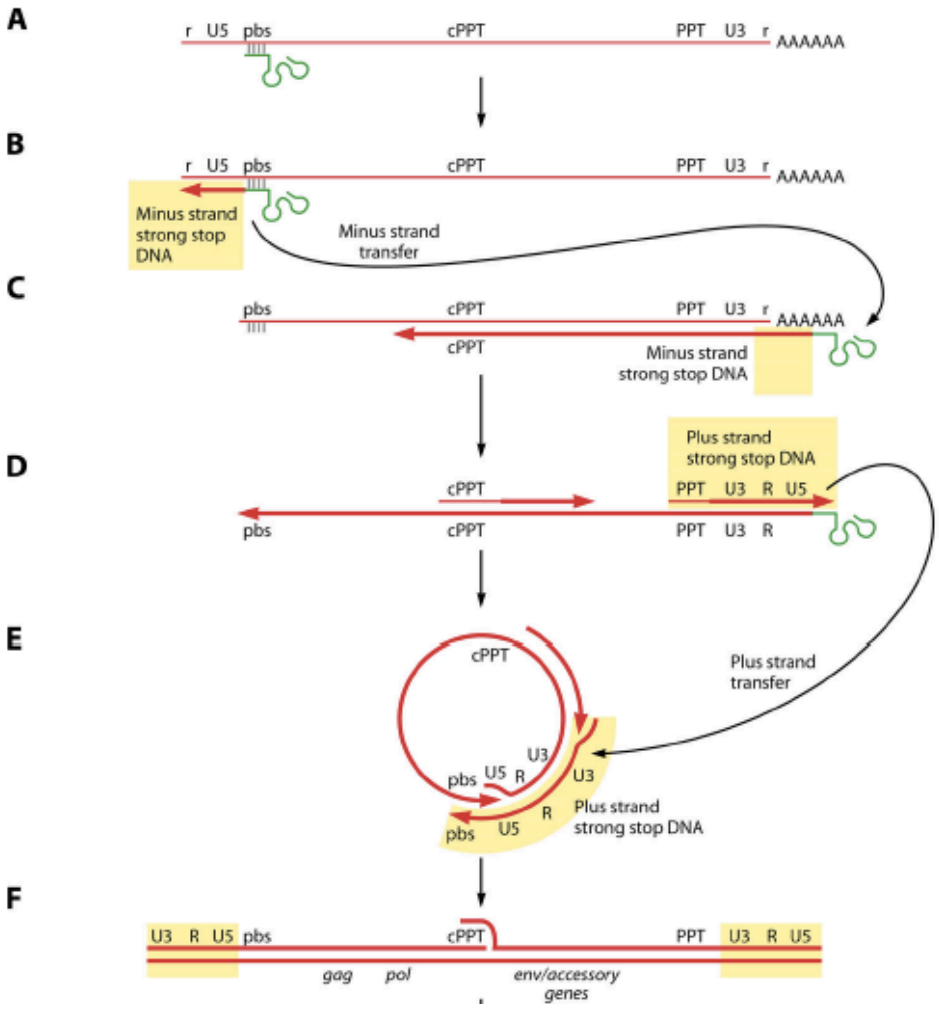


Fig.1-2 The process of reverse transcription

(A) Reverse transcription is initiated by tRNA (Lys3) binding to PBS. (B) Strand extension from the tRNA (Lys3) generates a minus strand strong stop composed of U5 and R. The RNase H function of RT degrades the template RNA within the hybrid. (C) The strong stop transfers to the 3' end of the RNA genome utilising another R sequence. Minus strand extension and subsequent degradation of template RNA, except cPPT and PPT, occurred. (D) Positive strand extension is initiated from cPPT and PPT. (E) PBS sequence on both strands leads to a circular form of the genome, which allows the completion of a DNA extension of a positive strand. (F) LTR is copied on each end of the synthesised DNA. Figure reprinted from (Onafuwa-Nuga and Telesnitsky, 2009).

Interestingly, the presence of multiple transcription initiation sites on the RNA genome affects the speed of the positive strand DNA synthesis. This was demonstrated by HIV infection on 293T and activated $CD4^+$ T cells, showing that the positive strand DNA synthesis was faster than that of the minus strand (Thomas et al., 2007). In addition to the contribution of the cPPT to increase the efficiency of reverse transcription, cPPT was proposed to be necessary for nuclear import of viral RTC (Zennou et al., 2000). Completion of reverse transcription and the formation of the three-stranded DNA flap on the cPPT lead to maturation of RTC into PIC (Arhel et al., 2007).

When vaccines or drugs to prevent or treat the AIDS epidemic are considered, the extensive HIV diversity caused by an accumulation of mutations presents a huge problem such as the emergence of resistant species (Clavel and Hance, 2004) (Praparattanapan et al., 2012). Examination of viral sequences in single infected individuals showed that viral sequences diverge rapidly after infection (Keele et al., 2008). This diversity is facilitated by a high error rate of reverse transcription and the diploid particles with two RNA genomes that allow a combination of mutations. In addition, complicated steps in reverse transcription also contribute to increasing the genomic diversity of HIV. Firstly, the template exchanges twice (tRNA transfers to 3' end (on plus-strand RNA) and PBS transfers to 5' end (on minus strand DNA)) causing recombination events at a high rate *in vivo*. Secondly, the RT enzyme has no proof-reading activity, causing high mutation rates at reverse transcription (Takeuchi et al., 1988) (Preston et al., 1988).

1.2.5 Viral uncoating

HIV uncoating is the next crucial step for producing an infectious virus and is structurally characterised by a loss of viral capsid prior to the entry of the viral genome into the nucleus (Arhel, 2010) (Campbell and Hope, 2015). Uncoating is known to take place at the early stage of HIV infection and accompanies with the transition between reverse transcription complex (RTC) and pre-integration complex (PIC) (Arhel et al., 2007). HIV conical capsid cone contains the viral

RNA genome, viral proteins (CA, NC, RT, IN, Vpr), and some cellular proteins such as cyclophilin A (CypA). Within the protected environment generated by capsid pores, successful reverse transcription takes place (Jacques et al., 2016). The shape and stability of the capsid cone affect viral infectivity by reducing the efficiency of reverse transcription, which was demonstrated by introducing point mutations in the capsid, which led to either hyperstable or unstable capsids or to capsids with aberrant morphologies and caused a reduction in infectivity (Forshey et al., 2002).

Although the functional importance of viral uncoating to successful virus replication is getting clearer, the location and timing of uncoating are still poorly elucidated because of difficulties in measuring the process (Campbell and Hope, 2015). When HIV proviral genome accesses host cellular chromatin through the nuclear pore, uncoating is definitely required because the diameter of viral capsid exceeds that of the nuclear pore (more details in 1.2.7 nuclear import of the reverse transcribed virus genome). A substantial difference in mass between HIV-1 complex in cytoplasmic and nuclear suggests the occurrence of uncoating before nuclear entry (Iordanskiy et al., 2006). Therefore, the failure of uncoating leads to an accumulation of HIV-1 complexes at the cytoplasmic surface of the nuclear membrane (Arhel et al., 2007). In addition, recent studies revealed that the timing of some degree of uncoating is associated with reverse transcription. Inhibition of reverse transcription delays uncoating suggesting that uncoating occurs later than the initiation of reverse transcription (Hulme et al., 2011). Another study showed that a remnant of the hexagonal CA lattice stays intact as the PIC traffics to the nucleus (Ambrose and Aiken, 2014). This uncoating model shows that CA is not completely removed in the cytoplasm and some capsid is associated with PIC in the nucleus. In contrast, a different study supports that CA removal may start when RTC arrives at the nuclear pore complex (NPC) to protect the replicating HIV genome from the host DNA sensor and induction of host immune reaction (Lahaye et al., 2013) (Rasaiyaah et al., 2013). The observation of intact cores at the NPC also supports this uncoating model (Arhel et al., 2007).

1.2.7 Nuclear import of the reverse transcribed virus genome

The HIV PIC, composed of virus genome MA, IN and Vpr, facilitates nuclear import (Matreyek and Engelman, 2013) allowing access to host cell chromatin and integration in non-dividing cells (Naldini et al., 1996). In contrast, gammaretroviruses, such as MLV, require a breakdown of the nuclear membrane during mitosis (Roe et al., 1993) explaining why they only infect dividing cells. Although the molecular mechanism of nuclear entry of PIC is still poorly understood, one possible hypothesis is nuclear localisation signals (NLSs) present in all of PIC components (Zennou et al., 2000) (Bukrinsky et al., 1993) (Gallay et al., 1997) (Heinzinger et al., 1994). However, HIV lacking Vpr (Agostini et al., 2000), or cPPT (Dvorin et al., 2002), could infect non-dividing cells productively suggesting that these NLSs are not essential for nuclear import. In addition, the role of the NLS in MA has been disputed (Freed et al., 1995). The IN NLS is therefore likely the most critical. When reverse transcription is completed, IN hydrolyses the extremities of the linear viral DNA to be ready for integration (see more details in 1.2.8).

The HIV PIC must pass through NPCs to traverse from the cytoplasm into the nucleus (Matreyek and Engelman, 2013). NPCs cross the nuclear membrane and are responsible for highly selective and bidirectional transport of various proteins and ribonucleoprotein (Wente and Rout, 2010). Theoretically, molecules up to 9 nm in diameter (~60 kDa for a globular protein) can be transported through NPCs by diffusion. However, the HIV PIC is estimated to be considerably larger so needs to take a different strategy for migration into the nucleus (Mattaj and Englmeier, 1998). One mechanism of nuclear transport involves the interaction between the NLS within IN in the PIC and members of an importin-a protein family. The NLS is initially recognised by importin-a, then a heterodimer of importin-a, and importin-ß interacts with the cargo-receptor complex. This complex is transported through the NPC in an energy-dependent manner (Fassati et al., 2003). After the translocation of the complex, the cargo molecules are released into the nucleoplasm when the GTP-bound form of the GTPase Ran associates with importin-ß in the complex (Mattaj and Englmeier,

1998).

Another mechanism that has been described involves the PIC docking to NPCs by some remaining CA molecules interacting with the nuclear transport factor NUP358 (Schaller et al., 2011). The PIC then engages other nuclear transport factors, NUP153 and cleavage and polyadenylation specific factor 6 (CPSF6). NUP358, NUP153, and CPSF6 at the nuclear pore likely act on PIC-associated CA to aid HIV infection. Transportin 3 (TNPO3) then allows nuclear import of PIC-associated CPSF6 possibly by binding to the C-terminal nuclear targeting the arginine/serine-rich domain of CPSF6 (Kataoka et al., 1999). Mutation of CA residues required for NUP358 interaction results in PIC nuclear import by the default importin a/b mechanism. Changes in the integration site preference were also observed (Schaller et al., 2011).

1.2.8 Integration of viral DNA into host chromatin

For HIV to maintain stable proviral DNA in the host cell and its progeny, integration is catalysed by the retroviral protein integrase (IN) (reviewed in (Krishnan and Engelman, 2012) (Craigie and Bushman, 2012)). The structure of retroviral IN was determined by crystal structure analysis of prototype foamy virus IN in complex with its cognate DNA (Hare et al., 2010). This revealed that integration complex is composed of a dimer of IN dimers and that each monomer plays a distinct role. In addition, an extended conformation of the inner monomer allows the contact with the viral LTR ends. HIV IN by inference also consists of three or four distinct structural and functional sub-domains, namely the N-terminal domain (NTD), the core catalytic domain (CCD), the C-terminal domain (CTD), and the N-terminal extension domain.

The reaction steps of the DNA cutting and joining during integration were discovered by structural analysis of integration intermediates (Engelman et al., 1991) (Fujiwara and Mizuuchi, 1988). The three steps to incorporate double stranded proviral DNA into host chromatin are described in Fig.1-3. Firstly, before the PIC enters the nucleus, IN removes two nucleotides from 3’ ends of

the double-stranded provirus DNA, resulting in CA-3' sequence; this step is called 3' processing. After nuclear entry of the PIC, the recessed 3' ends attack phosphodiester bonds in the target host cell DNA strands to join one DNA strand. Four to six bases (four for MLV, five for HIV) of the targeted site remain single stranded; this step is called strand transfer. Host cell DNA repair fills in those gaps, generating a four to six base-pair duplication flanking the provirus. This final step is known as gap repair.

During lentivirus integration, an interaction between the host protein lens epithelium-derived growth factor (LEDGF/p75) and IN is required to guide the PIC into the target sites on the host chromatin (Debyser et al., 2015). The N-terminal region of LEDGF/p75 possesses a nuclear localisation signal and a PWWP (Pro-Trp-Trp-Pro), chromatin binding elements, and A/T hook like elements, while the C-terminal of the LEDGF/p75 harbors the integrase-binding domain (IBD) that binds tightly to IN. The absence of LEDGF/p75 reduces virus infectivity (Llano et al., 2004) (Llano et al., 2006) and alters the degree of the HIV integration site in the transcription unit (Marshall et al., 2007).

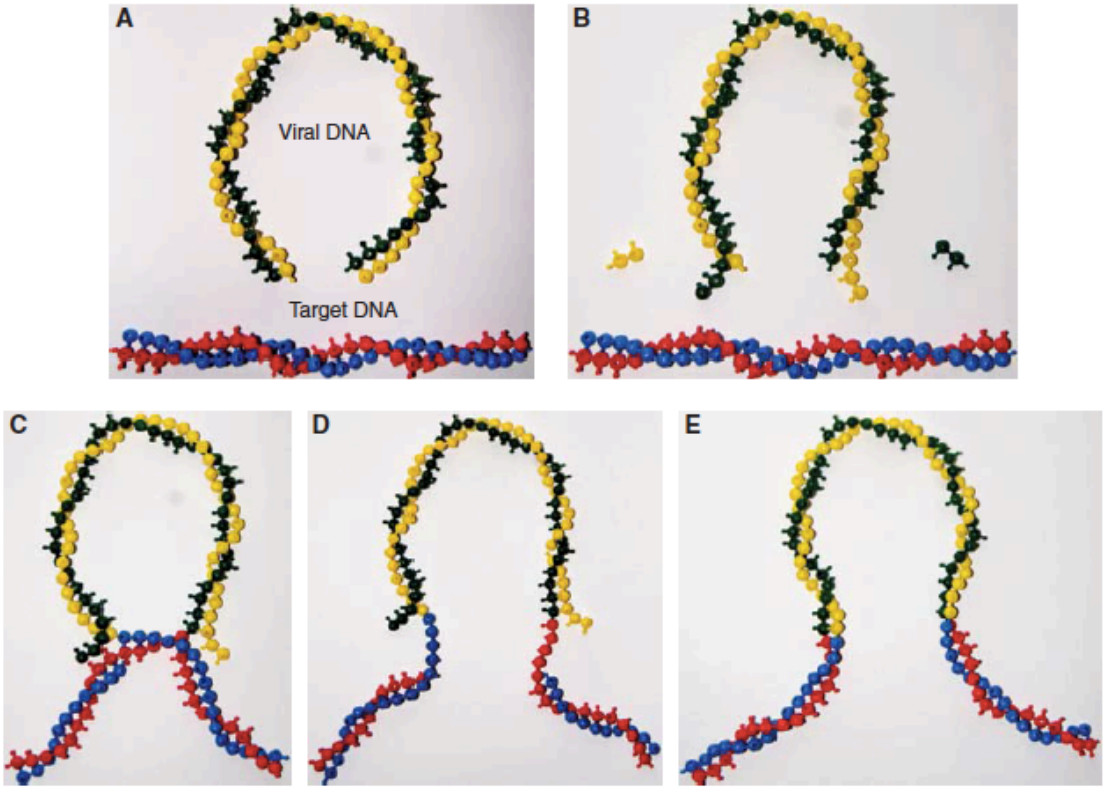


Fig.1-3 The principle of HIV integration

DNA bases are shown by coloured-balls. (A) The linear blunt-ended viral DNA (green and yellow) and target DNA (blue and red) prior to integration reaction. (B) 3' end processing of viral DNA. In most cases, two nucleotides are removed from each 3' end of the viral DNA. (C) The 3' ends created by 3' processing attack a pair of phosphodiester bonds in the target DNA. The 3' ends of the viral DNA are then joined to the 5' ends of the target DNA at the site of integration. The 5' ends of the viral DNA stay unjoined to the target DNA. (D) Provirus formation is completed when the two unpaired bases at the 5' ends of the viral DNA are removed, filling in the single-strand gaps between viral and target DNA, and the 5' ends of the viral DNA to target DNA is ligated. The 3' processing (B) and DNA-strand transfer steps to form the integration intermediate (C) are catalysed by IN. Subsequent steps are thought to be catalysed by cellular enzymes. (E) The integrated provirus in target DNA. Figure reprinted from (Craigie and Bushman, 2012) with permission.

1.2.9 Virus integration preference

Integration site analysis became available upon the arrival of the complete human genome in 2001 (Lander et al., 2001) (Venter et al., 2001). In addition, the development of next-generation sequencing accelerated analysis of virus integration in a quantitative manner by collecting a much larger number of integration sites (Craigie and Bushman, 2012).

Integration studies revealed that different retroviruses have different targeting bias in integration sites (Cavazza et al., 2013). For instance, MLV favors integration near transcription start sites and CpG islands, while HIV favors active transcription units (Wu et al., 2003). The favored integration sites of the retrovirus are either called “hotspots” or “common integration sites (CIS)”. CIS are associated with gammaretrovirus- or transposon-induced hematopoietic malignancies; therefore, those sites are selected by cell proliferation and oncogenesis (Suzuki et al., 2002). However, “hotspots” denotes the genomic region where integrations accumulate more than expected by chance in the absence of any selection process (Cattoglio et al., 2007).

Molecular mechanisms of integration site targeting are yet to be clarified. The difference in the selection of the integration site between MLV and HIV is possibly due to the difference in cell state when each virus can infect, as HIV can infect non-dividing cells and will encounter different states of chromatin compared to those by MLV. Another possible factor to distinguish integration site selection between HIV and MLV is IN. Chimeric viruses based on HIV generated by swapping HIV IN with MLV IN showed similar integration site targeting to MLV (Lewinski et al., 2006), which revealed that integration site preference could be governed by IN. Host proteins to guide PIC interacting with IN into the targeted genomic site are different between MLV and HIV (reviewed in (Craigie and Bushman, 2014)). As in 1.2.8, HIV utilises the interaction with LEDGF/p75 for integration, while MLV also has different tethering factors, bromodomain and extra terminal (BET) protein 2, 3 and 4 (De Rijck et al., 2013). In addition, transporter proteins in **1.2.7** are also involved in integration site selection. Upon

factor knockdown, host proteins crucial for nuclear entry, TNPO3, NUP153, and NUP358 were found to be important for targeting HIV PICs to gene-dense regions in the chromatin (Ocwieja et al., 2011) (Koh et al., 2013).

The integration site has been proposed as one factor that may influence vector safety. HIV-based vectors may be safer than MLV as they integrate into active genes but without preference for regions near the transcription start site. Other retrovirus family members, such as avian sarcoma-leukosis virus (ALSV), show even less specificity for gene-dense regions (Mitchell et al., 2004) and mouse mammary tumor virus (MMTV) is the most random integrator to date (Faschinger et al., 2008). However, integration site selection is only a preference and with a high number of integrants most genomic loci will be targeted. Also, the ability of a vector to affect neighbouring host gene expression is likely to be a greater problem than integration site selection.

1.2.10 Virus gene expression

HIV has a complex genome and requires the expression of accessory proteins as well as viral structure proteins; therefore, gene expression is controlled by more mechanisms than that of viruses with simple genomes such as MLV. HIV has two phases of gene expression; the initial expression of multiply spliced regulatory genes and the subsequent expression of singly/unspliced *env*, *gag-pol* and RNA genome (Karn and Stoltzfus, 2012).

1.2.11 Transcription initiation for producing infectious HIV

In order to produce infectious virus particles the integrated provirus needs to be transcribed then translated for viral protein expression. Transcription of HIV RNA is carried out by host RNA polymerase II. The HIV LTR possesses multiple DNA regulatory elements that participate in the cooperative binding of transcription initiation factor TFIID (Rittner et al., 1995). Transcription is initiated at the junction of U3 and R. In addition, NF-kB binding and other motifs within the LTR allow binding of inducible transcription factors, such as the NF-kB family and NFAT, so that transcription is coordinated with the cell activation status (Liu et al.,

1992) (Kinoshita et al., 1998).

One of the first proteins to be produced is the transactivator protein (Tat), translated from a multiply spliced RNA. Tat then provides a feedback loop to stimulate transcription initiation of HIV RNA by binding to a 59-nucleotide stem-loop structure on the HIV genome called the transactivation response (TAR) element (Gatignol, 2007). After the transcription through TAR, positive transcription elongation factor b (P-TEF-b) composed of cyclin T1 kinase (CycT1), cyclin-dependent kinase 9 (CDK9), and other accessory elongation factors are recruited to the TAR and increase the affinity of Tat with TAR (Zhang et al., 2000a). This results in hyper-phosphorylation of the C-terminal domain of RNA polymerase II and elongation factors, thereby stimulating efficient transcriptional elongation (Kim et al., 2002).

Transcripts of HIV serve distinct functions; firstly, unspliced viral genomic RNA is packaged into the daughter virions (9 kb) or acts as a template for translation of structural protein (Gag and Gag-Pol). This transcript is also spliced for the production of diverse subgenomic mRNAs (Vif, Vpr, Vpu/Env, Tat, Rev, and Nef,) (Pollard and Malim, 1998).

1.2.12 Viral RNA export from the nucleus to cytosol to express viral proteins

At the early stage of infection, only fully spliced viral transcripts such as those encoding Tat, Rev, and Nef are exported to the cytoplasm using the nuclear export factor NXF1 that couples splicing to nuclear export (Stutz and Izaurralde, 2003). Usually unspliced transcripts are retained within the nucleus and degraded by the host cellular surveillance system, nonsense mediated decay (NMD) (Chang et al., 2007). HIV needs a mechanism for the export of these unspliced transcripts into the cytoplasm; this is controlled by the expression level of the regulatory protein Rev. In later stages of infection the full-length unspliced HIV RNA genome, as well as 4-kb transcripts coding Env/Vpu, Vif, and Vpr, is exported by Rev binding to a cis-acting RNA element called Rev-response

element (RRE) within *env* (Pollard and Malim, 1998).

Rev protein is imported into the nucleus by interacting with importin ß through its NLS that contains an arginin-rich sequence (Daugherty et al., 2010b). Rev then binds to RREs in HIV transcripts via its RNA-binding domain (RBD), which induces a conformational change and oligomerisation of the Rev protein onto the RRE (Daugherty et al., 2010a) (Daugherty et al., 2010b). Then, this complex interacts with the nucleocytoplasmic transport factor chromosome region maintenance 1 (Crm1), in conjunction with the GTP-bound form of the Ran GTPase (Cullen, 2003). Crm1 was found to form a dimer when it interacts with Rev-mRNA which promotes nuclear export (Booth et al., 2014). Dissociation of this complex in the cytoplasm is achieved by hydrolysis of Ran-bound GTP to GDP.

1.2.13 Assembly of HIV viral components

Virus assembly is then required to bring viral proteins and genomic RNA into a viral particle which can bud from the plasma membrane (Freed, 2015). The main structural protein, Gag, is synthesised as a 55 kDa precursor polyprotein in the cytoplasm from full-length viral RNA. The Gag precursor contains matrix (MA), capsid (CA), nucleocapsid (NC), and p6 domains, as well as two spacer peptides, SP1 and SP2. During the translation of Gag RNA, the 160 kDa GagPol precursor protein that encodes viral enzymes is synthesised by programmed ribosomal frameshift and expressed at approximately 5 % of the level of Gag (Shehu-Xhilaga et al., 2001).

Viral RNA encapsidation is initiated by the NC domain of Gag recruiting viral genomic RNA following dimerisation of the RNA via its packaging signal ϕ (Kutluay et al., 2014). Once the multimerised Gag arrives at the plasma membrane, it is anchored to the plasma membrane by insertion of its amino-terminal myristate within MA. This mechanism of targeting to the plasma membrane by the MA domain in Gag was demonstrated by mutations in MA, which resulted in mistargeting to the late endosome or multivesicular body

(MVB) (Ono and Freed, 2004). The same mistargeting result was found by depletion of plasma membrane phosphatidylinositol 4,5-bisphosphate (PtdIns(4,5)P_2), highly enriched in the inner leaflet of the plasma membrane, suggesting the importance of targeting to lipid rafts (Ono et al., 2004).

The viral Env glycoproteins traffic via the secretory pathway from the rough endoplasmic reticulum (RER) through Golgi to vesicles that arrive at the plasma membrane. There are several possible non-mutually exclusive pathways for Env incorporation in retroviral particles (reviewed in (Checkley et al., 2011)); a passive process, co-targeting of Gag and Env to a common site on the plasma membrane, direct recruitment of Env by Gag, and the indirect recruitment of Env by Gag via host cell binding proteins. HIV interaction between Env and MA is suggested for Env incorporation in the assembling particle because MA mutants block Env incorporation (Murakami and Freed, 2000).

1.2.14 HIV budding and release

The assembled viral proteins undergo membrane fission for the release of daughter virions from the infected cells, which is mediated by the cellular endosomal sorting complexes required for the transport (ESCRT) pathway (reviewed in (Votteler and Sundquist, 2013)). The cellular function of this pathway is to sort ubiquitylated membrane proteins into the lumen of maturing endosomes or MVB and subsequent degradation or fusion with the plasma membrane to release extracellular exosomes. Recruitment of three broad classes of factors occurs in the ESCRT pathway: adaptor proteins, early-acting factors, and late-acting factors. Structural protein Gag is an example of the adaptor protein and recruits the early-acting factors that are related to subsequent recruitment of late-acting factors. Gag assembly recruits early-acting factors in the ESCRT pathway, tumour susceptibility gene 101 (TSG101, a part of ESCRT-I) a subunit of ERCRT-I, ALG2-interacting protein X (ALIX), ESCRT-I, and ESCRT-II to form a complex (Göttlinger et al., 1991). HIV Gag has two late domain motifs; PT/SAP (Pro-Thr/Ser-Ala-Pro) and YPXL (Tyr-Pro-X-Leu, where X can be any amino acid). TSG101 interacts with the

PT/SAP motif and ALIX interacts with YPXL. ALIX is a protein that also interacts directly with both ESCRT-I and ESCRT-III in mammals. Membrane scission is driven by ESCRT-III polymerisation at the virus budding site in concert with the activation of the ATPase vacuolar protein sorting 4 (VPS4) (Baumgartel et al., 2011). In addition, ESCRT-III protein CHMP4 subunits, possibly in a complex with CHMP2 and other ESCRT-III protein, form spiraling filaments within the neck of the budding virus, which promote membrane fission (Hanson et al., 2008).

1.2.15 HIV virion maturation

Newly formed virion particles undergo maturation concomitant with budding (or immediately after) to make the virus infectious. During the maturation process triggered by PR cleavage of Gag and GagPol (Kaplan et al., 1993), the locations of processed viral proteins are dramatically rearranged (Schur et al., 2015) (Freed, 2015). In immature particles, Gag molecules are packaged in a radial manner then, following the cleavage of Gag, CA proteins reassemble to form a conical structure to wrap the viral RNA genome (Schur et al., 2015).

1.2.16 Host restriction factors and accessory gene function

Although HIV exploits host proteins in the replication cycle, host cells have developed defence systems against pathogens including HIV, called innate immune sensing (Kajaste-Rudnitski and Naldini, 2015). In particular, the intrinsic antiviral immunity via host restriction factors is characterised by an immediate and direct antiviral effect (Yan and Chen, 2012). Below is a description of some of the host restriction factors that prevent HIV infection. These factors can also be an inhibitor of LV-based vector transduction, which may determine the outcome of *ex vivo* gene therapy.

The cytidine deaminase function of apolipoprotein B mRNA editing enzyme catalytic subunit 3G (APOBEC3G) protein catalyses the deamination of cytidine to uridine in single stranded DNA. This inhibits HIV infectivity by introducing

mutations in the viral sequence (Sheehy et al., 2002) (Zhang et al., 2003). In addition, APOBEC3G inhibits tRNA (Lys3) priming of reverse transcription and its removal process at the DNA minus strand synthesis (Guo et al., 2006) (Mbisa et al., 2007). However, HIV accessory protein Vif counteracts APOBEC3G and triggers its ubiquitilation that leads the degradation of APOBEC3G (Yu et al., 2003).

Tripartile motif protein 5 alpha (TRIM5a) blocks HIV infection in cells of Old World monkeys through recognition and degradation of the capsid lattice through its E3 ubiquitine ligase function. Interestingly, human TRIM5a has little or no antiviral effects against HIV (Stremlau et al., 2004). However, when choosing a primate model to test LV-based vector, this host restriction factor can affect transduction efficiency and impact the biological function of host cells. GTPase Mx2 does however interact with the HIV capsid and blocks HIV nuclear entry (Goujon et al., 2013) (Kane et al., 2013) (Liu et al., 2013).

Sam domain, and HD domain-containing protein 1 (SAMHD1), inhibits the reverse transcription process by reducing the level of nucleotide pool. Many simian immunodeficiency viruses (SIVs) encode Vpx to interfere with the effect of SAMHD1. However, HIV lacks this protein thereby a block of HIV infection is observed in resting dendritic cells and T cells (Laguette et al., 2011) (Baldauf et al., 2012).

A dimeric type II membrane glycoprotein tetherin inhibits the release of mature retrovirus particles and antagonised by a HIV accessory protein, Vpu (Neil et al., 2008). Tetherin expressed at the cell surface is downregulated by counteracting by Vpu, whose pathway is mediated by ESCRT lysosomal degradation (Janvier et al., 2011). In contrast, the majority of SIVs use Nef to counteract the tetherin proteins of their non-human primate hosts (Zhang et al., 2009). The domain on human tetherin that is important to tetherin/Nef interaction is absent (Sauter, 2014) and Nef was thought nothing to do with the antagonism of human tetherin. However, an HIV strain (called group O) common in West-Central Africa

demonstrated the similar inhibitory effect of Nef against human tetherin (Serra-Moreno, 2014).

The host integral membrane proteins, serine incorporator 3 (SERINC3) and 5 (SERINC5), were recently described as host restriction factors which can be counteracted by Nef (Usami et al., 2015) (Rosa et al., 2015). Viral infectivity was reduced by the incorporation of SERINC3 or 5 into a Nef-defective HIV (Usami et al., 2015). In addition, an inhibition model via SEINC3 and 5 was proposed, which is mediated by prevention of the pore expansion between viral and cell membranes (Rosa et al., 2015).

1.3 Viral vectors

1.3.1 Retroviral vectors as a delivery tool to introduce therapeutic genes into target cells

The idea of virus genome as a tool for gene transfer came from the discovery of bacteriophage genome in *E.coli* as well as from tumour formation in mammal cells by infection of oncogenic retroviruses. Oncogenic retrovirus, such as avian reticuloendotheliosis virus strain A (REV-A) and MLV, were initially engineered as such promising candidates because of their broad tropism as well as integration capacity into the host genome. Permanent vector integrations allow long-term and stable gene expression in transduced host cells. All highly oncogenic viruses were engineered into being replication-defective by deleting the coding regions of viral proteins or by replacing them with an exogenous gene in order to reduce the production of replication competent viruses (RCVs).

The first engineered vector tested on animal cells was based on oncoretrovirus vector MoMLV, carrying herpes simplex virus thymidine kinase gene (*tk*), driven by its own transcriptional promoter (Tabin et al., 1982). In the same study, the cell line producing MLV carrying the constructed vector plasmid as a genome was established by compensating with viral genes, whose cell line was later called packaging cell lines (PCLs). To reduce the risk of RCVs, the design of vector constructs was improved by separating *gag-pol* and *env* and splitting them on separate plasmids (Danos and Mulligan, 1988) (Miller et al., 1991). Later, env from another virus strain was borrowed to widen the tropism of produced vector viruses as described in 1.3.3. The ability to deliver genes via MLV vector to bone marrow stem cells from mouse and human were soon demonstrated (Williams et al., 1984) (Hock and Miller, 1986).

The history of the lentiviral vector (LV) started as a tool to screen anti-HIV-1 drugs (Sakuma et al., 2012). A replication-competent HIV vector encoding the chloramphenicol actyltransferase (CAT) gene, in the place of *nef,* was initially constructed for that purpose (Terwilliger et al., 1989). The infectivity of a

produced virus can be measured by its correlation with the enzyme activity of CAT. This vector was a promising model for screening anti-HIV drugs in a rapid and quantitative manner. Unlike MLV, LV was shown to be able to infect non-dividing cells, which widened the vector application to different cell types such as hematopoietic stem cells (HSCs) or neurons (Roe et al., 1993) (Naldini et al., 1996). This is a milestone of LV application to challenge a broader range of genetic diseases such as hematopoietic and neurodegenerative diseases (Naldini, 2015). In addition, in order to increase vector biosafety, the design of splitting virus structure-coding protein into multiple vectors was also applied to LV as described in the following section 1.3.2. Transient vector virus production is often exploited because the constant expression of gag-pol and env proteins can be cytotoxic in virus producer cells; however, some lentiviral vector-based PCLs have been developed, including WinPac that was established in our laboratory (Kafri et al., 1999) (Sanber et al., 2015).

1.3.2 Vector 'generations'- iterative improvements of vector safety and efficacy

Vector 'generations' have been developed to achieve better biosafety features of lentiviral vectors (Sakuma et al., 2012) (Fig.1-4). The "generation" is loose terminology that does not include the replication-competent prototype of the LV.

The first-generation replication-deficient recombinant HIV vectors are comprised of three vector components that provide viral genes necessary to produce virions: a packaging construct (*gag-pol*), viral envelope glycoprotein gene (*env*), and a gene of interest (Fig.1-4A). The vector carries the gene of interest together with HIV cis-acting elements such as LTRs, packaging signal (ϕ), and RRE. By providing those viral components on separate plasmids, production of RCLs can be reduced because more recombination events are required for the production (Richardson et al., 1995).

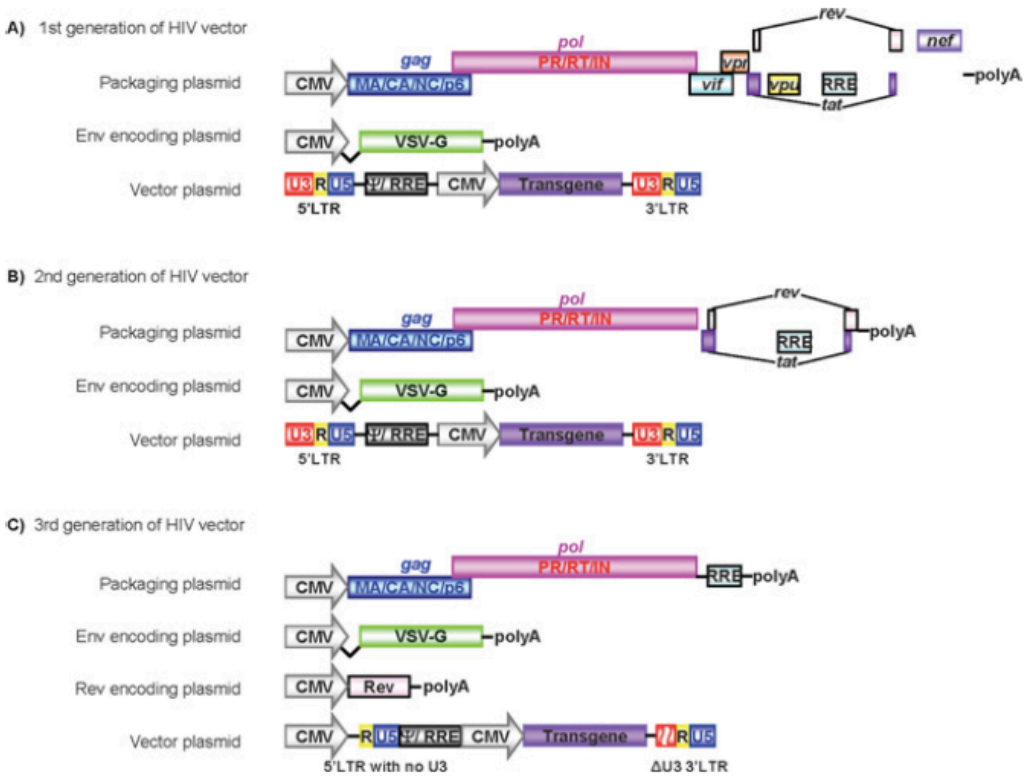


Fig.1-4 Vector generations

(A) First generation vector. All HIV proteins except Env are expressed in the packaging plasmid. Env is provided in a different vector. The vector plasmid has intact LTR. (B) Second generation vector removed most of the HIV proteins except Rev and Tat. (C) Third generation vector. Packaging plasmid encodes Gag and Pol. Rev is provided in a separate vector. The vector plasmid has self-inactivating LTRs. Figure reprinted from (Sakuma et al., 2012).

The second generation vectors are focused on the exclusion of HIV accessary genes that are included in the design of the first-generation vector (Fig.1-4B). As mentioned in 1.1 Retrovirus biology, the accessary proteins Vif, Vpu, Vpr, and Nef are necessary for efficient propagation and virulence of HIV in primary cells or *in vivo*. However, exclusion of those elements did not change vector virus infectivity; therefore, this vector design without those viral elements became commonly used (Zufferey et al., 1997).

These vector generations contained HIV regulatory proteins Tat and Rev in the packaging plasmid because transcription of the early vector generation was tat-dependent. To achieve further vector safety, the 5' U3 promoter region was replaced with a strong viral promoter from Cytomegalovirus (CMV) or Rous sarcoma virus (RSV), thereby tat is no longer necessary in a vector to generate vector genomic RNA (Fig.1-4C). This also led to the success in vector design without an enhancer/promoter region in the LTR, a self-inactivating (SIN) vector achieving high titre vector production. *Rev* is also provided in a separate vector. This four-plasmid system for vector production is nowadays called third generation vectors.

1.3.3 Pseudotyping viral vectors to circumvent limited viral tropism

The host range of retrovirus is determined and therefore limited by the nature of the retroviral envelope proteins (Hunter and Swanstrom, 1990). HIV can infect human cells expressing a CD4 receptor because HIV Env recognises it for the virus entry. To widen the tropism of the virus so that it is able to infect other cell types, Env glycoprotein can be replaced with that of a different virus strain such as the G protein from vesicular stomatitis virus (VSV-G). This process is called pseudotyping. Retrovirus is an attractive candidate as a vector vehicle because its envelope can be easily pseudotyped. The commonly used VSV-G has better stability than the retroviral or lentiviral envelope, therefore the VSV-G pseudotyped virus can be easily concentrated by ultracentrifugation without loss of infectivity (Burns et al., 1993). A recent study showed that the major receptor of VSV and VSV-G pseudotyped lentiviral vector was a low-density lipoprotein (LDL) receptor whose binding depends on the concentration of calcium ions (Finkelshtein et al., 2013). The drawback of VSV-G is cytotoxicity in high concentration (Hoffmann et al., 2010). This accelerated the study of discovering env proteins from other viral strains that can alternate with VSV-G. For instance, amphotropic MLV Env or RD114 Env-pseudotyped lentiviral vectors showed high vector titre in human $CD34^+$ hematopoietic stem cells (Relander et al., 2005). Again, the strength of VSV-G is its stability, hence those vectors pseudotyped with env protein of another retrovirus showed compromised

robustness and sensitiveness to the freeze-thaw cycle (Strang et al., 2004).

1.3.4 Lentiviral vector accessory elements

Exogenous elements are also playing an important role for transgene expression. The introduction of woodchuck hepatitis virus posttranscriptional regulatory elements (WPRE) in the vector can achieve improvement in titre and transgene expression (Zufferey et al., 1999). As examples for exogenous elements to prevent influences from the surrounding chromatin environment in which the vector integrates, insulator DNA sequences, such as chicken hypersensitive site 4 (cHS4) or human HNRPA2B1-CBX3 UCOE (A2UCOE), have been used. The promising element A2UCOE can block DNA methylation-mediated silencing; therefore, the target gene expression is position-independent. Our laboratory conducted the splicing site analysis of A2UCOE to find potential splice sites and develop the optimised form of the element by deleting them but without losing the gene silencing-resistance function (Knight et al., 2012).

Codon optimisation is another strategy to increase vector titre and transgene expression that affects the level of immune reconstitution in the treatment of X-linked severe combined immunodeficiency (Moreno-Carranza et al., 2009). In this study, the titre of a cDNA-optimised vector was increased 10-fold by changing the GC content and removing the RNA instability motif in *gp91*^{phox} cDNA for X-linked chronic granulomatous disease (X-CGD) treatment. Codon optimisation can also remove consensus and cryptic splice sites, which can contribute to reducing splicing-mediated side effects.

Previously, MLV-derived gammaretroviral vectors (GRVs), with enhancer/promoter region in the LTR, upregulated oncogenes near integration sites and caused adverse events at a high rate in several clinical trials (Hacein-Bey-Abina et al., 2008) (Howe et al., 2008) (Braun et al., 2014). Therefore the design of LVs has been developed to attain safe profiles by lowering vector-derived side-effects (Schambach et al., 2013) (Sakuma et al.,

2012). Oncogenic events were also observed by aberrant splicing between the integrated viral sequence and host sequence in fusion transcripts (Wang et al., 2010). Most of the vector splice sites are to be removed from the vector backbone sequence because they can potentially contribute to aberrant splicing. However, all splice sites cannot be deleted because some splice sites are required for efficient virus production, such as the HIV major splice donor (D1) and a splice acceptor in the RRE/Env fragment (A7). In addition, cryptic splice sites are hidden in the vector backbone and this can also be the cause of generating chimeric transcripts (Cesana et al., 2012). Such aberrant transcripts are possibly kept at a low level owing to the elimination by NMD or nuclear retention because mostly immature polyadenylation is introduced into the transcripts (Moiani et al., 2012). However, the unexpected chimeric transcript can lead to clonal dominance or cancerous events when the chimeric products stabilise or destabilise the gene product and dysregulate the original function of the gene (Cavazzana-Calvo et al., 2010).

1.4 Gene therapy

1.4.1 Gene therapy: gene addition and gene editing

Gene therapy is the use of somatic cell genetic modification for medicinal purposes. There are two main types of gene therapy; gene addition and gene editing (Kaufmann et al., 2013) (Naldini, 2015). Gene addition involves delivering therapeutic nucleic acids (DNA or RNA) into a patient's cells, usually by a viral vector, as these are at present more efficient than non-viral methods. Retroviral vectors (RVs), adenoviral vectors (AVs), and adeno-associated vectors (AAVs) are the most commonly used in clinical trials worldwide (http://www.abedia.com/wiley/) (Fig.1-5A). In contrast, site-directed gene-editing, using engineered site specific nucleases or CRISPR/Cas9, can be used to mutate unwanted genes or repair defective ones. So far, more than 60 % of gene therapy clinical trials have been performed to treat cancer while nearly 10 % have been for monogenic disorder worldwide (Fig.1-5B). Because my PhD study focuses on cell transformation by LV, the application of LVs in various clinical trials is introduced in contrast to the difficulties of treatment by conventional GRVs.

A **B**

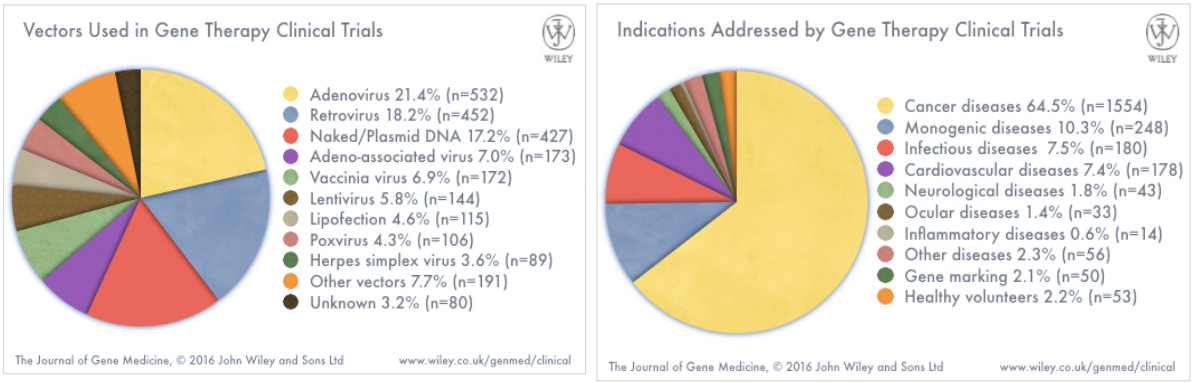


Fig.1-5 The technique of transferring therapeutic agents in gene therapy clinical trials (A) and the diseases treated by gene therapy clinical trials (B)

The data were extracted from http://www.abedia.com/wiley/.

1.4.2 *Ex vivo* gene therapy and *in vivo* gene therapy

In gene therapy, there are two ways to deliver therapeutic agents via viral vectors; *ex vivo* and *in vivo*. *Ex vivo* gene therapy delivers a therapeutic agent into patient-derived cells outside the body and re-infuses the engineered cells. My introduction will focus on the use of hematopoietic stem cells (HSCs). Some benefits of *ex vivo* gene therapy are that a lower dose of the vector is necessary and there are lower off-target effects such as ectopic expression and germ-line transmission. In contrast, *in vivo* gene therapy delivers a therapeutic agent directly into a patient's body by injection. Non-viral vectors such as lipid- or polymer-based vectors are also an option in *in vivo* gene delivery. However, few of these have been used clinically because of the low efficiency of gene delivery compared to that of viral vectors (Putnam, 2006; Yin et al., 2014).

1.4.3 Hematopoietic stem cell gene therapy (HSCGT)

Patients with inheritable disorders, such as primary immunodeficiencies (PID), first became a target for treatment by allo-transplantation from human leukocyte antigen (HLA)-matched donors. However, this is not universally applicable because every patient does not have an HLA-matched donor leading to problems with graft-versus host diseases (GVHD) (Gennery et al., 2010) (Kang and Gennery) (Mukherjee and Thrasher, 2013). So gene therapy using patient-derived HSCs became the alternative life-saving strategy, overcoming those limitations. To date, LV-mediated HSCGT clinical trials are ongoing in a number of fields such as inherited immune deficiencies (WAS, X-SCID, CGD, ADA-SCID), neurodegenerative disorders (X-ALD, MLD), hematological disorders (ß-thalassaemia), Parkinson's disease, and AIDS. LV-mediated gene therapy has progressed because of limitations in GRV-mediated gene therapy, in particular the detection of adverse events leading to cancer where the GRV LTR has caused upregulation of a gene neighbouring the integration site (Mitchell et al., 2004) (Yu et al., 1986). LV is potentially safer than GRV in this respect and is also more efficient in transducing HSC. To understand this historical shift of vector use in clinical trials I will introduce examples of LV use.

1.4.4 Gene therapy for primary immunodeficiencies (PIDs)

Many monogenetic disorders which impair the hematological and immunological system require phenotypic correction in HSC or early progenitors because the gene mutation affects the generation and differentiation of a specific cell lineage (reviewed in (Kaufmann et al., 2013)). As such examples, here I introduce adenosine deaminase (ADA) deficiency, X-linked severe combined immunodeficiency (X-SCID), and Wiskott-Aldrich syndrome (WAS). On the other hand, other types of monogenic disorders such as chronic granulomatous disease (CGD), introduced in this section, result in functional impairments that only affect mature cells.

1.4.4.1 Adenosine deaminase (ADA) deficiency and X-linked severe combined immunodeficiency (X-SCID)

The autosomal recessive inherited disorder ADA deficiency shows disturbance of leukocyte development because of dysfunction of the ADA enzyme. ADA is ubiquitously expressed and the impairment of the enzyme causes accumulation of toxic purine metabolites in the thymus. Before the HSC gene therapy was available, the only possible treatment was an injection of a recombinant enzyme which could lead to treatment failure when antibodies against the enzyme were produced (Chan et al., 2005). In 1990 the first gene therapy clinical trial was initiated by an autologous transplantation of *ex vivo* gene modified T-lymphocytes by GRV (Blaese et al., 1995). The two treated children had an improved number of T-lymphocytes and ADA enzyme activity; however, they could not cease enzyme replacement therapy. Later, HSC was efficiently transduced with GRV and long-term sustained engraftment was achieved (Aiuti et al., 2002). To date, more than 40 patients with ADA deficiency have been treated and more than 70 % of the patients are off enzyme replacement therapy. In this disease GRVs have not given rise to adverse events.

X-linked severe combined immunodeficiency (X-SCID), the most common form of SCID, has been another target since the early history of gene therapy.

X-SCID patients have dysfunction in the expression of the common gammachain of the IL-2 cytokine receptor that affects the development of T and natural killer (NK) cells and the function of B cells. In the late nineties, two GRV-mediated autologous HSC clinical trials were initiated (Gaspar et al., 2011) (Hacein-Bey-Abina et al., 2010). Over the follow-up period of nine years, the 20 treated patients with X-SCID reconstituted normal numbers of T cells and 50 % of the treated patients no longer needed immunoglobin replacement therapy. However, vector integration-related adverse events were detected in five of the children; therefore, this identified a risk in the use of GRV for gene transfer in this clinical setting. This risk of side-effects stimulated the study of other alternative vectors, LVs, which is described in 1.5.3.1.

1.4.4.2 Wiskott-Aldrich syndrome (WAS) clinical trial

In contrast to ADA deficiencies and X-SCID, Wiskott-Aldrich syndrome manifests further downstream in the haematopoietic cell lineage in several types of mature cells. WAS is a rare inherited immunodeficiency characterised by infections, microthrombocytopenia, eczema, autoimmunity, and lymphoid malignancies (Galy and Thrasher, 2011). The cause of this disease is loss-of function mutations in *WAS* gene encoding WASP that regulates the actin cytoskeleton in hematopoietic lineages and precursor cells (Thrasher and Burns, 2010) (Catucci et al., 2012). The mutations diminish WASP expression severely, which leads to abnormal thymopoiesis resulting in a low number of functional T cells. The first clinical trial for WAS was performed with a GRV (Klein et al., 2000) (Klein et al., 2003). The treated patients showed significant therapeutic benefit with reconstitution of B/T/NK cells and monocytes. However, patients developed acute leukaemia at a high rate (seven out of 10 patients) caused by upregulation of oncogenes via vector integrations (Braun et al., 2014). Because of the need for a novel vector to replace the GRV, the first LV-mediated clinical trial for WAS was conducted at two different clinical centres in London and Paris (Hacein-Bey Abina et al., 2015). $CD34^+$ HSCs transduction by a SIN LV carrying *WAS* cDNA whose expression is controlled by a 1.6 kb proximal WAS gene promoter (LV-w1.6W) was carried out (Charrier et al., 2007) (Scaramuzza et al.,

2013). Out of the seven treated patients, six patients had improved disease symptoms (one patient died of an opportunistic viral infection). Using the same LV, three WAS patients with slightly lower severity of the disease were treated by re-infusing autologous bone marrow-derived $CD34^+$ cells transduced with the LV-w1.6W (Aiuti et al., 2013). During the follow-up period of 20 to 32 months, all patients showed robustly multi lineage engraftment of gene-corrected cells in bone marrow and peripheral blood. All patients showed improved immune function, which resulted in amelioration of clinical manifestation of the disease. For the first time, the high-throughput integration site analysis was carried out in comparison with that of GRV in the same disease background (Boztug et al., 2010). The discovery of common insertional sites (CIS) of LV-w1.6W (*KDM2A*, *PACS1* and *TNRC6C*) may pose a concern in the biosafety although the typical GRV CIS was not observed in the *MDS1* and *EVI1* complex locus proteins EVI1 (*MECOM*) and *LMO2* that were reported in other clinical trials of PIDs.

There was another attempt to develop an LV vector for WAS with better performance than LV-w1.6W by optimising the promoter. As a candidate, viral-derived MND promoter (myeloproliferative sarcoma virus enhancer, negative control region deleted, dl587rev primer-binding site substituted) was selected. MND is a highly constitutively active promoter and resistant to transcriptional silencing (Challita et al., 1995) (Halene et al., 1999). In contrast to LV-w1.6W, higher expression of WASP was achieved by the MND promoter, which caused better rescue of the development of the MZ (marginal zone) B cell than LV-w1.6W (Astrakhan et al., 2012). However, one mouse developed clonal dominance by vector integration suggesting a potential risk of MND in an LV. However, the clinical trial for adrenoleukodystrophy (ALD) used the MND promoter in LV without any adverse events (Cartier et al., 2009); therefore, MND may not be the sole determinant for clonal dominance.

1.4.4.3 Chronic granulomatous disease (CGD) clinical trial

CGD is a rare inherited immunodeficiency characterised by the dysfunction of mature phagocytes (e.g. neutrophils, monocytes, macrophages, and

eosinophils) responsible for eradicating ingested microorganisms. Mutations in genes encoding subunits of the nicotinamide adenine dinucleotide phosphate (NADPH) oxidase complex (membrane-spanning subunits: gp91^{phox}, p22^{phox}, and cytosolic components: p47^{phox}, p67^{phox}, p40^{phox}) cause CGD (Segal et al., 2000) (Matute et al., 2009). 70 % of CGD patients are X-linked caused by mutations in the *CYBB* gene encoding catalytic subunit gp91^{Phox}, 30 % have defects in p47^{phox} (*NCF-1*), and only 5 % of the total patients have mutations in *CYBA* or *NCF-1* encoding p22^{phox} or p67^{phox}, respectively.

CGD is one of the most difficult diseases to treat by gene therapy, because therapeutic effects are not expected simply by correcting the expression of the defective gene (reviewed in (Grez et al., 2011)). Firstly, the transgene expression does not give a selective advantage to transduced cells; therefore, an intensive myeloablative conditioning is required. Secondly, correction in a large number of HSC is required owing to the short life-span of neutrophils that only last a few days. In the mid-1990s CGD clinical trials were initially attempted using GRVs at the National Institute of Health (NIH) (Malech et al., 1997). Five patients received an infusion of peripheral blood CD34^{+} cells transduced by p47^{phox}-expressed GRV without conditioning. However, the granulocytes with p47^{phox} expression were detected in the patients' peripheral blood at only a low level. Later, another clinical trial at NIH was carried out using a higher, but still not ablative, a dose of busulfan (10 mg/kg) prior to administrating CD34^{+} cells transduced by GRV, MFGS-gp91^{phox} (Kang et al., 2010). This improvement in conditioning led to improved initial gene-marking of the transduced cells. In one of the treated patients, the longer persistence of gene-marking was observed in neutrophils and monocytes for up to three years; however, the expression level was low in those transduced cells. The low survival rate of gene-corrected cells in these CGD trials could also be explained by the ectopic expression of the therapeutic genes. This leads to the production of a reactive oxygen species that may damage DNA after cell growth or induce apoptosis (Bedard and Krause, 2007) (Yahata et al., 2011).

Reflecting those treatment difficulties and the development of myelodysplasia by

GRV integration (more details in 1.5.3.2), nowadays LV is chosen for CGD treatment because of the highly efficient transduction of $CD34^+$ cells and the safety advantage over GRV. In order to overcome the off-target ectopic expression of the therapeutic gene and the subsequent reduction of gene-marked cells, a dual-regulated LV was developed (Brown et al., 2006). This LV encodes $gp91^{phox}$ whose expression is controlled by a myeloid-specific promoter and micro RNA 126 (miR126) for posttranscriptional regulation of $gp91^{phox}$ (Chiriaco et al., 2014). NOD/SCID and X-CGD mice xenograft model that received the transduced HSCs demonstrated restoration of NADPH activity by the development of stage-specific regulation of $gp91^{phox}$.

1.4.5 Lysosomal storage disorders

In this section, I introduce two inheritable neurodegenerative diseases: X-linked adrenoleukodystrophy (X-ALD) and metachromatic leukodystrophy (MLD). A patient with leukodystrophy shows abnormalities in white matter in the central nerve system (CNS) with or without the involvement of the peripheral nervous system (Vanderver et al., 2015).

1.4.5.1 X-linked Adrenoleukodystrophy (X-ALD) clinical trial

X-linked adrenoleukodystrophy (X-ALD) is a fatal demyelinating disease occurring at the frequency of 1:17,000 males (Biffi et al., 2011). X-ALD is caused by a deficiency in the *ABCD1* gene, encoding ALD protein that is an adenosine triphosphate-binding cassette transporter. This dysfunction causes accumulation of very long-chain fatty acids (VLCFAs) and the subsequent defective ß oxidation in peroxisomes disrupts myelin maintenance (Mosser et al., 1993) (Moser et al., 2007). ALD is a transmembrane peroxisomal protein; therefore, a cross-correction strategy using donor-derived cells to secrete the enzyme, as used for other storage disorders such as MLD, cannot be used. If the disease progression is at the early stage, disease progression can be stabilised by allotransplantation of HSCs if these are available.

The first beneficial autologous HSC gene therapy was carried out via LV expressing wild-type *ABCD1* cDNA (CG1711 hALD) under the control of MND (myelo-proliferative sarcoma virus enhancer, negative control region deleted, dl587rev primer binding site substituted) in 2009 (Cartier et al., 2009). In the preclinical experiments, murine Sca-1^+ cells used as a functionally equivalent of human $CD34^+$ cells were transduced with CG1711 hALD and transplanted into $ALD^{-/-}$ mice. Although transplanted ALD Sca-1^+ could replace 20 to 25 % of microglial cells, the neuropathological and clinical effects of the LV gene transfer were not evaluated because this mouse model does not develop cerebral demylination (Pujol et al., 2002). Therefore, the clinical trial was initiated without an accurate preclinical model for the outcome. However, after 24 to 30 months the two treated patients showed reconstitution of granulocytes, monocytes, and T and B cells expressing ALD protein. Integration sites were monitored over time and no obvious clonal skewing or dominance was observed. The patients also showed less progressive demyelination than historical controls.

1.4.5.2 Metachromatic leukodystrophy (MLD) clinical trial

Metachromatic leukodystrophy (MLD) is a neurodegenerative lysosomal storage disorder caused by a deficiency of arylsulfatase A (ARSA) because of mutations in *ARSA* (reviewed in (Fluharty, 2014)). ARSA deficiency is characterised by an accumulation of the enzyme substrate sulfatide in oligodendrocytes, microglia, and certain neurons of the central nervous system (CNS) and in Schwann cells and macrophages of the peripheral nervous systems (PNS). These accumulations lead to widespread demyelination and neurodegeneration leading to severe progressive motor and cognitive impairment. As a result, the affected children die within a few years. Diagnosis for MLD is based on measurement of the biological activities of the enzyme and/or characterisation of the gene mutations (Lorioli et al., 2014). The most common and severe form of MLD is the late infantile (LI) form (50 to 60 % of cases), with the onset of the disease within two years of life. As demonstrated in other gene therapy trials, allogenic hematopoietic stem cells transplantation (HSCT) is one therapeutic option for MLD, although the outcome seems only successful when treated at

the early stage of the disease course (Aubourg et al., 1990) (Shapiro et al., 2000). In the search for an alternative method for the treatment, the *Arsa*$^{-/-}$ mouse model demonstrated reconstitution of enzymatic activity of ARSA by a genetically modified HSC transplant (Biffi et al., 2004) (Biffi et al., 2006). The phenotype of *Arsa*$^{-/-}$ mice are mild, and not characterised by extensive demyelination, making it difficult to evaluate a clinical benefit; however, ARSA positive cells were successfully distributed in PNS as a possible source to provide the enzyme to the CNS.

The first successful *ex vivo* hematopoietic stem cell gene therapy for MLD was carried out using LV to treat three pre-symptomatic children, with one or more siblings, with LI-MLD (Biffi et al., 2013). Bone marrow derived CD34^{+} cells were transduced by LV carrying *ARSA* cDNA under the control of the human phosphoglycerate kinase (PGK) promoter. In order to validate a long-term therapeutic effect for this treatment, the treated patients were followed up for 18 to 24 months. During this time there was no abnormal cell expansion or clonal outgrowth caused by vector integration, which was profiled over time in different cell compartments such as progenitors, mature myeloid cells, and B/T lymphoid lineages. Although longer follow-up is required to measure safety and efficacy, this treatment had a successful outcome halting the disease progression of the demyelination in the patients' brain by providing a therapeutic level of ARSA enzyme.

1.4.6 ß-thalassaemia clinical trial

ß-thalassaemia (ß-TM) is a microcytic hemolytic anemia that is fatal in severe cases (ß-TM major) unless transfusions of life-long red blood cells and iron chelation are available (reviewed in (Finotti et al., 2015)). Based on the WHO estimation, about 1-5 % of the world's population might carry ß-TM and every year about 60,000 severely affected infants are born (Modell and Darlison, 2008). The imbalance of ß-globin and a-globin caused by mutations in the coding or control regions of the ß-globin gene results in ineffective erythropoiesis because the decreased amount of ß-globin produces free toxic a-globin in erythrocyte

cells that damages cell membranes. Among the severe ß-thalassaemia, $ß^{E}/ß^{0}$ is highly frequent. In this case, the one ß-globin allele is completely silent ($ß^{0}$), while the other ($ß^{E}$) has a missense mutation resulting in the amino-acid substitution of glutamic acid to lysine at position 26 of the ß-globin chain (GLU26LYS). Approximately 50 % of the ß-TM patients who have this combination of alleles require transfusions.

GRV was initially developed as a model vector for ß-TM. The introduction of the locus control region (LCR) upstream from the β-globin gene cluster could achieve successful gene expression in a site-independent manner in erythroid cells (Grosveld et al., 1987). In human β-globin gene locus, there are four DNase I hypersensitive sites (HS) distributed over 20 kb upstream from it and the length of the HSs is tailored, leaving the core element (Chung et al., 1997). However, these elements led to instability of the GRV vector with low titre (Emery et al., 1998). In contrast to the monogenic disorders described above, overall the GRV-mediated ß-globin gene transfer had limited success (Ellis and Pannell, 2001). Therefore, LVs were then engineered because of a superior vector property to GRVs, such as vector transduction in non-dividing cells (Naldini et al., 1996) and well-characterised nucleus-to-cytoplasmic export of unspliced mRNA that facilitated the inclusion of globin genes with intact introns, LCR fragments, and insulator elements (May et al., 2000).

The first clinical trial was carried out treating two patients with severe $ß^{E}/ß^{0}$-TM (Cavazzana-Calvo et al., 2010). Bone marrow $CD34^{+}$ cells transduced by the ß-$globin^{A\text{-}T87Q}$ vector (LentiGlobin® HPV569) were engrafted in the patients. This ß-$globin^{T87Q}$ works as a distinguishable protein marker against transfusion-derived ß-globin (Pawliuk et al., 2001). Although the first patient failed in the engraftments after full myeloablation, the second patient achieved transfusion-independence at 12-months after the treatment. In the latter patient, clonal expansion of erythrocytes by LV integration was observed (more details in the 1.5 insertional mutagenesis section); however, the transfusion -independence is still maintained after more than seven years with a decreased percentage of the dominant clone. A different ß-TM clinical trial was performed

using LV with a CMV promoter in place of the U3 promoter/enhancer in the 5' LTR (LentiGlobin® BB305). In contrast to HPV569, this vector does not have a cHS4 insulator in the 3' LTR. Two patients with ß^{E}/ß^{0}-TM were treated and both became successfully transfusion independent at 3.5 and 6.5 months after the treatment, respectively. It is assumed that the higher vector copy number in CD34^{+} HSCs may contribute to this better outcome. Based on the successful outcome, three additional ß^{E}/ß^{0}-TM patients were enrolled in the U.S. study and the treatment is ongoing (Negre et al., 2016).

1.4.7 Gene therapy for infectious diseases

Acquired immune deficiency syndrome (AIDS) is caused by infection with human immunodeficiency virus (HIV). The first recognition of the AIDS epidemic was in 1981 and still 37 million people are suffering from AIDS worldwide (http://www.who.int/hiv/en/) (Herrera-Carrillo and Berkhout, 2015). The majority of AIDS cases are caused by HIV-1 infection; however, some AIDS cases in West Africa are caused by HIV-2 infection (Nyamweya et al., 2013). The clinical trials for AIDS in this thesis deal with the treatment of HIV-1 infection. Immune function in a patient becomes impaired when the disease progresses because the number of CD4^{+} T cells is reduced. As a treatment, patients are assigned to life-long combinational antiviral therapy (cART) that significantly prolongs the patients' life-span; however, there are some limitations to be overcome, such as the incomplete elimination of the virus reservoir and the cytotoxicity by long-term cART.

The concept of the anti-HIV gene therapy is to interfere with crucial steps of the viral replication cycle by targeting viral protein or host cellular factors that are required for viral replication. The therapeutic strategy can be divided into two groups, RNA or protein-based therapies (Herrera-Carrillo and Berkhout, 2015). Autologous T cells, or CD34^{+} cells, have been engineered to express a surface fusion peptide, a mutant Rev molecule, or a ribozyme targeting viral *vpr*/*tat*. Safety and tolerance of these approaches have been demonstrated, however, the expression of therapeutic transcripts was not sustained.

The first LV-mediated clinical trial for AIDS was performed in five patients (who had failed at least two antiretroviral regimens previously) (Levine et al., 2006). Those patients were treated by a single infusion of $CD4^+$ T cells transduced with HIV derived vector (VRX496) carrying intact LTRs and a 937-base antisense sequence against HIV *env*. This antisense gene is expressed by the transcription initiation from the vector 5' LTR; therefore, tat expression is required which is only supplied by wild type HIV-1 in HIV-infected individuals. After the infusion of the transduced $CD4^+$ T cells, one subject had declined viral loads and four subjects had increased $CD4^+$ T cell counts. Three-year monitoring of the patients reported no adverse events. The same research group performed multiple infusions (three to six) of VRX496-Tcells in 17 HIV patients to test if the therapeutic efficacy is increased by multiple infusions (Tebas et al., 2013). Although efficacy was not enhanced by multiple infusions, in the plasma sample of eight patients A to G substitutions in HIV *env* sequence were observed as a cause of replication impairment of HIV. Integration site analysis also showed no evidence of enrichment of integration sites near cancer-associated genes (e.g. *LMO2*, *CCND2*, *SPAG6/BMI1*, and *EVI1*), suggesting this LV-mediated treatment protocol has good biosafety.

The CCR5 HIV co-receptor is a molecular target candidate to anti-HIV therapy by preventing the virus from entering cells. One intriguing case is known as "Berlin patient" (Hutter et al., 2009) (Hutter and Thiel, 2011). The patient who received a bone marrow transplant from a CCR5-Δ32 homozygous donor became free of detectable HIV load. CCR5-Δ32 is a natural mutation and the transplanted cells did not allow HIV cell entry because CCR5 is not expressed (Novembre et al., 2005). Therefore, HIV gene therapy targeting CCR5 became a promising strategy. Gene editing to target CCR5 by zinc finger nucleases (ZFNs) was also reported with its successful outcome including its biosafety (Tebas et al., 2014). 12 patients with chronic aviremic HIV infection were enrolled and treated by autologous $CD4^+$ T cell transplantation. Over the follow-up period, without the disruption of antiretroviral therapy (36-week), CCR-modified cells persisted; thereby the HIV load in the blood was controlled in most of the patients according to the DNA levels of HIV. Based on those results by

transplantation of a single dose of the modified CD4 cells, a trial with the increased dose will be the next step.

Another molecule candidate for anti-HIV therapy is a specific short hairpin RNA to CCR5 co-receptor and cell-membrane-anchored CD46 to inhibit virus fusion to the host cell membrane. The dual combination LV (LVsh5/CD46) that carries those elements was tested in the humanised bone marrow, liver, thymus (BLT), and mouse model (Burke et al., 2015). As previously shown (Wolstein et al., 2014), this vector could prevent both R5 and X4-tropic HIV-1 infection in comparison to an untransduced control. Stable transgene expression was also observed up to six months. Phase I/II clinical trial to test the feasibility of this method using $CD34^+$ HSPC and $CD4^+$ T cells transduced by LVsh5/C46 is underway.

1.4.8 T cell immune therapy

Beyond the rare genetic disorders, gene therapy became one option to treat cancer by immune-based strategies via the expression of engineered receptors on the surface of patients' T cells (reviewed in (Maus et al., 2014) (Naldini, 2015)). This exploits the cancer-specific antigen or antiviral molecule to boost the adaptive immune response and eradicate tumours. In early studies, exogenous TCR expression reported a partial effect on tumours and off-target expression that damaged tissues (Robbins et al., 2015). In recent studies, synthetic chimeric antigen receptors (CARs) have been used. These synthetic molecules are composed of a single chain antibody (scFv) composed of heavy and light chain variable regions with a synthetic linker expressed on a transmembrane region and intracellular domains from T cell signalling molecules. Expression of CARs on patients' T cells can circumvent tolerance of the T-cell repertoire and the requirement for MHC presentation (Maus et al., 2014). GRVs have been mainly used to introduce CAR into T cells because they can permanently and stably transduce T cells and safety is established from the integration standpoint in primary T cells (Scholler et al., 2012). LVs have also been used more recently. Hematologic malignancies are the first target of CAR T

cells because antigen receptors to be targeted on the tumours are well characterised. In addition, a relatively easy sampling of tumors and T-cells homing to hematologic organs such as blood, bone marrow, and lymph nodes are also the determinant of the choice of cancer. To date, CAR T cells were applied in B-cell malignancies by targeting CD19, CD20 or IgGκ, or acute myeloid leukaemia (AML) by targeting CD33 and Lewis-Y antigen. In addition, CD30 has also been investigated for Hodgkin disease treatment (Shakoor et al., 2002).

1.4.9 Parkinson's disease clinical trials

Parkinson's disease (PD) is a neurodegenerative disease caused by insufficient production of dopamine in substantia nigra pars compacta (within midbrain) (reviewed in (Coune et al.)). Dopamine is a neurotransmitter that helps regulate body movements and other motor/cognitive functions. Five million people are estimated to be in the effect of PD worldwide with the prevalence of around 1 % in people aged 60 years (de Lau and Breteler, 2006). Oral administration of dopamine precursor (levodopa) is effective, but only to the patients at initial stages of the disease.

AAVs have been mainly exploited to treat PD patients with advanced stage (LeWitt et al., 2011). Although the clinical trials demonstrated successful neuroprotection in such patients, AAV can accommodate up to 4.7 kb of an exogenous gene and this cloning capacity restricts the number of genes transferred into patients. So treatment with a non-primate LV based on the equine infectious anemia virus (EIAV) genome has also been studied (Mitrophanous et al., 1999) (Azzouz et al., 2002). The EIAV-derived vector ProSavin was tested in phase I/II clinical trial to treat PD patients in the advanced stage (Palfi et al., 2014). ProSavin carries three genes essential to synthesise dopamine (aminoacid decarboxylase (AADL), tyrosine hydroxylase (TH) and cyclohydrolase 1 (CH1)) in the SIN EIAV vector. During the 12-month follow-up, ProSavin was well tolerated in patients and all 15 treated patients showed an improvement in motor behavior. Importantly, no serious adverse

events were observed, suggesting the EIAV-based vector is promising for the treatment of PD.

1.5. Insertional mutagenesis

1.5.1 Insertional mutagenesis: brief introduction

Insertional mutagenesis (IM) is dysregulation of host cellular gene expression caused by retrovirus insertions into the host genome and can lead to clonal expansion or malignancies (reviewed in (Uren et al., 2005), (Knight et al., 2013)). IM was exploited in the forward genetic screening of novel oncogenes by slow transforming retroviruses that do not encode viral oncogenes so that tumour formation can be assessed by integrations. The identified oncogenes were then examined by overexpression in cell culture and lab animals to characterise the function and understand the molecular mechanisms of oncogenesis. In contrast to the beneficial aspect of IM, it can cause unwanted side effects such as leukaemia in treated patients at retrovirus vector-mediated gene therapy. Despite a positive effect by vector transduction at the early point of gene therapy, the follow-up study reported IM as a side effect of retroviral vector-derived gene therapy and led to life-threatening leukaemia in some patients. In such cases, the molecular mechanism of related genes affected by vector integration was studied by vector integration site identification and fusion mRNA expression. Furthermore, to prevent the IM from occurring in clinical trials, the methods for safety screening of tested vectors were developed.

In this section, firstly the accepted molecular mechanisms of IM are described for a better understanding of how the proviral sequence dysregulates neighbouring genes. Secondly, clonal expansion or malignancies reported in gene therapy clinical trials to date are described. Together with the previous section (1.4 Gene therapy), those adverse events are important as knowledge because a newly designed vector with superior biosafety to GRVs can be invented based on them. In addition, the mechanism of gene dysregulation by vector integration is crucial to discover the new function of genes. Therefore the methods to study IM are described including integration site and chimeric transcript identification. Finally, LV IM-based forward genetics is introduced as a useful tool to identify novel genes resistant to cancer treatment or viral latency.

1.5.2.Molecular mechanisms of retroviral IM

1.5.2.1 Activation of cellular gene transcription via viral elements in LTR

Retroviral LTRs function for effective transcription of the viral genome, however especially the enhancer and promoter elements contribute to dysregulate host cellular gene transcription. Enhancer sequences in the U3 of retroviral LTR augment the activity of cellular promoters, sometimes over large distances, resulting in the up-regulation of targeted oncogenes (Cavalli and Misteli, 2013). Such enhancer mutation is found predominantly upstream of the affected gene in the antisense orientation or downstream in the sense orientation (Uren et al., 2005). The reported malignancies at clinical trials by integrations of MLV-based GRV were also enhancer-driven because early tested GRV had intact viral LTRs (Hacein-Bey-Abina et al., 2008) (Howe et al., 2008) (Braun et al., 2014) (Stein et al., 2010).

The promoter element in the 5' LTR is another contributor to dysregulate host transcription by controlling transcription initiation within the 5' LTR and leads to the formation of chimeric transcripts with host sequences by read-through (Bokhoven et al., 2009) (Cavazzana-Calvo et al., 2010). This type of IM pattern can be achieved by the sense integration to host sequence. The chimeric transcripts have truncation at the 5' upstream part of the gene and splice sites within a vector.

Polyadenylation signal within LTRs can disturb cellular transcription and lead to cancer formation. Such mRNAs generate truncated protein that can cause deleterious gain-of-function or dominant negative activity (Chang et al., 2007). For instance, proviral insertion in the 3' untranslated region (UTR) of host mRNA removes regulatory elements, or destabilises motifs in it, by the sense polyadenylation signal in 3' LTR and the antisense cryptic polyadenylation signal in the enhancer of 5' LTR. As such, the IM example, the sense integration of MoMLV downstream of proto-oncogene serine/threonine-protein kinase Pim-1 (*pim-1*), caused T cell lymphomas in mice (van Lohuizen et al., 1989).

Prematurely terminated *pim-1* mRNA resulted in the truncation of the element of 3' terminal exon of the gene that reduces the stability of the mRNA, therefore the expression of *pim-1* was upregulated. Besides, the *pim-1* locus is also known as the most frequently targeted integration site by MoMLV (Bachmann and Moroy, 2005). Similarly, activation of *N-Myc* was observed in tumours by truncation of 3' of the mRNA by MLV integration in the 3' UTR and premature polyadenylation at the 5' LTR.

1.5.2.2 Aberrant splicing

The presence of a viral sequence in the genome by integration can generate chimeric mRNAs via known vectors or cryptic splice sites and the affected host sequence may be disrupted aberrantly. The disruption of the original host sequencing is called aberrant splicing. This can remove regulatory regions of the intact mRNA and contribute to the cancerous transformation of transduced cells by the up-regulation of protein expression. This type of IM was observed in cell culture, targeting the *Ghr* locus *in vitro* (Bokhoven et al., 2009) and in *HMGA2* locus *in vivo* (Cavazzana-Calvo et al., 2010). In both cases, the fusion transcripts contributed to the up-regulation of the gene product and led to clonal expansions. The initiation of fusion mRNAs, whether vector or host sequence is first, makes a difference in the splicing pattern which changes the use of the splice site between the vector and host sequence. For instance, the *Ghr* locus had integration of GRV with intact LTR in the same orientation with the host transcript. Transcription was initiated within the 5' LTR and spliced to the SA of *Ghr* exon 2. This pattern of aberrant splicing is splice-out. While SIN LV integration in *HMGA2* also had integration in the same direction, transcription started from the host sequence because the vector has a deletion in the U3 of LTRs. Therefore *HMGA2* exon 3 was spliced into a vector sequence within the 5' U3. This splicing pattern is called "splice-in".

1.5.2.3 Inactivation of tumour suppressor genes

Oncogene activation is also caused by inactivation of tumour suppressor genes.

As to tumorigenesis by inactivation of tumour suppressor genes, Knudoson's "two-hit" hypothesis is well known (Knudson, 1971). This theory is that a gene on both alleles in the diploid genome needs to be mutated for cancer formation. However, it was proved that the "two-hit" is not necessarily a pre-requisite. Instead, in some cases, gene mutation in one allele is sufficient to reduce the level of gene expression to inactive the gene. This is called haplo-insufficiency of tumour suppressor genes. For instance, the comparison study between p53 deficient heterozygous ($p53^{+/-}$) mice and wild type mice also demonstrated that the reduced amount of p53 could cause cancerous transformation of cells (Venkatachalam et al., 1998). Vector integration can be the cause of haplo-insufficiency and the subsequent cell transformation. In a mouse BM transplant model, SIN-LV integration in *Ebf1* caused acute B-lymphoblastic leukaemia (B-ALL) (Heckl et al., 2012). This integration into intron 8 of *Ebf1* resulted in initiating the transcript within the vector and truncation of exon 1 to 8 of this gene product, which led to the downregulation of overall *Ebf1*. The reduction of *Ebf1* expression contributed to strong STAT5 upregulation and caused the B-ALL. Another example was transduction of LV.SF.LTR (HIV-derived LV with spleen focus-forming virus (SFFV) promoter insertion in the U3 region) in *in vivo* genotoxicity assay using tumour prone $Cdkn2a^{-/-}$ murine hematopoietic stem/progenitor cells (Montini et al., 2009). In a myeloid and a lymphoid tumour generated by the same LV, *Nsd1* was downregulation by about 40 % by integration in the intron 5 and 6 in the opposite direction to the transcriptional direction, but the detailed mechanism by which the vector integration reduces the gene product is not described. Importantly, haplo-insufficiency of *NSD1* is the major cause of developmental disorder Sotos syndrome and is associated with malignancies (Martinez-Glez and Lapunzina, 2007).

1.5.3 Genotoxic events in clinical trials via vector integration

1.5.3.1 Acute lymphoblastic leukaemia in SCID and WAS clinical trials by vector-mediated IM

The first report of adverse events in a gene therapy clinical trial was the SCID-X1 trials in Paris and London using GRV with intact LTR (Hacein-Bey-Abina et al., 2008; Howe et al., 2008). Notwithstanding the successful immune reconstitution by the gene therapy, five of the total treated 20 patients developed T cell leukaemia two to 5.5 years after the treatment. When the expanded leukemic clones were examined, integration sites of GRV were identified in oncogenes *LMO2* or *CCND2* and those genes were up-regulated by the enhancer-mediated IM mechanism. In addition, other genetic abnormalities were observed possibly contributing to the leukaemogenesis, such as *SIL-TAL1* fusion transcripts, a deletion of *CDKN2A* tumour suppressor gene, a gain of function mutations in *NOTCH1,* or a translocation of the TCR-ß region of the *SIL-TAL1* locus. In the follow-up study, longitudinal pyrosequencing analysis was carried out using eight patients from the Paris SCID-X1 trial (Wang et al., 2010). Interestingly, four patients (two patients (P1, P2) without leukaemia development) had integration sites near *CCND2*. This resulted in a clonal expansion in those two healthy patients, suggesting that this integration event does not inevitably lead to leukaemia. In the same study integrations near *HMGA2* were observed in P1 and P7. Vector integration was found between exon 3 and 4 of *HMGA2* which generated host cell-virus fusion mRNA with truncation of exon 4 and 5. The lack of let-7 microRNA binding sites, the negative regulator of this gene coded in exon 5, stabilised the gene product and upregulated it. The same mechanism of IM via integration into *HMGA2* was also observed in the LV-mediated clinical trial of ß-TM (Cavazzana-Calvo et al., 2010)

In WAS clinical trials, a first phase I/II study was conducted with 10 patients. They were treated by GRV with intact LTR that drives WASP expression (Boztug et al., 2010) (Braun et al., 2014). Nine out of 10 patients who received the HSCT showed successful engraftments of transduced autologous HSCs and improved clinical symptoms by sustained expression of WASP. Nevertheless, seven patients developed acute leukaemia. Although the comprehensive analysis of the latest patient is currently ongoing, GRV integration was commonly found within or near *LMO2* in the six patients, *MDS1* for two patients or *MN1* for one patient. High throughput integration site identification without a restriction

enzyme (nrLMA-PCR) identified the four most targeted gene loci (*MDS1-EVI1*, *PRDM16*, *LMO2*, and *CCND2*) in all treated patients. These sites were observed in the clonal expiation of other GT clinical trials.

1.5.3.2 Myelodysplasia in CGD clinical trial by vector-mediated IM

Two adult patients treated with nonmyeloablative bone marrow conditioning received an infusion of peripheral blood $CD34^+$ cells transduced by a GRV carrying $gp91^{phox}$ (SF71$gp91^{phox}$) whose transcription was driven by SFFV-LTR (Ott et al., 2006). Treated patients had reconstitution of NADPH oxidase function in peripheral blood leukocyte and granulocyte and had improvement of antibacterial activity. Three months after the treatment, clones harboring integrations in *MDS1-EVI1*, *PRDM16,* or *SETBP1* were expanded in both patients and an elevated level of *MDS1-EVI1* expression was found in bone marrow cells and peripheral blood leucocytes. This clonal expansion was regarded as the contributor for the treatment at the time. However, longer follow-up of those patients showed the development of bone marrow disorders, myelodysplasia, owing to the increased expression of fusion transcripts *MDS1-EVI1* and *EVI1* via GRV integrations (Stein et al., 2010). Chromosomal aberration (monosomy 7; a loss of a copy of chromosome 7), possibly associated with clonal dominance, was also found in both patients. They also found that *EVI1* overexpression was dominantly contributing to centrosomal aberrations chromosomal instability. This study indicated that clonal expansion can lead to malignant myelodysplasia, thus long-term follow-up on treated patients is required to test if multi lineage populations have remained in patients' bodies.

1.5.3.3 Clonal expansion caused by virus integration: clonal dominance in ß-thalassaemia clinical trial by vector-mediated IM

In the case of ß-TM, clinical trials have been mainly performed with LVs and there are no reported malignancies to date. However, clonal dominance was reported as a first potential cancerous risk of using LVs for ß-thalassaemia

(Cavazzana-Calvo et al., 2010). In this trial, patients were infused with bone marrow $CD34^+$ cells, transduced by SIN lentiviral vector encoding ß-globin cDNA, driven by the human ß-globin locus control region with chicken hypersensitive site 4 (cHS4) insulators in the deleted 3' U3. In the dominant clones, a single vector integration in the same orientation as *HMGA2* was detected and the integrated vector lost one of the two copies of cHS4 insulators. The mechanism of the up-regulation of the gene was due to loss of the negative control region (let-7 microRNA binding sites) in *HMGA2*-vector transcripts by exon-skipping. The follow-up study reported the disappearance of the dominant clone and the clonal expansion did not cause malignancy (Leboulch, 2013).

1.5.4 Studies to understand the mechanisms of IM

1.5.4.1 Cell line-based assay

Murine cells are a robust tool to investigate IM mechanisms. The reason why mouse cells are chosen over human cells is that they can be immortalised at a higher frequency owing to a higher functional activity of telomerase (Rangarajan and Weinberg, 2003). To isolate mutants and assess the IM mechanisms, murine cytokine dependent cell lines have been exploited for a fast and quantitative tool. In our lab, mutant isolation was carried out by IL-3 dependent Baf3 (Palacios and Steinmetz, 1985) and its derivative clone with Bcl-2 overexpression Bcl15 cells (Collins et al., 1992) by RV integration. Both cell lines are IL-3 dependent and mutant isolation was successfully performed under the IL-3 starvation. With regard to the robustness of cells, Bcl15 cells are more advantageous than the parental Baf3 because Bcl-2 overexpression makes Bcl15 more resistant to apoptosis and a higher number of mutant recovery is expected in the protocol. Using those cell lines, our group established the *in vitro* quantitative IM assay and we studied the molecular mechanism by which isolated mutants gained growth advantage in an aspect of vector integration (Bokhoven et al., 2009). The vector integration sites were assessed by ligation-mediated PCR (LM-PCR) or inverse PCR and the expression of chimeric transcripts between cellular and viral sequences was confirmed by qRT-PCR

and 5’ RACE. This assay system discovered that LV integration-derived mutants had integration in *Ghr* locus and the virus-host fusion mRNAs from vector SD to the host cellular SA at *Ghr* exon 2. In contrast, GRV-derived mutants were generated by integration into retroviral vector common integration sites (CIS) and upregulation of oncogenes listed in RTCGD. In a later study from our team, the safety of the vector component was also examined in this assay system by comparing the number of generated mutants by LV integration (Knight et al., 2010). All used LV vectors have intact LTR. The tested vector components include SFFV, ubiquitin (UBIQ) promoter, and an SFFV vector counterpart without WPRE element at the 3’ side of the vector. The LV with SFFV promoter generated mutants at a high rate and the deletion of WPRE did not affect mutant generation. Intriguingly, there were no mutants by UBIQ vector transduction. These results suggested that SFFV promoter could increase or enhance transcription initiation from 5’ LTR. qRT-PCR demonstrated transcription initiation from 5’ LTR of the vector but the comparison date for the UBIQ vector is required to confirm this hypothesis. Furthermore, in the search of a promoter replaceable to a commonly used promoter such as CMV or SFFV promoter, Ubiquitous Chromatin-Opening Element (UCOE) element was raised as a candidate because of the resistance to gene silencing in a position independent manner (Antoniou et al., 2003). Therefore, UCOE element within LV was optimised to achieve the safest candidate with the low mutant formation (Knight et al., 2012). In IM assay, splice sites in the original UCOE sequence caused aberrant splicing. The vector candidate MA1082 and its derivatives that have mutations at both known and cryptic splice sites could abrogate aberrant splicing and result in a reduced number of mutants. These studies demonstrated that possible IM mechanisms of LVs can be assessed in the well-established murine cytokine independent cell-based assay and this assay can be further used to optimise vector safety.

In another study, to address the quantitative measurement of IM by mutant generation and molecular mechanism of IM in a relevant cell type, a protocol for re-plating cell assay was developed using lineage-negative (Lin^-) bone marrow (BM) cells (Modlich et al., 2006). In the protocol, Lin^- BM cells were derived from

C57B16/J mice and transduced by test vectors twice. The transduced cells were then expanded for two weeks and cells were re-plated at a density of 100 cell/well in 96-well plates and cultured for another two weeks. Based on the number of wells with proliferated clones, the frequency of mutagenesis was calculated. To validate the mechanism of the growth advantage of the transformed cells, LM-PCR and linear amplification–mediated PCR (LAM-PCR) were exploited. Using this model, vector safety was compared between SIN GRV and wild-type (WT) GRV (with intact LTR), showing a reduction in transformation risk by SIN GRV (Modlich et al., 2006). In expanded clones both vectors had integration in the *EVI1* locus in reverse-orientation relative to the *Evi1* transcription direction. This suggests that the clonal expansion by upregulation of *EVI1* was enhancer-mediated by WT GRV and was possibly aberrant-splicing by SIN GRV. Importantly, those results can draw the conclusion that enhancer/promoter-mediated IM has a stronger affect for mutant generation than splicing-mediated IM by SIN GRV. Later, from the same research group, LVs with different physiological promoters or with varieties of deletion in the U3 of the LTR were tested in the same assay system (Modlich et al., 2009). It revealed that SIN LVs with physiological promoters have a lower risk of cell transformation compared to SIN LV with SFFV promoter, suggesting the choice of promoter is important for the design of a safer vector.

1.5.4.2 Murine models

As a pre-clinical assessment of HSC gene therapy, the murine model has been used to validate the feasibility of the tested vector and find the potential risk of IM mechanisms via vector integrations (Montini et al., 2009). Mutagenicity of tested vectors was measured by the occurrence of clonal expansion or malignancies of transduced cells in transplanted mice. One example is the transplant of the transduced Lin^- cell derived from the BM of tumour prone *Cdkn2a*$^{-/-}$ mice by retroviral vectors (GRVs and LVs) with different designs into lethally irradiated wild-type FVB mice. Retroviral vectors with SFFV enhancer/promoter region in the LTR or SIN vectors with an SFFV internal promoter were assessed for genotoxic potential. This study showed that an active LTR is the major

determinant of genotoxicity in the recipient mice and LV with SFFV LTR required a 10-fold higher integration load to obtain the same genotoxic risk than GRV with SFFV LTR. The mechanism of IM by LV with SFFV LTR was examined by a further transplant of tumours generated by the vector into secondary recipient mice to obtain biological replicates. One transplanted myeloid tumour showed LV-*Braf* chimeric transcripts. The fusion transcripts were initiated from an SD in the 5' LTR and spliced out to the SA in the exon 13 of *Braf* (splice-out). The transcripts generated a truncated Braf molecule, which led to constitutive kinase activity. In addition, one myeloid and one lymphoid tumour showed *NSD1* haploin-sufficiency by the reduction of the product by about 40 %.

In contrast to cell-based assay, the mouse-based IM model is costly and time-consuming. However, the pre-clinical vector should be actually tested *in vivo* to confirm that the vector can work in the same manner as *in vitro* and no IM events are detected. The combinational use of both assay systems would not only assess the vector precisely and rapidly but also contribute to reducing the number of experimental animals.

1.5.4.3 Analysis of vector integration loci and chimeric transcripts between host and vector sequences

When the cause of vector-mediated IM is studied, firstly vector integration sites are identified, if such sites have hits in RTCGD, so that potential integration sites that could be the cause of IM would be narrowed down. One of the common methods uses restriction enzyme digestion such as inverse PCR (Carteau et al.) or LM-PCR (Wu et al., 2003). Digested gDNA is ligated with linker oligonucleotides, which allows us to obtain the host genome sequence next to the vector sequence using vector- and linker-specific primers. These techniques are sufficient to analyse monoclonal samples. However, to characterise complex samples with multiple target sequences another sensitive method, LAM-PCR (Schmidt et al., 2007), was developed. Prior to restriction enzyme digestion, the target vector-host junctions are amplified using 5' biotinylated primers, allowing for immobilisation on the solid phase and magnetic selection. The amplified DNA

is then digested by four-cutter enzymes and ligated to linker oligonucleotides. However, some studies argue that the use of a restriction enzyme creates a bias of retrieving integration sites (Harkey et al., 2007) (Wang et al., 2008b). Alternatively, nrLAM-PCR was invented (Paruzynski et al., 2010), which does not use restriction enzyme digestion. Nowadays it is increasingly common to combine LAM/nrLAM-PCR with next-generation sequencing technology to obtain a large number of sequence reads and analyse quantitatively.

Regarding the analysis of host-vector chimeric transcripts, rapid amplification of cDNA ends (RACE) is exploited for cDNA-based vector-host junction amplification of 5' or 3' mRNA (Frohman et al., 1988). Ligation of linker oligonucleotide to 3' cDNA or priming reverse transcription by oligo dT primers allows reading of the unknown host cDNA region fused to the vector sequence. The cDNA is ligated in a circular form, or concatemers, and the vector host boundary sequence is read by vector- and the linker/oligo dT specific primers. Another method recently proposed is cLAM-PCR (Cesana et al., 2012) whose protocol is based on LAM-PCR (Schmidt et al., 2007). mRNA is reverse transcribed by oligo dT primers and a biotinylated primer is used for linear PCR based on cDNA. Magnetic selection and digestion steps are the same in LAM-PCR. The shotgun cloned products are sequenced by next generation sequencing.

1.5.5 LV integration profile as solution for the treatment of cancer or infectious diseases

We highlighted retroviral vector-mediated IM as an adverse event in the previous section. However, IM can provide the genes of altered expression by integration as beneficial information to discover the strategies to treat cancer (Ranzani et al., 2014) or infectious diseases like AIDS (Maldarelli et al., 2014) (Wagner et al., 2014).

1.5.5.1 Cancer-related gene discovery

LV-mediated IM was firstly applied to the oncogene screen in hepatocyte as the model of liver cancer (Ranzani et al., 2013). They established the IM model to induce human hepatocellular carcinomas (HCC) in three different mouse models: tumour prone *Cdkn2*$^{-/-}$*ifnar1*$^{-/-}$ mice with higher permissiveness of transduction, *Pten* liver-null mice as a model of chronic liver oxidative damage, and wild type mice with CCl_4 treatment as a model of chronic liver injury. The tested LV in this model has enhanced transthyretin, hepato-specific enhancer-promoter sequences in the LTR, to induce tumours in the mice. Integration site analysis of cancer tissues by LAM-PCR found that *Fign*, *Braf, Sos1,* and *Dlk1-Dio3* were highly targeted integration sites whose expression was upregulated. All those loci had vector integration in the same direction as transcriptional direction. In the analysis of chimeric transcripts by RT-PCR, transcriptions start from the 5' LTR and are spliced in from the vector HIV major splice donor (SD) to the acceptor (SA) in each gene. In the case of *Fign*, *Braf,* and *Sos1*, the lack of N-terminal domain by vector integration contributed to the gene upregulation. Integration of the *Dlk1-Dio3* generated fusion transcript with full-length *Rtl1* caused *Rtl1* upregulation.

LV IM was also exploited in a model of breast cancer. Using two subtypes of $HER2^+$ breast cancer cell lines (SKBR3 and BT474), LV integration sites contributed to discover the genes involved in resistance to anticancer therapy (Ranzani et al., 2014). LV with SFFV enhancer/promoter in the LTR was transduced, followed by lapatinib (a tyrosine kinase inhibitor) selection. Cells with growth advantage gained by vector transduction can survive this drug selection step. The amplified vector-host sequence boundaries by LAM-PCR were sequenced by next generation sequence and *PIK3CA* and *PIK3CB* were identified as CIS. *PIK3CA* had integration in the first intron of the gene in the sense strand. *PIK3CB* had integration upstream of the gene in the sense integration by 87 %. The vector promoter-mediated IM overexpressed those genes; overexpression of those genes by SIN LV conferred the resistance to lapatinib. In the same study, they attempted to apply the same LV IM platform to discover novel drug-resistance genes in the context of pancreatic cancer. The LV integration in *SOS1* was isolated as a drug resistance gene to erlotinib

(clinically used EGFR inhibitor) in the pancreatic cell line. The overexpression of this gene was achieved in the integration of the intron 8 of the gene in the sense orientation, which led to truncation of the gene product.

1.5.5.2 HIV latency and the relation to virus integration sites

The HIV latency in AIDS patients is a barrier to eradicate infected cells under combinational antiretroviral therapy (ART). HIV in the latent stage is not affected by the ART and, once the treatment is stopped, the suppressed virus loads easily to increase to the level before the ATR is initiated because of the virus production from latent reservoirs (Weinberger and Weinberger, 2013). In order to understand the mechanisms of virus latency in regard to targeted gene locus within the infected cells, HIV integration sites in cells from AIDS patients under the ART were studied (Maldarelli et al., 2014) (Wagner et al., 2014). To test whether identical sequences are found in the HIV genome, single genome sequencing was performed (Kearney et al., 2008). This particular sequencing method permits individual cDNA molecules of a defined portion of the genome to be PCR amplified. As a result of genome analysis of HIV infected individuals, interestingly, clonally expanded cells were observed in plasma or peripheral BM cells of the five tested patients (Maldarelli et al., 2014). Integration analysis of $CD4^+$ T cells was performed in one patient who underwent 11.4 years of combinational ART prior to the cell collection. This revealed the characteristic virus integration in the intron 4 or 5 of *Bach2* (Liu et al., 2009) and in the intron 4 of *MKL2* (Ma et al., 2001) in the sense orientation to the transcriptional direction. HIV promoter-mediated IM possibly dysregulated those gene expressions and contributed to clonal expansion. Separately, they performed HIV integration site analysis in HeLa cells and $CD34^+$ cells. Because HIV did not have a preference of integration into *Bach2* and *MKL2*, those sites in the expanded clones were selected overtime. These novel findings revealed that HIV wisely uses gene loci that could be related to cell survival by disrupting the gene transcripts by integration. Overall, the cells with such integration can expand and persist without being affected by ART, and therefore contribute to achieving latency in patients.

1.6. Splicing

In my PhD project, we investigated the possible mechanism of LV-mediated insertional mutagenesis. Our focus is to look into host-vector fusion mRNAs expressed in LV-transduced cells that were isolated in an IM assay. The formation of such chimeric mRNAs is mediated by splicing. Therefore, the basic knowledge of splicing, the essential step for producing mRNA for gene expression, is crucial for understanding and estimating potential splicing patterns of host-vector chimeric mRNAs. For the formation of such various transcripts, use of different splice sites is related and the generation of multiple RNAs from one transcript is called alternative splicing. In this section, I start with the history of splicing and cover the basic knowledge of splicing including the chemistry and alternative splicing.

1.6.1 Genome composition of eukaryotic cells

Eukaryotic genes comprise both exons, expressed in the final mRNA, and introns which contain regulatory elements of gene expression. The density and size distribution of introns and exons vary between species (Zhu et al., 2009). In vertebrates, exons are relatively shorter and introns are longer.

The evolution of exon-intron structure and the evolutionary property of introns have been discussed in the context of the “introns-early” vs. “introns-late” debate (reviewed in (Rogozin et al., 2012)). The intron-early hypothesis proposes that (nearly) all introns present in eukaryotic cells were inherited from prokaryotic ancestors. In this theory, the difference in gene structure among homologous eukaryotic genes is due to differential intron loss (Gilbert, 1987). The intron-late hypothesis proposes that introns emerged during evolution (Logsdon, 1998). Comparative studies between different eukaryotic organisms were performed to see conserved intron positions, or loss of introns, and elucidate which theory could be correct. However, intron gain is not easy to detect and a defined mechanism of intron evolutions is yet to be proved.

1.6.2 Discovery of splicing

The discovery of splicing, the process which turns the nuclear primary transcript RNA to cytoplasmic RNA, was based on the availability of techniques to study unstable RNA and accumulation of knowledge about RNA (reviewed in (Darnell, 2013)). In the late 1950s, RNA labelling with isotopes in nucleated cells illustrated that RNA synthesis in the nucleus was not followed by a transfer of these RNA species to the cytoplasm because of their instability (Harris, 1959). To understand the process between nuclear and cytoplasmic RNA, the study of unstable nuclear RNA was challenged. In 1971-12, a 3′ unitary-sized polyA segment was observed in polyribosomal RNA (Darnell et al., 1971) which was shown to be added by polyA polymerase (Edmonds and Abrams, 1960). A few years later, a 5′-methylated GpppX$_m$p cap (p:phosphate; X: methylated base) at the 5′ end of animal virus mRNA was reported (Shatkin, 1974). Finally in 1977, electron microscopic (EM) analysis of adenovirus type 2 (Ad2) infection uncovered RNA splicing (Berget et al., 1977) (Berk, 2016). Many completed different late cytoplasmic Ad2 mRNAs were bound to Ad2 DNA at the same three regions close to the transcription start site and also to distant genomic segments. Thus the groundbreaking idea of splicing was born from the observation of the reduction in length from large primary Ad2 RNAs to shorter cytoplasmic RNAs (Bachenheimer and Darnell, 1975). In addition, the EM photo revealed Ad2 mRNAs with different lengths from each single large adenovirus pre-mRNA, whose event is called alternative splicing (Chow and Broker, 1978) (Nevins and Darnell, 1978).

1.6.3 Key factors and sequences for splicing

When splicing is carried out, the interaction between cellular sequences and proteins is required (reviewed in (Rogozin et al., 2012)). There are four key factors involved in splicing: splice sites located at both ends of the intron, a branch point in an intron, a polypyrimidine tract in an intron, and small nuclear ribonucleoproteins (snRNPs).

The splice site at the 5' and 3' ends of introns are defined as identical dinucleotides. The 5' dinucleotides in introns are known as splice donors (SDs) and the 3' dincleotides as splice acceptors (SAs). Intron removal is processed using those splice signals (see more details in 1.6.4). Splice site consensus sequences vary by species. For instance, for human (Churbanov et al., 2006) consensus sequence at a SD is (A/C)AG|GU (3' exon and a SD at the 5' intron underlined are indicated on the left or right of the vertical line, respectively), while, that at a SA is (C,U)AG|G (a SA at the 3' intron underlined and 5' exon are indicated on the left or right of the vertical line, respectively). The branch point locates anywhere from 18 to 40 nucleotides upstream of a SA and its sequence is YNYYRAY (Y: pyrimidine; N: any bases; R: purine; A: adenine) (Tazi et al., 2009). The reactive adenosine in the branch point is involved in the formation of the lariat-like structure, by the attachment of the cleaved 5' end of the intron to a branch point in the splicing intermediate. A polypyrimidine tract (C or U) that locates between the branch site and an SA plays an important role for an SA recognition via host protein U2AF (Sickmier et al., 2006). snRNPs catalyze the splicing reaction by recognising those crucial sites above and forming spliceosomes, then orchestrating the site-specific transesterification reaction (Clancy, 2008). Five snRNPs are known to form a spliceosomal complex; U1, U2, U4, U5, and U6.

In most Eukaryotes, the vast majority of spliceosomal introns contain GU at the SD and AG at the SA (Churbanov et al., 2006) and the splicing of such universally conserved U2-type introns is mediated by the listed five snRNPs above, a U2 spliceosome. However, atypical introns with distinct dinucleotides have also been discovered: AT at the SD and AC at the SA (Jackson, 1991) (Hall and Padgett, 1994). Splicing of this minor class of U12-type intron is processed by a specific spliceosome that includes low-abundant snRNPs, called a U12 spliceosome (reviewed in (Turunen et al., 2013)). The U2 and U12 spliceosomes share U5 snRNP and, especially, the U12 spliceosome contains U11, U12, U4atac, and U6atac as possibly equivalent to U1, U2, U4, and U6 of U2 spliceosome, respectively.

There are two principle machineries of splicing signal recognition: exon definition and intron definition (reviewed in (Ast, 2004)) (Fig.1-6). This theory describes that splicing is initiated either by recognition of exon or intron first. Both mechanisms are mediated by cis-elements either in the exon or intron. Here I describe the two recognition systems using exonic splicing enhancers (ESEs) that enhance the inclusion of exons with ESE (Ibrahim et al., 2005) as an example. Binding of serine/arginine (SR) protein to ESE on an exon recruits U1 to the downstream SD and U2 auxiliary factor (U2AF) to the upstream SA on the different introns, respectively, followed by recruiting U2 to the branch site (Robberson et al., 1990). As a result, the formation of the cross-exon recognition complex is promoted and the exon with the ESE is included in spliced mRNAs. This splicing signal recognition is called exon definition (Berget, 1995). Alternatively, intron definition can be explained by U1 being recruited to the upstream SD and U2AF/U2 to the downstream SA and the branch site on the same intron, respectively (Romfo et al., 2000).

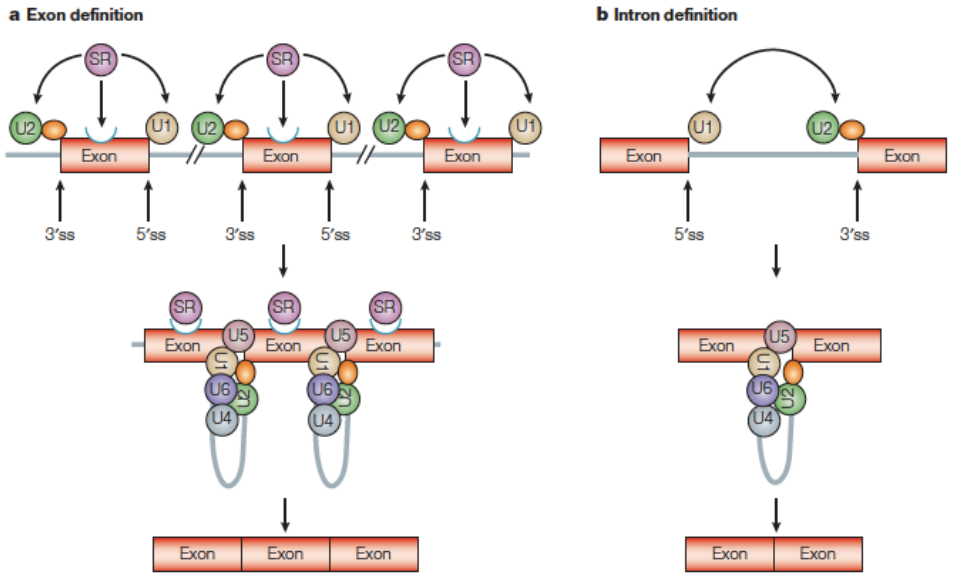


Fig.1-6 Exon and intron definition

(a) Exon definition. SR protein (purple) binding to exonic splicing enhancer (blue) enhances snRNPs binding on the 3' (U2: green; U2AF: orange) and 5' (U1) splice sites. This induces the recruitment of additional snRNPs subsequently to complete splicing. (b) Intron definition. 5' and 3' splice sites on the same intron are recognised by snRNPs. Figure reprinted from (Ast, 2004).

1.6.4 Splicing machinery

Splicing removes intron sequences between exons and joins the exons together via interaction with spliceosomal proteins. A few steps of a chemical reaction are required to complete splicing (Fig.1-7).

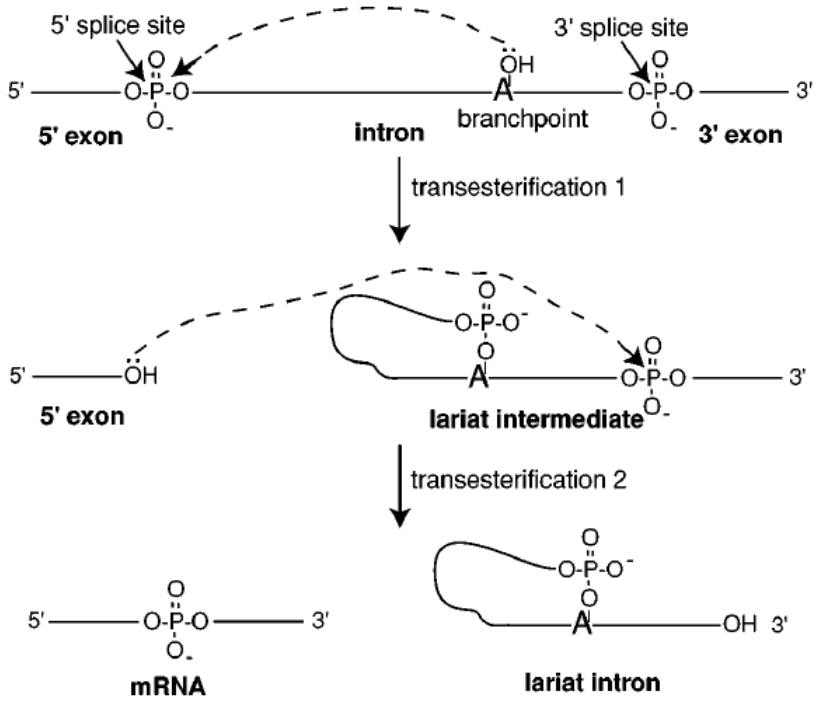


Fig.1-7 The process of splicing

(Top) The first transesterification for the formation of the lariat structure. From the left of figure 5' exon, intron and 3' exon are illustrated, the region of which is divided by 5' or 3' splice sites. The dashed arrow indicates the attack of the hydroxyl group to 5' splice site. (Middle) The second transesterification. The dashed arrow indicates the attack of the hydroxyl group to 3' splice site. (Bottom) The joining of 5' and 3' exons and the release of the lariat structure. Figure reprinted from (Brow, 2002).

Firstly, with the aid of binding of the U1 snRNP to the downstream 5' splice site, the cleavage at the 5' end of the intron in pre-mRNA occurs (Siliciano and Guthrie, 1988). The cleaved end of the 5' intron then attaches to a branch point

by pairing with guanine and adenine nucleotides to form a looped structure that is known as “lariat” (Berget et al., 1977). This attachment reaction occurs by attacking a hydroxyl group on a carbon atom of the adenosine residue of the branch point to the phosphorus atom of guanine nucleotide at the downstream 5’ splice site. This chemical reaction is called transesterification.

U1/5’ splice site base-pairing is weakened in an ATP-dependent manner (Staley and Guthrie, 1999) and this allows U2 binding to the branch point. U4/U5/U6 try-snRNPs are further recruited to the pre-spliceosome. The 3’ hydroxyl group of the released 5’ exon then attacks the phosphodiester bond of guanine of the upstream 3’ splice site. As a result, the lariat structure is removed and the5’ and 3’ exons are joined together to form mRNA.

1.6.5 Alternative splicing

Alternative splicing increases the diversity of gene expression (Ponting, 2008), for example in the IgM locus (Early et al., 1980), therefore alternative splicing is an essential process in the development and differentiation of cells. This explains the expression of over 90,000 proteins from approximately 23,000 genes (Hernandez-Lopez and Graham, 2012). 95 % of human pre-mRNA is estimated to undergo alternative splicing in different human tissues (Johnson et al., 2003) (Wang et al., 2008a). Throughout the eukaryotic evolutionary tree, alternative splicing is more prevalent in higher eukaryotes than lower eukaryotes and the number of genes that undergo alternative spicing is larger in vertebrates than invertebrates (Alekseyenko et al., 2007) (Artamonova and Gelfand, 2007). By alternative splicing, several forms of mRNAs can be generated by the following mechanisms: exon skipping, alternative splice site (3’ or 5’) selection, intron retention, mutually exclusive exons, alternative promoter usage, and alternative polyadenylation (Hernandez-Lopez and Graham, 2012) (Fig.1-8).

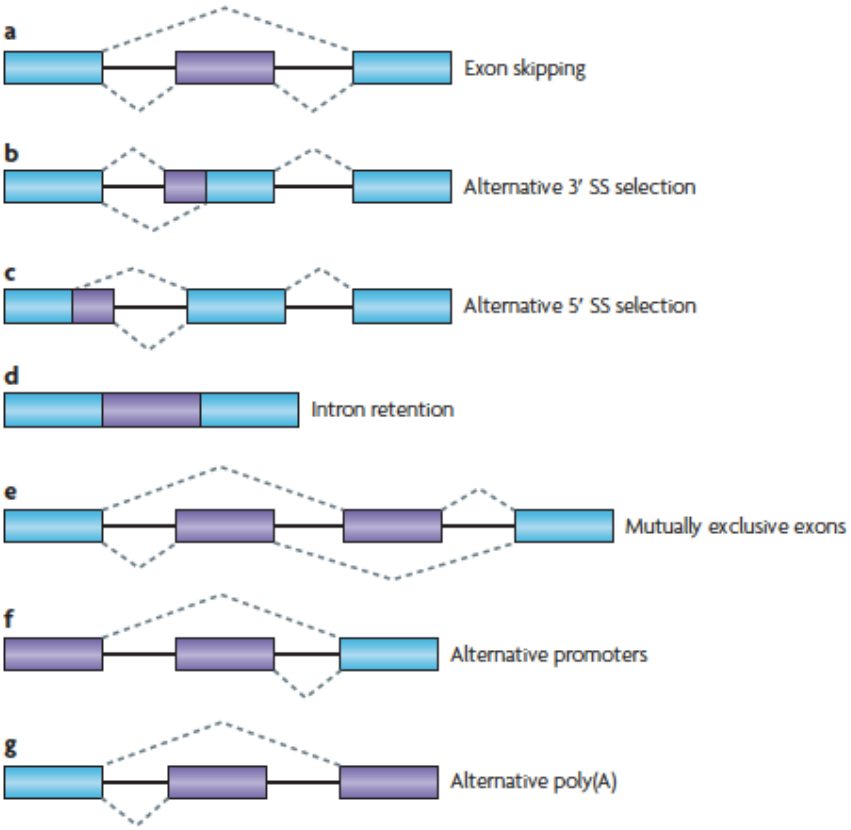


Fig.1-8 Patterns of alternative splicing

(a) Exon skipping. A cassette exon is spliced together with surrounding introns. (b) Alternative 3' or (c) 5' splice site selection. Two or more splice sites are recognised in the same exon. (d) Intron retention. An intron is retained in mRNA. (e) Mutually exclusive exons. Clustered internal exons are spliced in a mutually exclusive manner. (f) Alternative promoter usage. The generated transcripts have different 5′ untranslated regions (UTR) or open reading frames (ORF). (g) Alternative poly (A). The generated transcripts have different 3' UTR. Constitutive (blue) and alternatively spliced (purple) exons are illustrated in the figure. Introns are shown by a solid line. Dashed lines are splicing options. Figure reprinted from (Keren et al., 2010).

The first three mechanisms of alternative splicing are the major patterns and the rest is rare and more complicated (Keren et al., 2010). Exon skipping and alternative 3' splice site selection account for the major alternative splicing events in higher eukaryotes (Alekseyenko et al., 2007) (Sugnet et al., 2004), while intron retention is the most prevalent alternative splicing events in plants,

fungi, and protozoa (Kim et al., 2008).

In addition, alternative splicing can also introduce stop codons that can be targeted as a prematurely terminated gene product and degraded by nonsense-mediated decay (NMD) (Lewis et al., 2003). In this regard, alternative splicing can negatively regulate gene expression. In some cases, non-functional gene products generated by alternative splicing can be a cause of disease as introduced in the next section 1.6.6 (Cooper et al., 2009) (Tazi et al., 2009) (Scotti and Swanson, 2016).

1.6.6 Alternative splicing in disease and therapy

1.6.6.1 Neuronal disease

One example for alternative splicing as a cause of diseases is spinal muscular atrophy (SMA) characterised by degeneration of alpha motor neurons in the brain stream and spinal code (Wirth et al., 2006) (Tazi et al., 2009). SMA is the second most common autosomal recessive disorder and a leading genetic cause of infant mortality. The genetic cause of this disease is a homozygous loss of the survival motor neuron 1 (*SMN1*) gene. The human genome encodes *SMN2,* almost identical to *SMN1;* however, *SMN2* has a C to T change at position 6 in the exon 7, which leads to exon-skipping of the exon 7. As the result, the SMA2 protein is less stable and probably non-functional. Therefore it cannot be compensated for the loss of the SMN1 protein. The therapeutic approach for SMA focuses on increasing the expression level of functional *SMA2* by including the *SMA2* exon 7 by antisense oligonucleotides (AONs) that target the splicing silencer (Spitali and Aartsma-Rus, 2012).

Secondly, Duchenne muscular dystrophy (DMD) is a severe muscle-wasting disorder caused by mutations in the *DMD* gene encoding dystrophin that links the cytoskeleton to the extracellular matrix of muscle fibers. Point mutations and frame-shifting deletions insert stop codon in the exon 51, resulting in the formation of a premature dystrophin protein. In order to restore reading fame,

AON targeting the exon 51 is applied to hide it from the splicing machinery and skip the exon in pre-mRNA (Spitali and Aartsma-Rus, 2012).

Finally, Hutchinson-Gilford progeria syndrome is a rare autosomal-dominant disease characterised by accelerated ageing, growth impairment, and shortened life span (Spitali and Aartsma-Rus, 2012). Most patients carry a mutation in exon 11 of the *LMNA* gene encoding lamin A [a C-to-T transition in the *LMNA* coding region at the nucleotide position 1824 but no putative change in the protein sequence], which introduces a cryptic SD, and the protein lacks the region coded in exon 11 (Osorio et al., 2011). This shortened protein is called progerin and is expressed during normal ageing, albeit at a lower level. Reduction in progerin production by targeting exon 10 SD and the mutation in exon 11 with phosphorodiamidate morpholino oligomers (PMOs) were proven in the $LMNA^{G609G}$ knock-in mouse model, resulting in an improvement in survival rate compared to untreated knock-in mice (Osorio et al., 2011).

1.6.6.2 Cryptic splicing as a cause of ß-thalassaemia

The majority of ß-thalassaemia patients have mutations in cryptic splice sites in the ß-globin gene (Busslinger et al., 1981). This mutation activates the cryptic splice sites, which causes aberrant splicing of the gene and subsequent generation of nonfunctional transcripts, leading to defective ß-globin expression. As a treatment option, 2'-*O*-methyl RNA AONs were developed in a pre-clinical model to target the cryptic splice sites in ß-globin to block the access of spliceosomes to those sites and restore cryptic splicing (Dominski and Kole, 1993) (Svasti et al., 2009). However, the treatment for ß-thalassaemia by AONs is not clinically established yet and viral vector-mediated treatment is being further developed to be more durable (Arechavala-Gomeza et al., 2014).

1.6.6.3 Alternative splicing and cancer

Abnormalities of alternative splicing were reported in cancer (reviewed in (Tazi et al., 2009)). Tumour generation is related to a mutation in cis- or trans-acting splicing regulatory factors including RNA binding heterogeneous nuclear RNP

(hnRNP) and SR proteins. Changes in the concentration, localisation, composition, or activity of such regulatory proteins alter the splicing site selection. For instance, a splicing isoform of the tyrosine kinase receptor for the Macrophage-stimulating protein Ron is related to breast and colon cancer (Zhou et al., 2003). The exon 11 (147 nucleotides) skipping form of Ron (Δ Ron) is constitutively active and abnormal accumulation correlates with a metastatic phenotype. The inclusion or exclusion of exon 11 is controlled by two regulatory sequences, an enhancer and silencer, in the exon 12 and by the level of SR protein SF2/ASF that binds to those elements (Ghigna et al., 2005). Mutations that affect splice site selection can also produce non-functional tumour suppressor transcripts. In colorectal cancer (Markowitz and Bertagnolli, 2009) mutations in regulatory elements of adenomatous polyposis coli (*APC*), the tumour suppressor gene, affect exon skipping of the exon 4 (Neklason et al., 2004) or a single nucleotide deletion at the beginning of the exon 8 (Kurahashi et al., 1995).

In contrast, alternative splicing could produce isoforms that prevent tumorigenesis. For instance, persistent phosphorylation of STAT3, a transcription factor implicated in cytokine and growth factor signalling, was found in the majority of human cancer (Spitali and Aartsma-Rus, 2012). STAT3 has two isoforms: the full-length STAT3-alpha, and a shorter STAT3-beta. The STAT3-beta is produced by a cryptic SA within the exon 23, resulting in a lack of C-terminal transactivation domain. The STAT3-beta downregulates the expression of genes involved in proliferation and cancer and promotes apoptosis (Caldenhoven et al., 1996). This idea was utilised as a strategy to anti-cancer therapy by applying a PMO targeting the cryptic SA in exon 23 to increase the production of STAT3-beta (Zammarchi et al., 2011). Similarly to STAT3, production of a short form of anti-apoptotic protein Bcl-x, called Bcl-xS, is known to lead to apoptosis in different cancer cell lines (Sumantran et al., 1995) (Lindenboim et al., 2000) (Hossini et al., 2003) (Singh et al., 2016). The selection of an SD proximal to exon 2 results in Bcl-x, while an SD distal to the exon 2 results in Bcl-xS formation.

1.6.6.4 Alternative splicing and HIV replication

Alternative splicing is a key mechanism that is used by most DNA viruses and nuclear replicating RNA viruses including HIV (reviewed in (Karn and Stoltzfus, 2012)). HIV has three classes of transcripts distinguished by the length (1.8, 4 and 9 kb) that are produced by alternative splicing from its polycistronic pre-mRNA (Tazi et al., 2010) (Fig.1-9). These transcripts are produced by the combinational use of different splice sites on the genome. On the HIV genome there are four splice donors (D1, D2, D3 and D4) and eight splice acceptors (A1, A2, A3, A4a, A4b, A4c, A5 and A7) which result in more than 40 different spliced mRNA species (Karn and Stoltzfus, 2012). In addition, cis-acting elements, namely exonic/intronic splice silencer (ESS, ISS) and exonic/intronic splice enhancer (ESE, ISE), further regulate the splicing of transcripts negatively or positively.

In general, HIV-1 SAs are relatively inefficient compared to constitutive cellular SAs (Martin Stoltzfus, 2009). However, those splicing controlling cis-acting elements make a difference in the relative abundances of each viral mRNA species. For instance, two ESS elements within the first *tat* coding exon repress A3 SA (Jacquenet et al., 2001). This results in a reduction in the level of both incompletely and completely spliced *tat* mRNA. In contrast, a guanosine-adenosine-rich ESE element within the exon 5 facilitates splicing at weak SAs (A4a, A4b, A4c, and A5) to increase the level of incompletely spliced *env/vpu* and completely spliced *nef* and *rev* mRNAs (Purcell and Martin, 1993). In addition, mutations in splice sites also affect viral infectivity by altering the relative amount of each transcript. For instance, 5' splice site D2 mutation decreases the affinity of U1 snRNP and the inclusion of exon 2, therefore the level of *vif* mRNA and the protein is also reduced (Madsen and Stoltzfus, 2006). The produced virus with low Vif protein expression would be more sensitive to APOBEC3G inhibition than wild-type HIV (1.2.16).

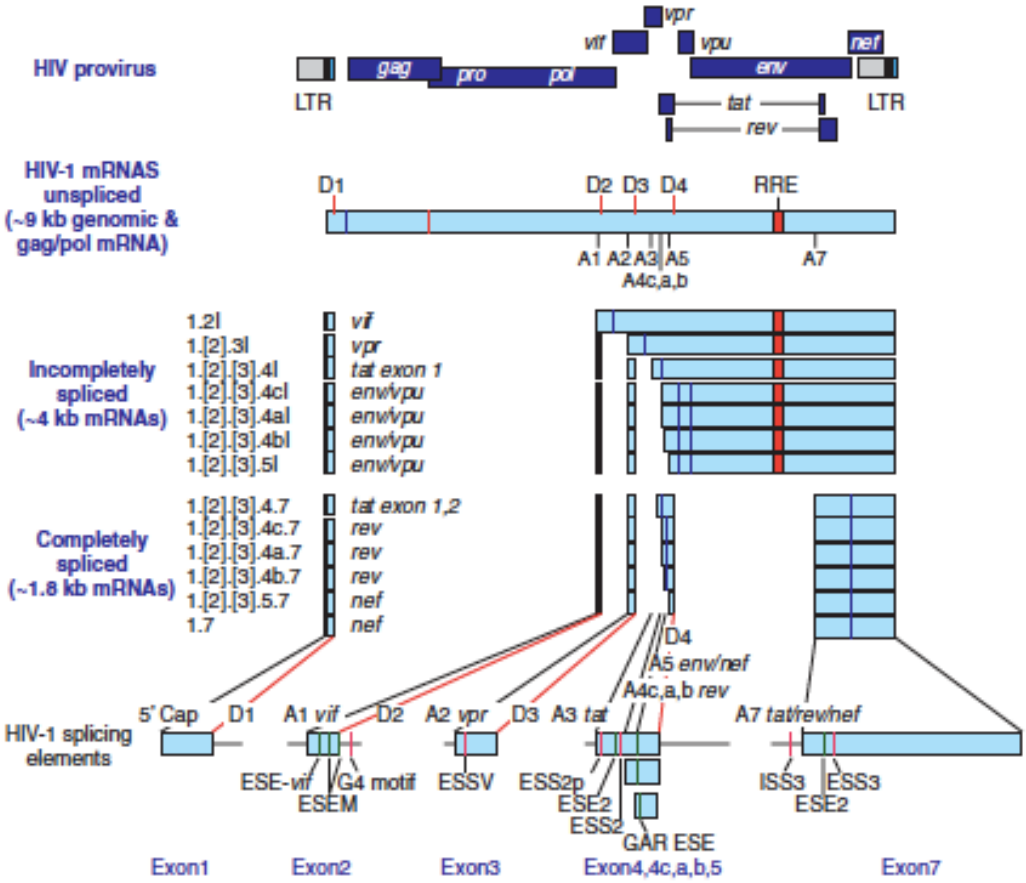


Fig.1-9 The location of splice sites, elements, and the variations of spliced HIV transcripts

(Top) HIV RNA genome. LTR is divided into three elements, U3 (grey), R (black), and U5 (blue), respectively. The position of RRE is indicated by the red rectangle. (Middle) Three classes of HIV transcripts. The first nine kb transcripts are unspliced mRNAs. Splice sites (D with red bars: splice donor; A with black bars: splice acceptor) are indicated. The second 4 kb transcript is an incompletely spliced mRNA. The third 1.8 kb transcript is a completely spliced mRNA. The use of splice sites and genes encoded on each transcript are on the left or right of the first rectangle, respectively. Each form of spliced transcripts is illustrated on the right. (Bottom) The known HIV regulatory elements. Splicing enhancers or silencers are designated by vertical green or red bars, respectively. Each rectangle corresponds to each numbered exon. Figure reprinted from (Karn and Stoltzfus, 2012).

1.7 Aims of thesis

To obtain more knowledge of IM by LV, we aimed to establish the model LV to investigate splice-in fusion mRNAs. This splicing form is generated by splicing from a host cellular SD to a vector SA as in the first report of mutagenic potential of LV (Cavazzana-Calvo et al., 2010). To investigate this particular type of splicing, we introduced a strong SA into a LV sequence for inducing splice-in efficiently. In addition, the model LV is promoterless to make marker gene express from a host promoter and in a chimeric form by splice-in. The construction of the model LV and the vector function in respect of vector titration was tested in Chapter 3. We then employed an IM assay established by our research group to test mutagenicity of the model vector and generated a mutant likely caused by integrations of the model LV. We took an approach of interrogating integration sites and characterisation of fusion mRNAs to understand the mechanism of the cellular transformation in Chapter 4. Because of the lack of information about host gene sequences in Chapter 4, next generation sequencing in mRNA was utilised. In the same assay, we used bulk transduced cells from a step before mutant isolation (STEP 4) by pooling recovered cells after IL-3 addition to see skewed gene expression that could lead to cell survival and transformation.

Throughout the experiments presented in the result chapters, we also aimed to find some further modification in this model LV and the flow of mutant isolation to make the assay system optimal. This model vector could be used to assess the mutagenicity of future clinical vectors by comparing the frequency of mutant generation and expression of fusion mRNAs via splice-in.

2 Materials and Methods

2.1 Cell culture

2.1.1 Cell lines used in this PhD project

HEK293T, WEHI3B, and Bcl15that were used in this study were a kind gift from Dr. Sean Knight. HEK293T cell line is immortalised by the expression of SV40 largeT antigen (Thomas and Smart, 2005). WEHI3B cell line originated from Balb/c mouse mylomonocytic leukemia cells (Ymer et al., 1985) that have intracistarnal A particle (IAP) insertion in two locations, *IL-3* and *HoxA* loci. These insertions maintain continuous IL-3 secretion and its immortalisation, respectively. Bcl15 cell line originated from Baf3 cells; that is, Balb/c mouse pro-B cell lines with the expression of B220 surface antigen not being expressed on mature B/T cells and monocytes (Palacios and Steinmetz, 1985). Bcl15 cells are apoptosis resistant by the continuous expression of human Bcl2 (Collins et al., 1992).

Through this PhD project, HEK293T cells were used for transient virus production and LV transduction. Bcl15 cells were used for LV vector transduction and insertional mutagenesis assay. WEHI3B cells were used for harvesting the supernatant to support parental Bcl15 cell growth. In the routine culture, HEK293T and WEHI3B cells were cultured in D10 medium which is DMEM (Dulbecco's Modified Eagle Medium) with Glutamax (Gibco, Carlsbad, CA, USA) supplemented with 50 U/ml Penicillin 50 µg/ml Streptomycin (Gibco) and 10 % Fetal bovine serum (FBS) (Labtech.com, Uckfield, UK) at 37 ^{o}C in 10 % CO_2. Bcl15 cells were also cultivated in the same condition with an additional supplement of 10 % WEHI3B supernatant in the culture medium. WEHI3B supernatant was harvested four days after the initial culture condition at 1 x 10^5 cells/ml of cell density in 120 ml in a T175 flask (Corning, St Louis, MO, USA). The supernatant was filtrated with a 0.45 µm filter and stored at -20 ^{o}C. HEK293T cells are adhesive. At the cell passage, HEK293T cells were firstly trypsinized with 1 ml 25 % trypsin-EDTA (Gibco) by incubation for a few minutes

in 5 % CO_2 and then the reaction was neutralised by adding 1 ml fresh D10 medium. The 1:4 cells were seeded in a new flask, while WEHI3B and Bcl15 cells were suspension cells. 1:4 and 1:20 cells, respectively, were seeded into a new flask every three to four days.

Cell stocks were prepared by centrifugation followed by resuspending in FBS with 10 % DMSO (Sigma, St Louis, MO, USA). They were aliquoted into a cryotube (Corning) and stored at -80 ^{o}C in an isopropanol container.

2.1.2 Cell viability/proliferation assay

Cell viability/proliferation assay was performed for three purposes in this PhD study. Firstly, cell viability of the Bcl15 cell under different puromycin concentration was measured by trypan blue staining (Gibco) to decide the critical puromycin concentration that kills off the majority of parental Bcl15 cells. 3.4 x 10^4 Bcl15 cells were cultured in duplicate wells (24-well plates) under different puromycin concentrations (0.1-1 μg/ml) for seven days. The cell number was counted on a hemocytometer. 10 % WEHi3B supernatant was kept in the culture.

Secondly, cell viability/proliferation assay was performed to test if the isolated IL-3 independent clone (the IL-3I clone) secretes autocrine factors in the supernatant. We tested cell viability of parental Bcl15 cells cultured in the supernatant from WEHI3B cells or the IL-3I clone. For this purpose, two techniques were employed independently: trypan blue staining (Gibco) and colorimetric assay CTK-8 (Dojindo, Munich, Germany). The supernatant from both WEHi3B cells and the IL-3I clone was harvested 24 h after seeding cells at 5 x 10^5 cells/ml. 3.4 x 10^4 cells were seeded in 1 ml in a 24-well plate in duplicates. We tested 20 and 100 % of the IL-3I clone supernatant and 0.01, 0.05, 0.1 (the concentration used only in Fig.4-3B), 0.5, 1 and 10 % of WEHi3B supernatant. The cell number was counted on a hemocytometer every 24 hours by trypan-blue staining. Cell expansion was estimated by the absorbance of the

culture at 450 nm by Multiskan^{TM} FC Microplate Photometer (Thermo Scientific, Wilmington, DE, USA) that was measured after the reagent was added and incubated for 4h. The absorbance was monitored every 24 hours from when cells were initially seeded. The absorbance number over the experimental period was normalised by the absorbance number obtained on Day 0. The absorbance in empty wells was considered as background absorbance and was subtracted from the other test wells with cells. All experimental conditions were performed in duplicate wells.

Finally, cell viability/proliferation assay was performed to see the difference between cell density and the rate of cell proliferation of the Il-3I. Three initial cell densities (5.6 x 10^4 (low), 1.7 x 10^5 (medium), and 5 x 10^5 (high)) were prepared in duplicate wells (24-well plates). Cell proliferation was measured by colorimetric assay CTK-8 (Dojindo) every 24 hours up to three days. The method of measurement was as described in the previous paragraph.

2.2 Construction of lentiviral vectors for splice-in study

pSIN-BX-IRES-Em (pSFFV-Em in this thesis)

The splice-in model vector plasmid used in this study, pGhr IRES-Em, was constructed based on pSIN-BX-IRES-Em that was kindly provided by Dr. Y Ikeda. pSIN-BX-IRES-Em was used as a positive control vector for Em expression at FACS. It contains a spleen focus-forming virus (SFFV) promoter, followed by internal ribosomal entry site (IRES) to drive the expression of Emerald (Em) (Demaison et al., 2002)). This vector also contains the woodchuck hepatitis virus posttranscriptional regulatory element (WPRE) to enhance gene expression (Zufferey et al., 1999). The 3' LTR of the vector has a deletion in promoter/enhancer sequence in the U3, thus this SIN vector theoretically has no transcriptional initiation from the 5' LTR. The restriction enzyme sites important for other vector construction in this thesis are indicated in Fig.3-2. EcoRI locates upstream of SFFV. BamHI and XhoI locate between SFFV and IRES. NcoI locates between IRES and Em. NotI locates between Em and WPRE.

pUBIQ-Puro

This construct is a kind gift by David Escors (Escors et al., 2008). The vector encodes a ubiquitin promoter to express a puromycin resistant (puroR) gene; therefore, it was used as a positive control for puro drug selection of transduced cells by puro construct.

pGhr IRES-Em/Puro

In order to construct a vector model that can induce splice-in insertional mutagenesis (IM), a strong splice acceptor (SA) derived from *Ghr* exon 2 was introduced. The SA of the *Ghr* exon 2 was detected as a dominant SA site that was used in virus-host cellular transcripts expressed in IL-3 independent mutants to support its survival without IL-3 supply into the culture (Bokhoven et al., 2009). Using the gDNA of parental Bcl15 cells (2.4.2), the growth hormone receptor (*Ghr*) locus was amplified by PCR primers (Ghr Fw and Ghr Rc, Table 2-1) which includes 150 nt of the intron 1 and 56 nt of exon 2, followed by three stop codons in three translation frames to separate the marker protein from the

host-vector fusion protein. The synthetic stop codons are located between BcmHI and XhoI. The DNA fragment of *Ghr* exon 2 with three stop codons was replaced with the SFFV promoter in pSIN-BX-IRES-Em by EcoRI and XhoI. pGhr IRES-Puro was constructed by replacing the Em with puro that was amplified on pUBIQ-Puro by PCR primers (Puro Fw and Puro Rc, Table 2-1) and replaced by BspHI (compatible restriction enzyme site of NcoI) and NotI.

pIRES-Em/Puro

The derivative vectors of pGhr IRES-Em/Puro, pIRES-Em/Puro, were constructed to test background transgene expression compared to that from pGhr IRES-Em/Puro. In order to remove the introduced sequence of the *Ghr* exon 2 locus, but leave the synthetic three stop codons, the PCR-amplified fragment 5' NruI-RRE-cPPT-BamHI 3' (1591 nt) by primers NruI RRE F and BamHI cPPT Rc was replaced with the corresponding region of the vector.

pGhr IRES-Em/Puro SV40 rev

Another derivative vector of pGhr IRES-Em/Puro, pGhr IRES-Em SV40 rev, has a reverse-oriented region within the vector as described (Fig.3-2E) to test its mutagenicity. Initially, SV40 late polyA signal flanked by 5' NotI and 3' EcoRI was introduced at the downstream of the transgene. pGhr IRES-Em/Puro has another EcoRI restriction enzyme site at 152 nt upstream of SA of the *Ghr* exon 2. The sequence flanked by the EcoRI sites was digested and ligated again. The vector with the reverse orientated fragment was screened by sequencing.

All PCR-amplified DNA fragments during vector construction, were initially subcloned into a pJET1.2/blunt cloning vector (Thermo Scientific) and sequenced using the commercial pJET1.2 forward/reverse sequencing primer (pJET Fw/Rv, 10 μM) to confirm if any mutation was not inserted (Table 2-1). In order to confirm that a final vector construct upon DNA ligation was valid, diagnostic DNA digestion, as well as the confirmation of vector sequence of the boundary of each ligated fragment, was performed if any mutations or deletions occurred. The obtained sequences were analysed using DNAdynamo (http://www.bluetractorsoftwa re.co.uk/).

Table 2-1 Primers used for vector construction

Primer name	Primer sequences (5' to 3')
Ghr Fw	CTCGGAATTCTGAAGTCTGAGGC
Ghr Rc	CAGTCTCGAGCTAGTTATTCAGGATCCCTGGTGACTGCCAGTGC CAA
Puro Fw	AGTCTCATGACCGAGTACAAGCC
Puro Rc	ATGCGGCCGCTAATTCCCTCGAGTCA
NruI RRE F	GGGAGAATTAGATCGCGATG
BamHI cPPT Rc	TCAGGGATCCGATATCAAGCTTATC
SV40 F NotI	GACTAGCGGCCGCAGCTTATAATGGTTACA
SV40 RC	ACTCCGCGGGAATTCCAGACATGATAAGATAC
pJET Fw	CGACTCACTATAGGGAGAGCGGC
pJET Rv	AAGAACATCGATTTTCCATGGCAG

2.3 Molecular cloning

2.3.1 Polymerase chain reaction (PCR)

General PCR on DNA was employed for vector cloning, semi-quantitative mRNA expression (RT-PCR), vector integration site identification (2.5.5), and detection of host-vector fusion transcripts (2.5.6). Different polymerases were used for different purposes as described below.

For plasmid cloning (2.2), or sample preparation for RNAseq (2.6), high fidelity KOD polymerase (Novagen, Gibbstown, NJ, USA) was used to minimise the occurrence of mutations in target sequences. For colony screening after bacterial transformation with cloning vectors (2.3.8), GoTaq G2 Hot Start Green Master Mix (Promega, Madison, WI, USA) was used to test if the target DNA fragment was successfully inserted. GoTaq G2 Hot Start Green Master Mix contains the high-performance GoTaq G2 DNA polymerase that binds to a proprietary antibody that blocks polymerase activity; hence, the pre-incubation at 94 or 95 ^{o}C before PCR reaction is required to restore the polymerase activity. Ligation-mediated PCR (LM-PCR) (2.5.5) and 5' RACE (2.5.6) was performed by Taq polymerase (Qiagen, Hilden, Germany). This polymerase was chosen because the reaction needs amplification with sufficient fidelity, but not too high.

Table 2-2 Composition of reactions for KOD polymerase PCR

Components	Concentration	Amount per reaction (µl)
10 x Buffer	10 x	5
$MgSO_4$	25 mM	3
dNTPs	2 mM each	5
Primer (Forward)	10 µM	1.5
Primer (Reverse)	10 µM	1.5
KOD polymerase	-	1
Nuclease-free H_2O	-	Up to 50 µl
Template DNA	1 to 50 ng (plasmid DNA)	X**
Total amount (µl)		50

X; valuable amount

Table 2-3 Cycle conditions for KOD polymerase PCR

Temperature (^oC)	Time	Cycle number
95	2 mins	1
95	20 sec	20-40
X	10 sec	
70	Y sec*/kb	
70	7 mins	1

* Annealing temperature X is decided as -3 to -5 ^oC of the melting temperature (Tm) of a primer. The elongation time Y depends on the length of the fragment. If the fragment length is less than 500 bp, the elongation time is 10 s/kb. If the length of the amplicon is 500 to 1000 bp, 1000 to 3000 bp, or more than 3000 bp, the elongation time is 15 s/kb, 20 s/kb, or 25 s/kb respectively.

Table 2-4 Composition of reactions for GoTaq G2 DNA polymerase PCR

Components	**Concentration**	**Amount per reaction (µl)**
GtoTaq G2 Hot Start Green MasterMix	2 x	12.5
Primer (Forward)	10 µM	2.5
Primer (Reverse)	10 µM	2.5
Bacteria colony	1 colony	-
Nuclease-free H2O	-	7.5
Total amount	-	25

Table 2-5 Cycle conditions for GoTaq G2 DNA polymerase PCR

Temperature (^{o}C)	**Time**	**Cycle number**
95	2 mins	1
95	30 sec	30
50	45 sec	
72	1 min	
72	5 mins	1

2.3.2 Restriction enzyme digestion for plasmid cloning

This technique was used for cutting out a DNA fragment in a vector and cloning it into another vector or diagnostic purpose. The digestion reaction was performed in 50 µl total reaction volume for plasmid cloning, followed by gel excision and purification (2.3.5), while in 20 µl for diagnostic digestion to test whether a size of DNA fragment is as expected.

Table 2-6 Experimental components and conditions for restriction digestion

Component	Diagnostic purpose	Cloning purpose
10 x Buffer*	2	5
100 x BSA**	0.2	0.5
DNA	200-500 ng	>1 µg
Nuclease-free water	Determined by the amount of DNA and RE	Determined by the amount of DNA and RE
Restriction enzyme (RE, Promega)	Enzyme unit*** number for the amount of input DNA	Enzyme unit number for the amount of input DNA
Total volume (µl)	20	50
Incubation temperature (^{o}C)	According to enzyme (e.g.37 ^{o}C)	According to enzyme (e.g.37 ^{o}C)
Incubation time	1 to 2 hours	a few hours to overnight

10 x buffer*; according to the choice of restriction enzyme, the company-suggested buffer was chosen, BSA**, Bovine Serum Albumin, to stabilise enzymes in the reaction mix, Enzyme unit***; one enzyme unit can digest 1 µg DNA in an hour.

When using different REs in one reaction mix, the amount of RE was adjusted according to the enzyme efficiency in the chosen buffer (e.g. if the enzyme efficiency is 50 %, twice the amount was added). The digested product was electrophoresed for the diagnostic test or for gel purification.

2.3.3 Agarose gel electrophoresis

In order to prepare 1 % of agarose gel for instance, 1.0 g agarose (Sigma) was prepared in 100 ml 1 x Tris base, acetic acid, and ethylenediaminetetraacetic acid (EDTA) (TAE) buffer (the stock concentration: x 50). This mixture was heated in a microwave until all the agarose particles were dissolved homogenously in the 1 x TAE. Once the gel had cooled down, ethidium bromide (Dutscher Scientific, Essex, UK) was added to the final concentration at 0.25

μg/ml. The agarose gel was poured into a tray with an appropriate size and a good comb was inserted. Once the gel was solidified, the gel was set inside an electrophoresis tank (BIO-RAD, Dublin Ireland) filled with 1 x TAE. Upon electrophoresis, samples were mixed with 6 x loading buffer (Thermo Scientific) to make the final concentration at 1 x and load in each well. 1 kb plus DNA ladder (Invitrogen, Carlsbad, CA, USA) was run in parallel to indicate the DNA size. When better band separation was required, a gel with higher concentration (2 %) was used. The electrophoresis was performed with an electric current of 100 V from 45 minutes to 1 hour for DNA fragments to migrate to an anode. In order to visualise the DNA fragments the gel was exposed to UV light in the GelDoc-It®^{e} Imaging System (UVP, Cambridge, UK).

2.3.4 Purification of electrophoresed PCR product

When a PCR product was used for cloning, 5 μl was loaded in “a reference well” and 45 μl in “an excision well”. After the electrophoresis, the reference well was exposed to UV light on High Performance Transilluminators (UVP) to confirm the location of the DNA band to be excised in the excision well that was not exposed to UV light. Using a blade, the corresponding part in the excision well was excised. The excised agarose gel was purified by a Gel Extraction Kit (Qiagen) according to the manufacturer’s protocol. The final eluted sample was used for ligation or cloning. The concentration of the purified DNA was measured by Nanodrop 3300 spectrophotometer (Thermo Scientific). The sample was kept at -20 ^{o}C.

2.3.5 DNA dephosphorylation treatment for plasmid cloning

Upon plasmid cloning, DNA dephosphorylation was performed to avoid the self-ligation of a DNA fragment that was digested by the one restriction enzyme at both ends. Calf intestinal alkaline phosphate (CIAP, Promega) was used to remove phosphate groups from both 5’ termini. The CIAP (1 u/μl) was used by

1:100 dilution up to 10 pmol of 5' ends of DNA. The reaction was performed by incubation at 37 ^{o}C for 30 minutes, twice, by adding the same amount of CIAP before the second round of the incubation. For the following purpose, such as ligation, this reaction mix was purified by the Gel Extraction Kit (Qiagen) to get rid of any inhibitory compounds in it.

2.3.6 DNA ligation for vector cloning

T4 DNA ligase (Promega) was used to ligate a DNA fragment with another piece that has compatible ends to be ligated. The general condition for ligation was set up in 10 µl according to the manufacturer's instructions. 10 x ligation buffer that was provided by Qiagen was added to the reaction mix at the final concentration of 1 x. In general condition, 1:3 molar ration of backbone to insert was used. 0.1-1 u of ligase (3 weiss unit/µl, the original enzyme stock) was used for 100 ng DNA vector. Nuclease-free water was used to adjust the reaction mix to 10 µl. Ligation reaction mix was incubated overnight at 4 ^{o}C. The ligation mix was stored at -20 ^{o}C.

2.3.7 Subcloning digested/amplified fragments in pJET cloning vector

Subcloning in this study was performed using CloneJET PCR Cloning Kit (Thermo Scientific). pJET1.2/blunt cloning vector was designed for the subcloning of blunt-ended DNA such as DNA amplified by KOD polymerase. Even DNA has sticky end by restriction enzyme digestion or PCR amplification by *Taq* polymerase. DNA blunting enzyme (thermostable DNA blunting enzyme from *E.coli* with proofreading activity) can make the sticky ends blunt end prior to ligation. Firstly, 2 x reaction buffer was added to make the final concentration at 1 x into the reaction mix with the DNA template. This reaction mix was incubated at 70 ^{o}C for 5 minutes and immediately placed on ice. Secondly, the blunted DNA fragment was ligated into the multiple cloning sites in a pJET1.2/blunt cloning vector (50 ng) with T4 DNA ligase. The reaction mix was incubated at a

room temperature (22 ^{o}C) for five minutes to complete ligation. 5 µl of the reaction mix was used for bacterial transformation (2.3.8).

2.3.8 Bacterial transformation by heat shock method

XL1-blue competent bacteria (Invitorgen) were used to incorporate foreign vector plasmids and transform them to produce and purify the plasmids. XL1-blue competent bacteria that were stored at -80 ^{o}C were thawed on ice for 15 minutes, followed by inoculation of an aliquot of ligation mix (5 µl) or DNA plasmid (100 ng). This XL1-blue bacteria and DNA mix were incubated on ice for 20 minutes, allowing them to take up the exogenous DNA in the meantime. In order to enclose the DNA into the competent bacteria, heat shock was given at 37 ^{o}C for two minutes precisely. Immediately after the heat shock, the bacteria and DNA mix were incubated on the ice again for a few minutes and seeded on an LB agarose plate with ampicillin. The plate was incubated at 37 ^{o}C overnight to obtain bacterial colonies that potentially carry target DNA plasmids. Before preceeding those plasmids to 2.3.10, colony PCR was performed to identify which colony carries the target plasmid.

2.3.9 Antibiotics selection of transformed bacteria

Ampicillin (Sigma) was used to screen transformed bacteria that have a plasmid with an ampicillin-resistant gene. The ampicillin-resistant gene catalyses the hydrolysis of the ß-lactam ring of ampicillin to detoxify, thus bacteria which have DNA plasmid with an ampicillin resistant gene can survive in the ampicillin-containing culture. In order to purify plasmids, bacteria were cultured in LB medium with ampicillin (2.3.10).

2.3.10 Plasmid purification in cultured bacteria

Seeding the colonies obtained in 2.3.8, bacteria were cultured with a vigorous

shake at 37 ^{o}C for 14 to 16 hours in 2 ml (miniprep) or 50 - 100 ml (midiprep) LB medium with ampicillin at 100 µg/ml. Pelleted bacteria were lysed and then column-purified by the Qiagen Miniprep Kit or Midiprep Kit (Qiagen). The concentration of purified DNA was measured by a Nanodrop 3300 spectrophotometer (Thermo Scientific). The purified plasmid was digested with a restriction enzyme to confirm that the purified DNA was the target product.

2.3.11 *In silico* splice site identification of our test vectors

Prior to the vector virus titration, test vectors including pGhr IRES-Puro, pIRES-Puro, and pGhr IRES-Puro SV40 rev were analysed by the NetGene2 splice sites identifier (http://www.cbs.dtu.dk/services/NetGene2/) (Brunak et al., 1991) (Hebsgaard et al., 1996). Each proviral sequence was used as input. Splice sites were identified on both strands with nucleotide position and confidence that defines the likeliness of the site used (where 1 = a consensus splice site).

2.4 Gene transfer to Mammalian cells

2.4.1 Transient lentiviral vector (LV) production by three plasmids on HEK293T cells

The all virus supernatant used in this project was produced by transient transduction by three plasmids in HEK293T cells. The vector genome plasmids were p8.91 (packaging plasmid encoding *gag-pol*) and pMDG (VSV-G envelope-coding plasmid) as described previously (Zufferey et al., 1997). On day one, 2 x 10^6 HEK293T cells were seeded in 12 ml D10 medium in a 100 mm^2 dish (Corning) and cultured in 10 % CO_2 overnight. On day two, premixed-vector components of 1.5 µg of vector plasmid and 1 µg each of p8.91 and pMDG were then mixed with 200 µl Optimem (Gibco) and 10 µl FuGene 6 (Promega) and incubated at room temperature for 15 minutes. Meanwhile, the old medium was replaced with 8 ml fresh DMEM. Then, the incubated mixture was added onto the HEK293T cells. Cells were cultured at 37 ^oC in a 10 % CO_2 incubator. At 24 hours post-transfection, the medium was replaced with 8ml fresh D10 medium. At 48 and 72 hours post-transfection, virus supernatant was harvested and filtrated with a 0.45 µm filter and cryopreserved at -80 ^oC. Those supernatants were used for transduction to titrate it (2.4.3).

2.4.2 Genomic DNA (gDNA) extraction to obtain vector copy number in transduced cells

Extracted genomic DNA (gDNA) was used for vector titration to determine the vector copy number in transduced cells and for identification of vector integration sites by ligation-mediated PCR (LM-PCR). Generally, 1 x 10^6 cells (the maximum cell number: 5 x 10^6) that were counted by trypan blue staining were used for gDNA extraction using the DNeasy Blood and Tissue Kit (Qiagen). When the number of cells was harvested, cells were washed a few times with 1 x PBS to remove DMEM and centrifuged to make cell pellets. The supernatant was discarded and the gDNA was extracted according to the manufacture's

protocol. The concentration of the eluted sample was measured by Nanodrop 3300 spectrophotometer (Thermo Scientific). The sample was kept at -20 ^{o}C until its use.

2.4.3 Vector titration to estimate vector infectivity to host cells

2.4.3.1 The vector titration of EmGFP vectors by fluorescence activated cell sorting (FACS)

HEK293T cells have higher vector infectivity in comparison to Bcl15 cells according to the previous research carried out by our group. Therefore, virus titration experiments were performed onto both HEK293T cells and Bcl15 cells. Firstly, HEK293T cells were prepared at the concentration of 2 x 10^6 cells/ml in D10 medium that contained 8 µg/ml polybrene (Sigma). Separately, serial dilutions (1:5 reduction serially, four points for SA and SA-less vector, and four points for the positive control vector) of test virus supernatant were prepared with 8 µg/ml polybrene. Secondly, 300 µl of cell-contained D10 medium (6 x 10^5 total cells) and 200 µl viral supernatants were added into 12-well plates (Corning). When cells were plated, the cell number was measured again to obtain the accurate cell number to use for a more accurate calculation of titre. 1 ml of fresh D10 medium was added 24 hrs post-transduction to dilute polybrene and avoid any of its inhibitory effects on cell growth.

Using the similar protocol with HEK293T cells, Bcl15 cells (24-well plates) were re-suspended in D10 medium with 8 µg/ml polybrene to make the cell concentration of 10^5 cells/ml. Separately, serial dilutions (1:5 reduction serially, four points for an SA and an SA-less vector, and five points for a positive control vector) of test virus supernatant were prepared with 8 µg/ml polybrene. When cells were plated, the cell number was measured again to obtain the accurate cell number to use for a more accurate calculation of titre. The transduced cells were scaled up in a T25 flask to 10 ml four hrs post-transduction.

FACS was performed 48 hrs post-transduction. Before processing transduced cells to measure the fluorescence, they were fixed. Firstly, HEK293T cells were trypsinized. Then, HEK293T and Bcl15 cells were re-suspended in 0.5 ml 2 % paraformaldehyde (PFA, Sigma) and transferred into a FACS tube (FALCON, St Louis, MO, USA). Gene expression in the transduced cells was detected by BD FACSCalibur™ (BD Biosciences, San Jose, CA, USA) and analysed by CellQuest Pro software (BD Biosciences).

The titre for Emerald GFP (EmGFP, the further enhanced eGFP) vectors was estimated using the following formula.
Titre (IU/ml) = the number of cells exposed to virus (initial cell number) x proportion of GFP positive cells (%) x dilution factor*
* Virus dilution factor was calculated by dividing the virus input (µl) by 1000

2.4.3.2 The measurement of vector titration by quantitative PCR (qPCR)

In addition to titre measurement by FACS, vector titre was estimated based on the vector copy number transferred into the host genome. Using the same transduced cells that were processed by FACS, an aliquot of HEK293Tor Bcl15 cells were kept passed for seven to 10 days for gDNA extraction. When cells were harvested, the cell number was counted by trypan blue (Gibco) staining and 1 x 10^6 cells were processed by gDNA extraction. After cells were washed once with 1 x PBS (Gibco) to remove culture medium, gDNA was extracted via column-purification by the Qiagen DNeasy Blood and Tissue kit (Qiagen) according to the manufacturer's protocol. SYBR green dye (Qiagen) was chosen for quantitative PCR (qPCR). The qPCR was performed by 7500 Fast Real-Time PCR System (Applied Biosystems, Warrington, UK) and analysed with 7500 Software version 2.0.6. The analysis software calculated the correlation coefficient (r^2) of the standard curve, standard deviation of triplicates, and copy number of samples based on the standard curve. Standard curves were accepted as an efficient reaction when r^2 was >0.98 and the slope of the standard curve was between -3.2 and -3.9.

Primers for HIV-1 leader sequence were routinely used to obtain the total number of vector integrations into the host cellular genome, while human beta actin (HEK293T cells) or mouse beat actin (Bcl15 cells) were also used to normalise the vector copy number (Table 2-7).

Table 2-7 The primer list for SYBR green-based qPCR

Amplification target	**qPCR primer name**	**Primer sequence (5' to 3')**	**Standard plasmid**
HIV leader sequence	GT248	TGTGTGCCCGTCTGTTGTGT	SIN pHV
	GT249	GAGTCCTGCGTCGAGAGAGC	
Human beta actin	HB-actin-F	TGGACTTCGAGCAAGAGATG	pJET-HB actin
	HB-actin-RC	TTAAGTAGGCCGTCTTGCCT	
Mouse beta actin	MB-actin F	AGAGGGAAATCGTGCGTGAC	pJET-mB actin
	MB-actin Rc	CAATAGTGATGACCTGGCCG	

Table 2-8 Reaction components for SYBR green-based qPCR

Components	**Concentration**	**Standard samples (µl)**	**Test samples (µl)**
SYBR green master mix	2 x	12.5	12.5
Primer (forward)	20 µM	0.5	0.5
Primer (reverse)	20 µM	0.5	0.5
Nuclease-free H_2O	-	6.5	X
Standard vectors gDNA	10^n copies/µl 100 ng	5	Y
Total (µl)	-	25	25

Reaction was set up in an optical 96-well plate (Applied Biosystems). The top of

the plate was sealed by an adhesive sheet (Applied Biosystems) and centrifuged briefly before loading the plate on the machine 7500 Fast Real-Time PCR System (Applied Biosystems).

Table 2-9 Cycle conditions for SYBR green-based qPCR

Temperature (^{o}C)	Time	Cycle number
95	15 mins	1
95 (denature)	15 sec	40
55 (annealing)	30 sec	
72 (extension)	30 sec	
Melting curb analysis		

Data retrieval was automatically set at the extension step by the machine.

In order to calculate titre from the obtained vector copy numbers, per sample, the cell number to 100 ng gDNA was calculated using the molecular weight of double stranded DNA as follows. The length of the haploid human genome per cell is 3,099,734,149 bp (http://www.ncbi.nlm.nih.gov/projects/genome /assembly/grc/hum an/data/). HEK293T cells are hypotriploid cells, containing less than three times the number of chromosomes compared to a normal human haploid cell (Lin et al., 2014). Hence, the total genome length per cell was three times the haploid genome length. The single DNA mass (675 Da per bp) was applied to convert the unit of bp to Da. Total DNA mass (Da) was converted into g using 1.66 x 10^{-27} kg per Da. 100 ng of gDNA was used for each sample, which theoretically contains 9594 cells. Multiplicity of infection (MOI) is defined as the vector copy numbers per single cell. Therefore, MOI was obtained by dividing the raw vector DNA copy number by the cell number to 100 ng gDNA. The length of the haploid mouse genome is 2,803,568,840 bp (http://www.ncbi.nlm.nih.gov/assembly/GCA000001635.6). Bcl15 cells are diploid; therefore the total genome length per cell was twice the size of the

haploid genome length. According to the conversion of bp to the cell number, as used in HEK293T cells, 100 ng of samples were assumed to be derived from 15911 cells.

In order to draw a standard curve to obtain the vector DNA copy number in test samples, SIN pHV plasmid (lentiviral vector with SIN LTR, a kind gift from Dr. Sean Knight) (Knight et al., 2010), that contains the HIV-1 leader sequence, was serially diluted (10, 10^2, 10^3, 10^4, 10^5 copies per µl). The target copy in each sample was proportional to test sample amplification that was shown by Ct value. Vector titre was defined by the following formula.

Titre (IU/ml) = the number of cells exposed to virus (initial cell number) x MOI (raw vector copy number that was divided by cell number in 100 ng gDNA) x dilution factor*

* Virus dilution factor was calculated by dividing virus input (µl) by 1000

2.4.3.3 The measurement of vector titration of puromycin vector by puromycin selection

Transduction of puromycin vectors was also performed on HEK293T and Bcl15 cells. In the case of HEK293T cells, on day 1, 2 x 10^5 cells were seeded in a 12-well plate in 2 ml D10 medium per well. On day two, one well was chosen to trypsinize HEK293T cells to obtain the cell number as the initial cell number for titre calculation later. Virus dilutions were serially made (1:5 reduction serially, four points for SA and SA-less vector, and five points for positive control vector) in D10 medium with 8 µg/ml polybrene (Sigma). The old culture was removed and 500 µl of the diluted virus supernatant was added into each well. 24 hours later, in order to select puromycin resistant transduced cells, a quarter of the cells (2 x 10^5) were transferred into 6-well plates onto three ml D10 medium that contains 1 µg/ml puromycin. Separately, an aliquot of the rest of the cells were passaged in a 12-well plate for seven to 10 days without puromycin to use for qPCR. Cell growth in the puromycin culture was kept monitored under a

microscope for 10 to 14 days. Half of the culture medium was changed every three to four days to keep puromycin in the culture.

1 x 10^5 Bcl15 cells in D10 medium with WEHI3B (D10WEHI20%) supernatant and 8 μg/ml polybrene were transduced with virus supernatant (1:5 reduction serially, four points for SA and SA-less vector, and five points for positive control vector) containing 8 μg/ml polybrene in a 24-well plate. At 24hr post-transduction, cells were diluted at 1:100 with D10WEHI10% containing 1 μg/ml puromycin and seeded into a 96-well plate. At the preliminary stage, when pruomycin concentration was decided, different concentrations were tested (0.5 to 1 μg/ml) to culture transduced cells by pGhr IRES-Puro. Positive control cells transduced by pUBIQ-Puro were cultured in 0.6, 0.8, 1 μg/ml puromycin. The rest of the cells were passaged in D10WEHI10% without puromycin for seven to 10 days to use for qPCR. The cell growth in the puromycin culture was kept monitored under a microscope for 10 to 14 days. Half of the culture medium was changed every three to four days to keep puromycin in the culture.

Toward the end of the monitoring period in both cell lines, cell colonies (foci) for HEK293T and cell expansion in wells for Bcl15 were observed. Especially, in order to count the number of foci, the supernatant was completely removed from each well and wells were dried briefly under the clean bench. The number of foci, or wells, was considered as the number of single puromycin resistant (puroR) clones. Using this number, “transduction rate” is the ratio of puroR number against the total cell number seeded at puromycin selection. The transduction rate was obtained by dividing the obtained puroR clone number by the total cell number seeded when puromycin selection started (Table 2-10).

Table 2-10 The method of calculating titre based on the number of puroR cells

Steps	Cell number
A: Initial cell number at transduction	87,500/well
B: Cell number 24 hrs post-transduction	350,000/well
Make 1:100 cell dilutions and 100 µl was plated in each well	
C: Initial cell number at the start of puro selection	14,000/20 wells

Puromycin titre was calculated using the formula below.

Titre (IU/ml) = the number of cells exposed to virus (initial cell number) x (transduction rate)/(virus dilution factor*)

*Virus dilution factor was calculated by dividing virus input (µl) by 1000

2.5 Insertional mutagenesis (IM) assay to obtain IL-3 independent clones that could be caused by expression of splice-in fusion mRNAs

2.5.1 IM assay protocol

IM assay used in this study was based on the previously reported protocol that was established by our group (Bokhoven et al., 2009). The vector model in this study has a puromycin-resistant gene. Therefore, puromycin selection was performed to provide a better chance of screening potential IM events caused by splice-in. The virus supernatant used in the IM assays was primarily titrated before applying it. In order to achieve the aimed puroR clone number, a cell number and volume of the virus supernatant were determined based on this preliminary titration.

On day one, Bcl15 in D10WEHI10% medium cells was transduced by pGhr IRES-Puro with 8 µg/ml polybrene in a t175 flask at 37 ^{o}C in 10 % CO_2. 24 hrs post-transduction, the same number of cells with the transduction were transferred into a new t175 flask. Puromycin was added to make the final concentration at 1 µg/ml and kept cultured for three days. On the same day, puromycin selection of cells was performed using an aliquot of transduced cells to measure vector titre. The cells were seeded in 100 µl medium in 96-well plates with 1:2 serial cell dilutions from 2000 cells/well down to 62.5 cells/well. The leftover cells that were not used for any downstream experiments were cryopreserved as a backup.

On day five, IL-3 that was kept in the culture was removed (IL-3 starvation). In order to remove IL-3 from the culture medium, cells were washed thrice with D10 medium with the interval of 15-minute incubation at 37 ^{o}C. The washed cells were again re-suspended in D10 medium that contains puromycin (1 µg/ml) and seeded at 1-1.4 x 10^5 cells per well in 24-well plates with 1:3 cell serial dilutions. This IL-3 starvation was performed for two weeks, monitoring cell condition under the microscope. Because this step is known as being harsh to transduced

cells, preventing them with proliferation, then potential cell populations that could expand without IL-3 were rescued by adding WEHI3B to make the final concentration at 10 %. The cells were kept until cell proliferation was observed (up to one month). In order to verify whether the expanded cells were IL-3 independent, IL-3 starvation was performed again in a t25 flask without IL-3 supplement for two weeks to a month. The isolated IL-3 independent clone underwent the downstream analysis to examine integration site identification and fusion mRNAs expressed in the clone.

2.5.2 RT-PCRon puromycin transcript to see if splice-in between the introduced *Ghr* exon 2 SA and an upstream SD is a major event

Total RNA was extracted by the Qiagen RNeasy mini kit (Qiagen) from 1.5 x 10^6 Bcl15 cells transduced by pGhr IRES-Puro. cDNA was synthesised by reverse transcription on the 1 µg total RNA by the Quantiscript Reverse Transcription kit (Qiagen) (M.M 2.3.3). PCR was performed with two different forward primers paired with a reverse primer on the synthesised single-stranded cDNA as a template. RT-PCR was carried out by GoTaq G2 DNA polymerase (Promega). The amplified samples were run in 1 % agarose gel and electrophoresed for DNA band detection.

Table 2-11 The list of primers to test splice-in the introduced *Ghr* exon 2 SA with an upstream SD

Amplicons	Forward primers (5' to 3')	Reverse primers (5' to 3')
Ghr (upstream of the *Ghr* SA)	TCAGGTCTTCTTAACCTTGGCA	ACACCTTGCCGATGTCGAG
Ghr (downstream of the *Ghr* SA)	CTCGGAATTCTGAAGTCTGAGGC	ACACCTTGCCGATGTCGAG
Puro	AGTCTCATGACCGAGTACAAGCC	ATGCGGCCGCTAATTCCCTCGAGTCA

2.5.3 Analysis in the IL-3 independent clone (the IL-3I clone) about the expression of IL-3 mRNA as an autocrine factor

3 x 10^6 cells (the maximum cell number: 1 x 10^7) were washed with 1 x PBS and pelleted by centrifugation for total RNA extraction. Total RNA was extracted by following the manufacturer's protocol of the Qiagen RNeasy Mini Kit using the column purification method (Qiagen). The concentration of extracted total RNA was measured by a Nanodrop 3300 spectrophotometer (Thermo Scientific). The rest of the extracted total RNA was kept at -80 ^oC to avoid its degradation.

1 µg of total RNA of the IL-3I clone was reverse transcribed by the Qiagen Quantiscript Kit (Qiagen) to synthesise complementary DNA (cDNA). Firstly, extracted total RNA was treated with a gDNA wipe-out buffer to eliminate the gDNA contamination in RNA samples. This reaction is performed in 14 µl. Then, the gDNA-free template was reverse transcribed using the optimised cocktail of oligodT and random primers to increase the yield of the cDNA from all regions of RNA transcripts. This reaction is performed in 20 µl. The recombinant reverse transcriptase that contains the RNase inhibitor has three enzyme activities; RNA-dependent DNA-polymerase activity, RNaseH activity, and DNA-dependent DNA polymerase activity. Therefore, the final product can be

used directly in the next step without treating the sample with additional RNaseH.

According to the manufacturer's suggestion, 2 µl of the reverse transcribed product was used for RT-PCR and qRT-PCR to detect IL-3 mRNA. Both reactions were performed in 20 µl. As control samples, two negative controls (water control at PCR and from reverse transcription), WEHI3B cells, parental Bcl15 cells, and two step four survivors from the first IM assay were tested. WEHI3B and parental Bcl15 cells were tested as positive and negative controls, respectively, for IL-3 mRNA detection. In both RT-PCR and qRT-PCR, the amplification of mouse beta actin was tested as a standard control. In RT-PCR, amplified samples were run in 1 % agarose gel and electrophoresed for DNA band detection.

Table 2-12 The list of primers used for RT-PCR and qRT-PCR in the IL-3l clone for IL-3 mRNA detection

Amplicons	Forward primers (5' to 3')	Reverse primers (5' to 3')
IL-3 mRNA	IL-3 Fw	TATCCCGGGAATGGTTCTTGCCAGC
	IL-3 R	GTCGACTTAACATTCCACGGTTCC
Mouse beta actin	MB-actin F	AGAGGGAAATCGTGCGTGAC
	MB-actin Rc	CAATAGTGATGACCTGGCCG

2.5.4 Ligation mediated PCR (LM-PCR) to identify integration sites on gDNA

This technique was employed to identify vector integration sites in transduced cells as previously described (Wu et al., 2003). Firstly, gDNA was extracted by the Qiagen DNeasy Blood and Tissue Kit from the isolated IL-3 independent clones (10^6 cells). 1 µg gDNA was digested by a four-cutter enzyme, NlaIII

(5'-CATG^-3'), overnight at 37 ^{o}C. The reaction mix was column-purified to remove any inhibitory agents within samples by a Gel Extraction Kit (Qiagen) for downstream use. A pair of linkers with NlaIII overhang sequence (100 pmol each) was synthesised by incubation at 95 ^{o}C for five mins then by cooling down to 15 ^{o}C (1 ^{o}C decrease/min). 0.5 pmol of the annealed linkers and 10 µl of digested gDNA were ligated by T4 DNA ligase (Promega) at 15 ^{o}C overnight. The reaction mix was again column-purified by the Gel Extraction Kit (Qiagen). This purified product was subjected to the first round of PCR. 1 µl of the first round of the PCR product was used in a nested PCR. HotStart Taq polymerase (Qiagen) was used for both PCRs. Primers used in this experiment are listed below.

Table 2-13 The list of primers used for LM-PCR

Purpose	**Primers**	**Primer sequence (5' to 3')**
Linker generation	Linker positive	GTAATACGACTCACTATAGGGCTCCGCTT AAGGGACTACATG
	Linker negative	TAGTCCCTTAAGCGGAG
1st PCR	3'LTR-F	AGTGCTTCAAGTAGTGTGTGCC
	Linker F	GTAATACGACTCACTATAGGGC
Nested PCR	Nested 3'LTR	GTCTGTTGTGTGACTCTGGTAAC
	Nested F	AGGGCTCCGCTTAAGGGAC

Table 2-14 Composition of reactions for HotStart Taq polymerase PCR for LM-PCR

Components	Concentration	Amount per reaction (µl)
10 x PCR buffer (contains 15 mM $MgCl_2$)	10 x	10
dNTPs	10 µM each	2
Primer Fw	10 µM	5
Primer Rv	10 µM	5
Template DNA	≤1 µg/100 µl reaction	30
Nuclease-free water	-	Up to 100
HotStart Taq polymerase	2.5 units/reaction	0.5

Table 2-15 Cycle conditions for HotStart Taq polymerase PCR for LM-PCR

Temperature (°C)	Time	Cycle number
95	15 mins	1
94	45 sec	30
50	30 sec	
72	1 min	
72	10 mins	1

The final PCR products were electrophoresed on 2 % agarose gel containing ethidium bromide (Dutscher Scientific) and the excised DNA fragments were column-purified by the Gel Extraction Kit (Qiagen). The purified fragments were ligated into pJET1.2 cloning vectors (Thermo Scientific). The ligated vectors were transformed into XL1-Blue competent cells. The obtained bacterial colony was seeded and propagated in ampicillin-containing 2 ml LB medium. The sample plasmids were column-purified by Qiagen miniprep kit (Qiagen) and

sequenced with pJET commercial primers (Table2-1) by University College London sequencing service, Beckman or GATC Biotech. Obtained sequences were analysed by DNAdynamo (http://www.bluetractorsoftware.co.uk/) and vector and host cellular genome sequences were annotated. The defined host cellular sequence was analysed by BLAT (https://genome.ucsc.edu/cgi-bin/hgBlat?command=start) to validate the location of the sequences on mouse genome (chromosome coordinate). The region of the chromosomal number and position obtained by BLAT were used as input in Ensembl (http://www.ensembl.org/index.html) to gain the information of genes surrounding the integration sites in 1 mega base (Mb) window.

2.5.5 Rapid amplification of 5' cDNA ends (5' RACE) to identify host cellular sequence adjacent to vector sequence on fusion puromycin mRNAs

5' RACE was performed to identify host cellular sequences that make splice-in fusion transcripts with vector sequences. Total RNA was extracted by the Qiagen RNeasy mini kit (Qiagen) from puroR clones isolated in the IM assay. 1 μg of the RNA was reverse transcribed by the Quantitect reverse transcription kit (Qiagen). RT primers in this kit were the optimised cocktail of oligo-dT and random primers. In order to remove any possible inhibitors in the sample, cDNA was purified by ethanol precipitation. The purified cDNA was ligated by T4 RNA ligase (TaKaRa, New York, NY, USA) at 15 ^{o}C for 15 to 18 hours. 1 μl of the ligated cDNA was used for the first round (1:10 dilution of the sample from the previous reaction) and nested PCRs. HotStart Taq polymerase (Qiagen) was used for both PCRs. Primers used in this experiment are listed below. The first 5' RACE targeted the transcript with a puromycin resistant gene. The second 5' RACE was primed further upstream compared to the first trial, 44 nt downstream of 5' LTR. We aimed to obtain the host cellular sequence next to the viral sequence more easily at the later trial.

Table 2-16 Primers used for rapid amplification of 5' cDNA ends (5' RACE) at the first trial

Purpose	Primer name	Primer sequence (5'->3')
Reverse transcription (5'phos)	RT PURO	ACACCTTGCCGATGTCGAG
1st PCR	PCR1 PURO F	ATGACCGAGTACAAGCCCAC
	PCR1 SA	ATTCAGGATCCCTGGTGACT
Nested PCR	Nested PURO F	GTCACCGAGCTGCAAGAAC
	Nested SA Rc	GCCAGTGCCAAGGTTAAGAA

Table 2-17 Primers used for rapid amplification of 5' cDNA ends (5' RACE) at the second trial

Purpose	Primer name	Primer sequence (5'->3')
Reverse transcription (5'phos)	GT249 (5'Phos)	GAGTCCTGCGTCGAGAGAGC
1st PCR	GT248	TGTGTGCCCGTCTGTTGTGT
	PCR1 R RC	GAAGCACTCAAGGCAAGCTTTA
Nested PCR	U5 Nested F	GAAAATCTCTAGCAGTGGCG
	DelU3 Nested Rc or R Nested RC	AGCAGATCTTGTCTTCGTTGG or CCAGGCTCAGATCTGGTCTA

Table 2-18 Composition of reactions for HotStart Taq polymerase PCR (5' RACE)

Components	Concentration	Amount per reaction (µl)
10 x PCR buffer (contains 15 mM $MgCl_2$)	10 x	10
dNTPs	10 µM each	2
Primer Fw	10 µM	5
Primer Rv	10 µM	5
Template DNA	≤1 µg/100 µl reaction	1 (1:10 sample dilution of the previous reaction)
Nuclease-free water	-	Up to 100
HotStart Taq polymerase	2.5 units/reaction	0.5

Table 2-19 Cycle conditions for HotStart Taq polymerase PCR (5' RACE)

Temperature (°C)	Time	Cycle number
95	15 mins	1
94	45 sec	30
50	30 sec	
72	1 min	
72	10 mins	1

The final PCR product was column-purified by the Gel Extraction Kit (Qiagen) and the purified DNA fragments were cloned into the pJET1.2 cloning vector (Thermo Scientific). The obtained bacterial colony was seeded and propagated in ampicillin-containing 2 ml LB medium. The sample plasmids were column-purified by the QIAprep spin miniprep kit (Qiagen), and sequenced with pJET commercial primers (Table2-1). Obtained sequences were analysed by DNAdynamo (http://www.bluetractorsoftware.co.uk/) and vector and host cellular

genome sequences were annotated. This allowed us to exhibit splicing form within fusion transcripts. Host cellular sequences were identified by BLAT (https://genome.ucsc.edu/cgi-bin/hgBlat?command=start) and further information of host cellular genes, such as the composition of genes and gene function, was obtained by Ensembl (http://www.ensembl.org /index.html).

Based on the identified three genomic sequences (Ch4, 7 and 17, respectively) by the second 5' RACE trial, we tested whether those were genuine integration sites. We designed primer pairs for those three sites as listed in Table 2-20. Initially, the primer pairs were tested on gDNA to amplify those genomic sequences properly. Then, using primer annealing to 5' and 3' side of the sequence of pGhr IRES-Puro we performed PCR on gDNA of the IL-3I clones by GoTaq G2 DNA polymerase (Promega) (see 2.3.1 for PCR protocol). PCR negative control, PCR water control, Bcl15, and a step 4 survivor from the first IM were tested.

Table 2-20 The list of primers to confirm vector integration sites based on the second 5' RACE

Purpose	Primer name	Primer sequence (5'->3')
Ch4	Fw	GACTCCAATATGATGGTCTT
	Rc	GAGCACGAAGCAACATAGTG
Ch7	Fw	CAGGATCACCATAGATACAC
	Rc	CCAACAATTTAGCGGCAATC
Ch17	Fw	AGAGTCTGTGTTCTTGCCTC
	Rc	GTGAAGGAATCTCTGATGAC
Vector 5'	GT249	GCTCTCTCGACGCAGGACTC
Vector 3'	WPRE Fw	CTCAGACGAGTCGGATCTC

2.6 Next generation sequencing (NGS) by Illumina Miseq to identify host cellular genes on fusion puromycin transcripts and characterise its splicing form

2.6.1 The work flow to run Illumina Miseq

The workflow of the next generation sequencing is based on 16S Metagenomic Sequencing Library Preparation revision B issued by Illumina (http://support.illumina.com/downloads/16s_metagenomic_sequencing_library_preparation.html). This protocol includes RNA sample preparation through to the initial analysis of the raw sequence.

2.6.2 mRNA sample preparation; mRNA extraction using oligotex beads from total RNA and direct cell lysis

For mRNA extraction from total RNA, initially 2 x 10^8 transduced Bcl15 cells were pelleted and lysed for total cytoplasmic RNA extraction by an RNeasy Midi Kit (Qiagen) according to the manufacturer's protocol. This resulted in 600-1200 µg of total RNA yield. Then, 300 µg of total RNA was used to extract mRNA by Oligotex mRNA Kits (Qiagen). Oligotex suspension is an affinity reagent to detect nucleic acids containing polyadenylic acid sequences. dC10T30 oligonucleotides are covalently linked to the surface of the component of oligotex suspension, polystyrene-latex particles via condensation reaction. After capturing mRNA by the beads, mRNA was dissociated and eluted in RNase-free H_2O.

The step of total RNA extraction can be omitted and mRNA can be extracted by direct lyses of cells. I used 2 x 10^7 cells for the direct lysis, and the total yield of mRNA was 400 to 2000 ng. 10 ng of this mRNA was directly used as I prepared for NGS samples; however, on the 1 % electrophoresed gel unexpectedly higher background DNA bands were observed in untransduced Bcl15 cells at nested PCR (Fig.5-4). In addition, the direct mRNA extraction from cells requires a

greater amount of oligotex beads and this is not cost-effective. Therefore, the RNA samples processed in the next generation sequencing were extracted from total RNA.

2.6.3 Reverse transcription (RT) with a vector specific RT primer with an overhang tag sequence

Using the vector-specific RT primer i5_R1_SA27 (SA1), reverse transcription was performed by the SuperScript ® III First-Strand Synthesis System (Invirtogen) according to the manufacturer's protocol. Reaction mix (mRNA, the RT primer and 10 mM dNTP mix) was prepared in 10 µl and mRNA was heated at 65 ^{o}C for five minutes to denature unwanted secondary structures. cDNA synthesis mix that contains reverse transcriptase was added into the reaction mix that contains mRNA, and was incubated at 50 ^{o}C for 50 minutes followed by incubation at 85 ^{o}C for five minutes to terminate the reaction. In order to degrade mRNAs in the sample, RNaseH (2 U/µl) was added and incubated at 37 ^{o}C for 20 minutes.

Table 2-21 Condition of reactions for reverse transcription by SuperScript® III reverse transcriptase: elimination of unwanted RNA secondary structure

Components	Amount per reaction (µl)
Poly(A)$^{+}$ RNA	X (1 pg-500 ng)
Gene specific primer (2µM)	1
10 mM dNTP mix	1
DEPC-treated water	To 10

Table 2-22 Condition of reactions for reverse transcription by SuperScript® III reverse transcriptase: cDNA synthesis

Components	Amount per reaction (µl)
10 x RT buffer	2
25 mM $MgCl_2$	4
0.1 M DTT	2
RNaseOUT ™ (40 U/µl)	1
SuperScript®III (200 U/µl)	1

2.6.4 Optimisation of NGS sample preparation: choosing an optimal RT primer

Among three RT primers, SA1 was designed initially. In order to test the function of primers, PCR was performed on gDNA from transduced and untransduced Bcl15 cells. The gDNA extracted from Bcl15 cells and the IL-3I clone was serially diluted (1:5 dilution, 5 points) starting from 100 ng gDNA. Three primer pairs were tested to detect the amplification of the region in the puromycin gene, the *Ghr* locus, and vector backbone sequence. PCR was performed by GoTaq G2 DNA polymerase (Promega).

Table 2-23 Primers used for PCR on gDNA to test the function of SA1 RT primers

Target	Primer	Primer sequence
Puro	Puro Fw	TGACCGAGTACAAGCCCAC
	Puro Rc	CTTCCATCTGTTGCTGCGC
The *Ghr* locus	Ghr Fw	CTCGGAATTCTGAAGTCTGAGGC
	SA1	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG GCCAGTGCCAAGGTTAAGAA
Vector sequence	RRE-cPPT	GAGACAGATCCATTCG
	SA1	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG GCCAGTGCCAAGGTTAAGAA

As next optimisation, SYBR green qPCR (Qiagen) was performed using three RT primers, including the SA1, to detect amplification in the region downstream of the *Ghr* exon 2, puromycin gene, and mouse beta actin. Prior to SYBR green qPCR, the mRNA extracted from Bcl15 cells and Bulk-Puro was reverse transcribed by the SuperScript®III First-Strand Synthesis System (Invirtogen).

Table 2-24 Candidate RT Primers for next generation sequencing

RT Primer name	The distance from the *Ghr* exon 2 SA (nt)	Primer sequence
SA1	47	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG GCCAGTGCCAAGGTTAAGAA
SA2	77	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG CGAGCTAGTTATTCAGGAT
IRES	119	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG CTTCGGCCAGTAACGTTAG

Table 2-25 Primers used for SYBR green qPCR to choose an optimal RT primer for NGS

Target	Primer	Primer sequence
The region downstream the *Ghr* exon 2 SA	Ghr exon 2 Fw	GTCTCAGGTATGGATCTTTGT
	Tag primer	TCGTCGGCAGCGTC
Puro	Puro Fw	TGACCGAGTACAAGCCCAC
	Puro Rc	ACACCTTGCCGATGTCGAG
Mouse beta actin	MB-actin F	AGAGGGAAATCGTGCGTGAC
	MB-actin Rc	CAATAGTGATGACCTGGCCG

2.6.5 Double-stranded DNA synthesis by a random octamer

In order to flank the amplified target cDNA with a known sequence, double-stranded cDNA synthesis was initiated using a primer with random octamers with a tag sequence. This second strand cDNA was synthesised by a large fragment of DNA polymerase I (Invitrogen). 10 µl of the previous reaction mix that contains cDNA (2.4.3) was used. The following incubation cycle was used for the reaction.

Table 2-26 Condition of reactions for the second strand cDNA synthesis by Large fragment of DNA polymerase I

Component	Amount per reaction (µl)
RACT2 Buffer	5
dNTPs (10 mM)	0.75
Primer (100 µM)	1
Reverse transcribed product	10
Klenow Large fragment (3-9 U/µl)	1
Nuclease-free water	Up to 50

Table 2-27 Incubation conditions for the second strand cDNA synthesis by Large fragment of DNA polymerase I

Temperature (°C)	Time (minutes)
15	10
20	10
25	10
37	60
75	20

2.6.6 Amplification of target fusion transcripts

Since the synthesised double-stranded DNA has the tag sequences at both ends, the target fusion transcripts were amplified using them by KOD polymerase (Novagen) with a cycle number of 22.

Table 2-28 Primers used for RNA sequencing

Purpose of use	Primer name	Primer sequence
Reverse transcription	i5_R1_SA27 (SA1)	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGCCAGTGCCAAGGTTAAGAA*
The second strand cDNA synthesis	i7_R2_N8 (a random octamer)	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGNNNNNNNN
Tag PCR	i5_R1 (a tag primer)	TCGTCGGCAGCGTC
	i7_R2 (a tag primer)	GTCTCGTGGGCTCGG

*Vector specific sequence

The PCR-amplified products were indexed to distinguish each sample replicate

by Nextera XT Index Kit v2 set C (Illumina, San Diego, CA, USA), followed by a purification step by Ampure XP beads (Beckman, Brea, CA, USA) according to the manufacturer's protocol. The concentration and size distribution of the purified product was assessed by a Qubit 2.0 fluorometer (Invitrogen) and Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA). Sequencing was run on the DNA libraries using a 500-cycle paired-end sequencing reagent kit (Illumina). This sequencing generated the average read length of 250 pb. Low quality sequencing tails were trimmed by Phred quality score of 30 suing trimmomatic (http://www.usadellab.org/cms/?page=trimmomatic). Separately, to analyse host-vector fusion transcripts, all forward sequence reads were filtered with vector-specific 47 nt sequence (5' GTCTCAGGTATGGATCTTTGTCAGGTCTTCTTAACCTTGGCA CTGGC 3') by trimmomatic to confirm that analysed sequence reads are genuine fusion transcripts between the host and vector sequences. Those experimental steps were kindly helped by Dr, Edward T. Mee (NIBSC) and initial sequencing analysis by Dr. Mark Preston (NIBSC) (Table 5-3 and 5-4).

I then performed the later study to identify the host sequence on fusion mRNAs. Initially, the extracted forward sequences with 47 nt vector sequence were clustered by a multiple sequence alignment tool, Clustal X (Jeanmougin et al., 1998). A representative sequence in each cluster was selected and was blasted to the mouse genome (GRCm38.p4) using Ensembl (http://www.ensembl.org/index.html) to obtain the information of host sequences. Reverse sequences of the selected forward sequences were blasted against the mouse genome (GRCm38.p4) in Ensembl. Separately, all forward sequences were blasted in NCBI nucleotide blast (https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome) to find new host cellular sequences that could not be identified by forward sequence screen.

In addition, the biological and molecular function of each identified gene was extracted from the Universal Protein Resource called UniProt (http://www.uniprot.org/), a comprehensive resource for protein sequence and data annotation (Consortium, 2010).

2.7 Buffers

Purpose of use	Reagent name	Component
Fixing solution for FACS	2 % paraformaldehyde (PFA)	1 x PBS, 37 wt. % in H_2O paraformaldehyde (Sigma)
Electrophoresis gel running buffer and gel making	Tris-acetate EDTA (TAE)	40 mM Tris (pH 7.8), 20 mM sodium actate, 1 mM EDTA
Electrophoresis sample loading	6 x loading dye	0.25 % bromophonol blue, 0.25 % xylene cyanol FF, 30 % glycerol in water (pH 6.8)
Preparation for competent bacteria cells (for resuspension)	Transformation buffer (TFB)-I	30 mM potassium acetate, 100 mM rubidium chloride, 10 mM $CaCl_2$, 50 mM $MgCl_2$, 15 % glycerol, acetic acid to pH 5.5
Preparation for competent bacteria cells (for final resuspension)	TFB-II	10 mM MOPS, 75 mM $CaCl_2$, 15 % glycerol, KOH to pH 6.5
Bacterial colony formation	Luria-Bertani (LB) Agar	LB broth, 15 g/L bacto-agar (pH.7.5)
Bacterial culture for plasmid purification	Luria-Bertani (LB) Broth	1 % bacto tryptone, 0.5 % bacto yeast, 0.5 % NaCl (pH 7.5)
Elusion buffer for extracted gDNA	AE	10 mM Tris-HCl (pH 9.0). 0.5 mM EDTA
Elusion buffer for purified DNA plasmids/fragments	EB	10 mM Tris-HCl (pH 8.5)
gDNA extraction	lysis buffer (ATL)	10 mM Tris-Cl (pH 7.4), 10 mM EDTA, 10 mM NaCl, 0.5 % SLS, 1 mg/ml Proteinase K

Total RNA extraction	lysis buffer (RLT buffer) 1 % ß-mercaptoethanol)	RLT buffer composition (confidential by Qiagen), 1 % 14.3 M ß-mercaptoethanol (Sigma)

Chapter 3

Characterisation of a lentiviral vector to study splice-in fusion transcripts

3.1 Introduction

Insertional mutagenesis (IM) caused by vector integration and the subsequent alteration of the host cellular gene splicing pattern is a potential concern in the clinical use of lentiviral vectors. Several *in vitro* studies demonstrated that LV could disrupt cellular gene splicing. The paper from our group by Bokhoven et al (Bokhoven et al., 2009) showed gene upregulation by splicing out from the LV (see below). Moiani et al used a splice trap vector, without an internal promoter but using an internal ribosomal entry site (IRES) for transgene expression, and also a clinical candidate beta-globin vector (Moiani et al., 2012) and detected a variety of fusion transcripts, including read-through, splice-in and splice-out. Cesana et al (Cesana et al., 2012) performed a similar study with a vector carrying the PGK promoter driving GFP expression, detecting read-through transcripts and a number of cryptic splice donors and splice acceptors. The potential risk of splicing alteration by a lentiviral vector was also reported in a clinical trial for beta thalassaemia (Cavazzana-Calvo et al., 2010). The vector integration contributed clonal expansion of myeloid cells by truncation and activation of transcription factor-associated gene, the high mobility group AT-hook 2 (*HMGA2*) (Modell and Darlison, 2008). This dominant clone disappeared after 6 years of clinical follow-up (Leboulch, 2013). The safety risk of this single clinical event is hard to estimate. On the one hand the expanded clone contributed to beta-globin expression in the patient allowing them to attain transfusion independence, however such expanded clones might ultimately become leukaemic after additional mutations.

Since the splicing event that caused a clonal expansion in the ß thalassaemia clinical trial was a splice-in, using an SA in the LV, we decided to develop a vector design to study splice-in. We assumed that the insertion of a potent splice acceptor (SA) upstream from an IRES driving a marker gene could increase the chance of inducing splicing by splice-in, as in the study of Moiani et al (Moiani et al., 2012).

3.1.1 Selection of a potential SA

Previously, our research group discovered distinct mechanisms of cell transformation to IL-3 independence caused by integration of gammaretroviral vectors (GRVs) or lentiviral vectors (LVs) respectively (Bokhoven et al., 2009) (Knight et al., 2010). In these studies, LV integrated into the *Ghr* locus, at sites identified by inverted PCR and LM-PCR (Fig.3-1A). Those vector integrations altered not only the original *Ghr* transcriptional pattern by inclusion of vector sequence in the transcript (Fig.3-1B), but also mRNA and protein expression, which was detected by 5'RACE, qRT-PCR (Fig.3-1C) and immunoblotting. The novel transcripts used a splice donor (SD) within the vector specific and the *Ghr* exon 2 SA (Fig.3-1B). Based on this finding, we hypothesised that the *Ghr* exon 2 SA could be a potential strong SA that can be introduced in this novel splice-in vector design to induce and detect splice-in splicing.

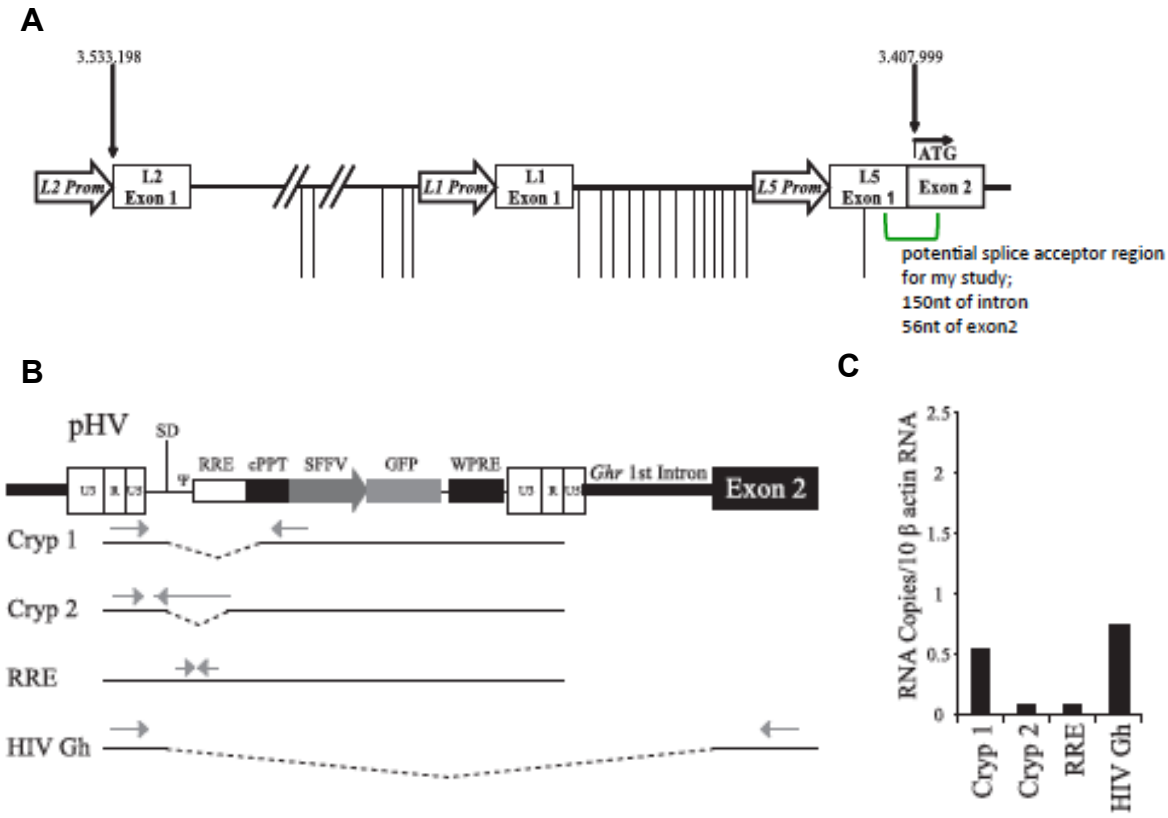


Fig.3-1 The identification of the *Ghr* locus as the mechanism by which LV make Bcl15 cells IL-3 dependent by splice-out fusion mRNA, which is described in (Bokhoven et al., 2009) (Knight et al., 2010).

(A) LV integration sites were clustered in the *Ghr* locus in the IL-3 independent mutants caused by LV integration. Each vertical line shows the integration site of an individual mutant, identified by inverse PCR. (B) mRNAs that were expressed in mutants were examined using a different set of primers to classify patterns of transcript formation. (C) Each form of fusion mRNAs that is shown in B was examined in a quantitative manner by q-RT-PCR. The splice acceptor (SA) of the *Ghr* exon 2 was used frequently with the HIV-1 major splice donor (MSD) site for virus-host fusion transcripts that were expressed in the isolated IL-3 independent mutants. This bar graph shows the relative copy number normalised by the number of mouse beta actin copy number.

3.2 Aims

1. **Construction of a novel LV** with a strong splice acceptor (SA) site introduced upstream of an IRES-transgene in a promoterless self-inactivating (SIN) vector.
2. Measure the titre of this vector, both by detection of transgene expression and by vector genome integration in the host genome, these parameters to be compared with a control vector with an internal promoter driving transgene expression. The prediction would be that both vectors would have similar titre measured by genome integration, but the test vector would have a much lower titre when transgene expression is measured. This is because such expression will only occur in the small number of integrants adjacent to a cellular promoter, driving expression of a fusion transcript.
3. Optimal puromycin culture condition for selection of puromycin resistant clones that may transform into cytokine-independent in the IM assay (see Chapter 4).

3.3 Results

3.3.1 Construction of splice-in lentiviral vectors

All vector constructs in this study are shown in Fig.3-2 (Materials and Methods (M.M) 2.2) and the splice-in model vector is pGhr IRES-Em/Puro (Fig.3-2C), designed to express a marker gene by splice-in.

pGhr IRES-Em came from pSFFV-Em (Fig.3-2A) (Demaison et al., 2002). This original vector is a bicistronic vector that contains a spleen focus forming virus (SFFV) promoter, followed by IRES downstream to control transcription of Em (Emerald, Em, mutation-enhanced GFP). The SIN design in its backbone was used to construct pGhr IRES-Em. In order to make the vector promoter-less, the SFFV promoter was removed. Then the *Ghr* exon 2 locus (150 nt of intron 1 upstream of exon 2 and 56 nt of the beginning site of exon 2) was amplified by PCR from genomic DNA (gDNA) of Bcl15 cells and was introduced with the three stop codons, followed by IRES. This vector design has the potential strong SA site in the SIN LV backbone. A marker gene Em or Puro (puromycin resistant gene)) within the vector is located downstream of the *Ghr* exon 2 SA. The marker gene expression is under the control of an internal ribosomal entry site (IRES). Theoretically, the marker gene may be expressed when there is a host promoter upstream of vector integration site in the same orientation relative to the direction of the integrated vector. Under this condition, the marker gene can be expressed and a host SD can select an SA within the vector to generate a fusion mRNA. Between the introduced *Ghr* exon 2 and the IRES, three stop codons to cover three reading frames were introduced. It is assumed that the separation of host gene open reading frame (ORF) from the marker gene ORF by introducing these stop codons might avoid any host-vector fusion protein. The final construct was sequenced by primers (Ghr Fw: CTCGGAATTCTGAAG TCTGAGGC, BamHI Ghr Rc: CTCGGAATTCTGAAGTCTGAGGC, Ghr Rc: CAGTCTCGAGCTAGTTATTCAGGATCCCTGGTGACTGCCAGTGCCAA).
pGhr IRES-Em was utilised to construct pGhr IRES-Puro by replacing Em with a PCR fragment encoding the puromycin resistance gene. The final construct was

sequenced by the primer (Puro Fw: AGTC TCATGACCGAGTACAAGCC, Puro Rc: ATGCGGCCGCTAATTCCC TCGAGTCA).

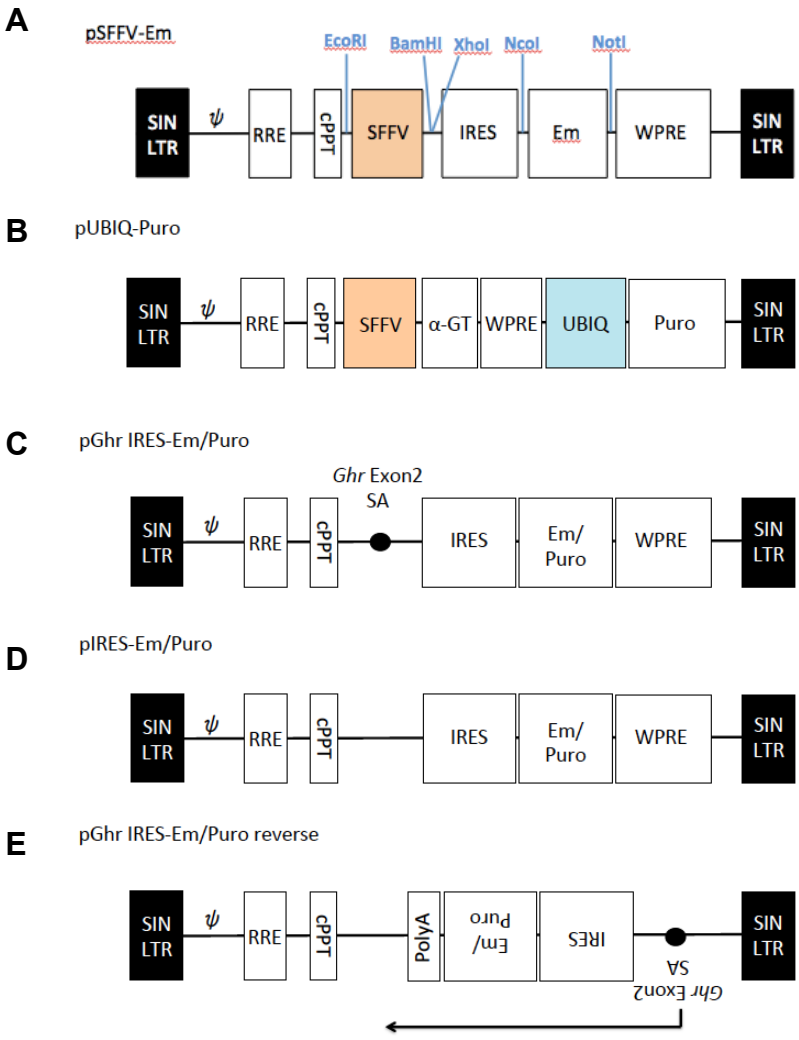


Fig.3-2 The lentiviral vector constructs that were used in the PhD project.

The provirus form of each vector is shown. All vectors have self-inactivating long terminal repeats (SIN-LTRs) with deletions in the U3 (the promoter and enhancer regions). (A) pSFFV-Em has a spleen focus-forming virus (SFFV) promoter to express Emerald GFP (Em) that was used as a positive control for Em expression. Restriction enzyme sites important to construct other model vectors (C, D, E) are indicated on the vector in blue. (B) pUBIQ-Puro has a ubiquitin promoter to express puromycin resistance gene (puroR) and was used as a positive control for puromycin expression

(Escors et al., 2008). (C) pGhr IRES-Em/Puro is a derivative vector of pSFFV-Em. In order to allow gene transcription by an upstream host promoter, pGhr IRES-Em/Puro does not have an internal promoter, but an internal ribosomal entry site (IRES) that controls the marker protein translation. The *Ghr* exon 2 splice acceptor (SA) with its surrounding genomic sequence was introduced upstream of the IRES (150 nt of intron 1 upstream of the exon 2 and 56 nt of the beginning site of the exon 2). (D) pIRES-Em/Puro is a derivative vector of pGhr IRES with deletion of the *Ghr* exon 2 SA. This vector was used to observe background gene expression by a post-transcriptional mechanism that was unrelated to the *Ghr* exon 2 SA. (E) pGhr IRES-Em/Puro reverse is another derivative vector of pGhr IRES. This vector has reverse-oriented sequence from downstream the central polypurine tract (cPPT) to upstream of the vector 3'LTR. This vector design has a potential to investigate different chimeric transcripts compared to the pGhr IRES vector.

SFFV, spleen focus forming virus; LTR, long terminal repeat; IRES, internal ribosomal entry site; WPRE, Woodchuck Hepatitis Virus Posttranscriptional Regulatory Element; SIN-LTR, self-inactivating LTR; RRE, rev response element; polyA, polyadenylation signal; α–GT, alpha-galactosyltransferase.

To confirm the function of the SA introduced to increase transgene expression in pGhr IRES-Em/Puro, pIRES-Em/Puro was designed as a control vector based on pGhr IRES-Em/Puro. This vector does not contain the *Ghr* exon 2 SA (Fig.3-2D). We hypothesised that this SA-less vector would express a marker gene less than the SA vector because less host-vector fusion mRNAs would be generated by splice-in. Additionally, this SA-less vector might give an insight into the background expression of a marker gene that is not related to the *Ghr* exon 2 SA, such as splice-in via cryptic SAs in the vector backbone or read-through transcripts without splicing. The final construct was sequenced by primers (DelSA F：GCAGTTAATCCTGGCCTG、BamHI cPPT Rc：TCAGGGATCCGA TATCAAGCTTATC) for its confirmation.

pGhr IRES-Em/Puro reverse was designed based on pGhr IRES-Em/Puro, as

another candidate vector to study splice-in (Fig.3-2E). The vector sequence flanked by two EcoRI sites, the downstream of cPPT to upstream of 3'LTR is reversed. This vector had a potential to test the effect of the reversed vector sequence on fusion mRNA generation with host sequence and marker gene expression. Initially, the SV40 late polyA signal was introduced downstream of a marker gene to terminate the transcript as well as introducing another EcoRI site. This intermediate vector was then digested with EcoRI to obtain the vector with the reverse-oriented vector sequence. The final construct was sequenced by primers (Em Fw: ATGGTGAGCAAGGGCGAG, Puro Fw: AGTCTCATGACCG AGTACAAGCC, Ghr Rc: CAGTCTCGAGCTAGTTATTCAGGATCCCTGGTG ACTGCCAGTGCCAA) for its confirmation.

3.3.2 *In silico* prediction of splice sites within model vectors showed the introduced *Ghr* exon 2 SA was a potential strong SA

The provirus sequence of each promoter-less puromycin vector was tested in the *in silico* programme NetGene2 (http://www.cbs.dtu.dk/services/NetGene2/) to confirm that the *Ghr* exon 2 SA was a possible dominant SA site in the vector provirus sequence (Fig.3-3) (M.M 2.3.11). The number of predicted SA and SD sites in each vector was counted and summarised (Fig.3-3A). The number of SA sites is relatively larger than that of SD sites in each test vectors. When looked at each predicted SA, the *Ghr* exon 2 SA was predicted as the strongest SA in the provirus sequence of pGhr IRES-Puro. The *Ghr* exon 2 SA showed the highest confidence number (0.69) out of other 21 SA candidate sites (Fig.3-3B). On the other hand, the pIRES-Puro vector does not contain any outstanding SA sites (< 0.33). pGhr IRES-Puro reverse showed the introduced SA as the strongest SA in the negative strand of the provirus sequence, as expected.

A

	Splice acceptor (S)	Splice acceptor (AS)	Splice donor (S)	Splice donor (AS)
pCSGW Ghr IRES-Puro	21	16	8	4
pCSGW IRES-Puro	19	16	7	4
pCSGW Ghr IRES-Puro SV40 rev	23	8	7	8

B

	confidence	pos 5'->3'	phase	strand	5' intron exon 3'
pCSGW Ghr IRES-Puro	0.69	1967	0	+	TGTCTTGCAG^GTCTCAGGTA
	0.33	3860	2	+	TCGCCCTCAG^ACGAGTCGGA
	0.26	1507	2	+	ATCGTTTCAG^ACCCACCTCC
pCSGW IRES-Puro	0.33	3651	2	+	TCGCCCTCAG^ACGAGTCGGA
	0.26	1507	2	+	ATCGTTTCAG^ACCCACCTCC
	0.25	2105	1	+	CCCTTTGCAG^GCAGCGGAAC
pCSGW Ghr IRES-Puro SV40 rev	0.69	3226	0	-	TGTCTTGCAG^GTCTCAGGTA
	0.43	376	0	-	CTCCTTCTAG^CCTCCGCTAG
	0.39	1234	2	-	CTCCATCCAG^GTCGTGTGAT

Fig.3-3 *in silico* prediction of the *Ghr* exon 2 SA as a potential strong SA in vector provirus sequence.

Using the NetGene2 (http://www.cbs.dtu.dk/services/NetGene2/), potential splice sites (SD and SA) on both DNA strands were detected on the provirus sequences of pGhr IRES-Puro, pIRES-Puro and pGhr IRES-Puro reverse. (A) The predicted total number of SA and SD of each vector was summarised. S, a sense strand; AS, an antisense strand (B) The top three SAs with a highest confident number of each vector were extracted. The row that is highlighted in red is the *Ghr* exon 2 SA.

3.3.3 The constructed promoterless vectors were characterised by measuring vector infectivity and marker gene expression

Prior to the test of vector infectivity by measuring vector DNA transfer and transgene expression in transduced cells, a VSV-G pseudotyped virus containing promoter-less vector genome was produced by HEK293T cells by transient transfection of three plasmids (M.M 2.4.1). RNA genome of this model vector is driven by a cytomegalovirus (CMV) promoter that leads the robust Em

expression, which was confirmed in the virus producer cells of pSFFV-Em, pGhr IRES-Em and pIRES-Em at 72 hours post-transfection (Fig.3-4).

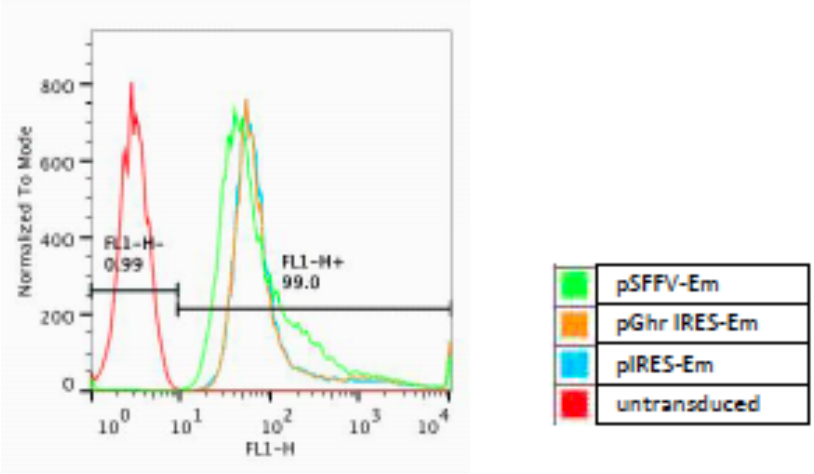


Fig.3-4 FACS analysis of vector producer cells

Virus producer cells (HEK293T) of pSFFV/pGhrIRES/pIRES-Em were analysed by FACS 72 hours post-transfection (M.M 2.4.1). The histogram shows the distribution of Em expression intensity of transfected cells by each vector. X axis shows Em expression, and Y-axis shows the number of live cells being analysed. Each colour in the histogram indicates the transfection result of each vector as listed in the small table next to the histogram. The Cells were fixed in 2 % PFA prior to the detection by FACS.

The scheme of vector titration was described in the Fig.3-5. The serially diluted virus supernatant was transduced on HEK293T or murine pro-B cell line, Bcl15 cells (M.M 2.4.3). HEK293T have higher transduction efficiency than Bcl15 cells. The vector infectivity was tested by two titration methods: the titre based on vector DNA transfer and marker gene expression. Depending on the marker gene in the vector, the method to measure the protein expression differed. For an Em vector construct, fluorescence activated cell sorting (FACS) was employed to measure the transgene expression at 48 hours post-transduction (M.M 2.4.3.1). On the other hand, for a puromycin vector construct, drug

selection of transduced cells (1 µg/ml puromycin) was performed (M.M 2.4.3.3). Choosing in the linear range of the number of puroR clones under certain virus inputs (µl), the occurrence of puroR cells per virus dose was calculated as a titre of puromycin vectors.

The marker gene expression of the promoter-less vector would rely on the presence of a host cellular promoter upstream of vector integration. Therefore, the FACS-based titre of the promoter-less construct would be expected to be much lower than the titre estimated by vector genome transfer detected by SYBR green qPCR (Qiagen) (M.M 2.4.3.2). For this measurement, in order to eliminate unintegrated episomal vector DNA transduced cells were kept cultured for 7 to 10 days post-transduction, then gDNA was extracted. The vector copy number of 100 ng gDNA was estimated by a standard curve of HIV-1 leader sequence of SIN pHV. Five points of known vector copy numbers of SIN pHV are plotted against the corresponding Ct values (Fig.3-6, top). Using the standard curve, experimentally detected amplification of each sample estimates a vector copy number. As a representative result of vector transduction efficiency measured by qPCR, the similar efficiency of vector DNA transfer was achieved among the tested vectors (Fig.3-6, bottom).

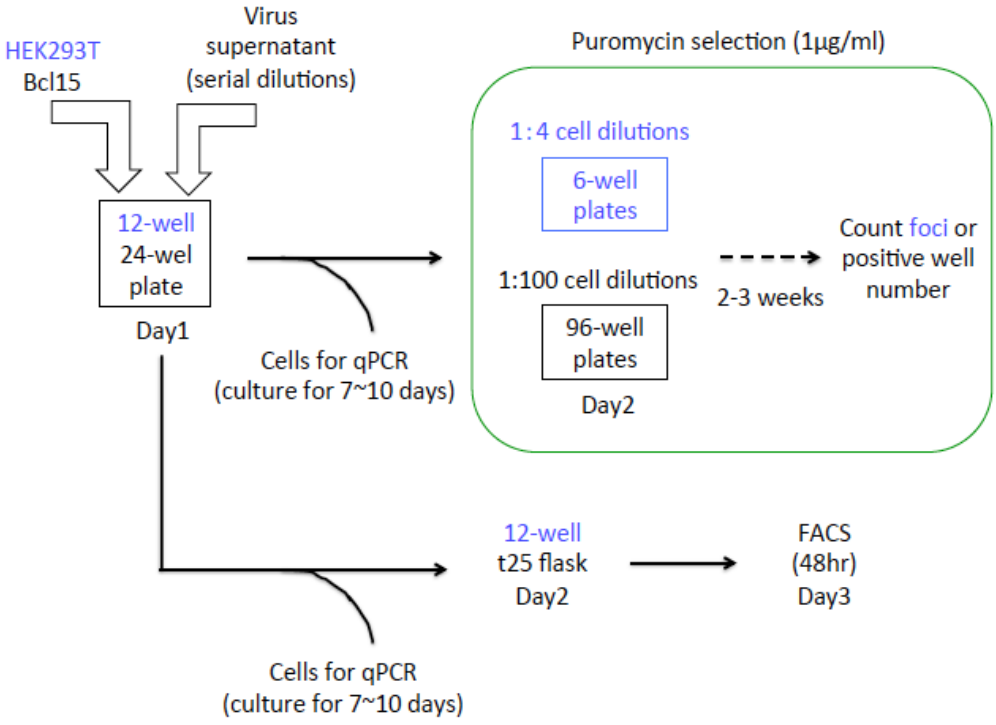
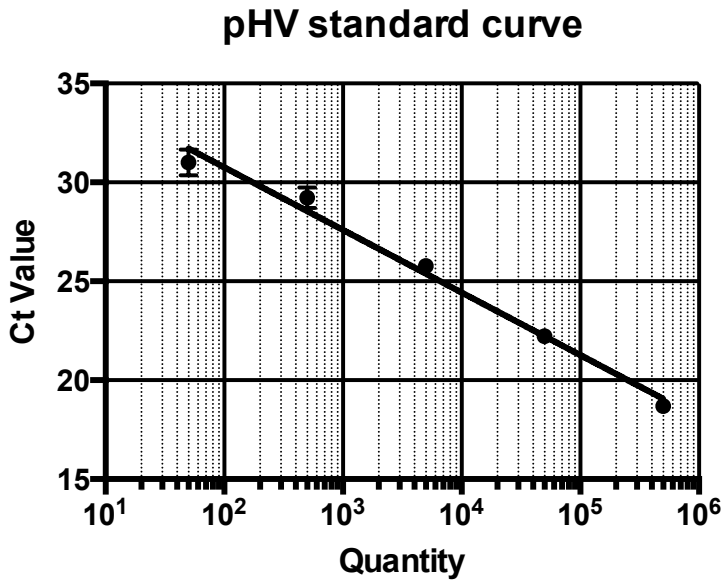


Fig.3-5 The experimental protocol of vector titration in HEK293T or Bcl15 cells for estimating virus infectivity.

pSFFV-Em and pGhr IRES-Em or pUBIQ-Puro and pGhr IRES-Puro were titrated to estimate the efficiency of vector genome transfer by qPCR and transgene expression following the successful gene transfer by FACS or puromycin selection (M.M 2.4.3). The serially diluted virus supernatant was transduced. In Bcl15 cells, 24 hrs after the transduction, puromycin selection was initiated on transduced cells by 1:100 dilution. The cells were kept cultured in puromycin medium (1 μg/ml) for 2 to 3 weeks to obtain puro-resistant clones in 96-well plates. For Em vectors, Em expression was detected by FACS 48-hour post-transcription. Prior to the detection, cells were fixed in 2 % paraformaldehyde (PFA) solution (Materials and Methods). Upon both vector transduction, genomic DNA (gDNA) was extracted from transduced cells that were cultured for 7 to 10 days post-transduction to avoid the detection of non-integrated episome DNA by SYBR green qPCR (Qiagen). The experimental condition specific for HEK293T cells was highlighted in blue.



Vector construct	pHV copies	gDNA input (ng)	Bcl15 cell number in 100ng gDNA input	MOI	Virus input (µl)	Initial cell number	Titre (IU/ml)
pSFFV-Em	41109.63672	100	15911.31172	2.58	250	96250	994,714
pGhr IRES-Em	45125.20313	100	15911.31172	2.84	250	96250	1,091,877
pIRES-Em	73085.96094	100	15911.31172	4.59	250	96250	1,768,433

Fig.3-6 Titration of vector based on measuring gene transfer by SYBR green qPCR.

The standard curve was drawn by plotting Ct value for amplification of HIV-1 leader sequence on the Y axis against the known copy numbers of the SIN pHV plasmid on the X axis (M.M 2.4.3.2). The standard curve allows the calculation of the vector copy number of each test sample according to an amplification result shown in Ct values. Once the raw vector copy number was obtained, the cell number corresponding to the gDNA input (100 ng) was estimated based on genome size (bp) of a mouse. Using the estimated cell number in 100 ng gDNA and the raw copy number of pHV, the vector copy number per cell was calculated and defined as a multiplicity of infection (MOI). Using the transduction condition of each sample (virus input (µl) and an initial cell number), the vector titre was calculated.

3.3.4 Promoter-less SIN LV-Em vectors expressed lower levels of marker genes

In order to characterise the marker gene expression, the vector was titrated on HEK293T cells. As a positive control, pSFFV-Em with a strong promoter driving constant expression of Em was used.

The percentage of live cells with Em expression was detected by FACS (Fig.3-7A). The percentage increases in a dose-dependent manner for all three vectors, though that with pSFFV-Em is highest. The mean fluorescence intensity (MFI) of marker gene expressing cells was also compared to the three vectors (Fig.3-7B). The MFI in transduced cells by pSFFV-Em was much higher than that of promoter-less vectors at each virus dose (μl) as expected.

As also described in Fig.3-6 on Bcl15 cells, vector integration upon transduction was similar in HEK293T cells for both promoter-less Em vectors and pSFFV-Em (Fig.3-7C). However, the titre measured by GFP expression was much lower for the two promoterless vectors. Nearly 85-fold reduction in the FACS titre was observed in transduction of pGhr IRES-Em compared to pSFFV-Em (Fig.3-7D). Transduction of pIRES-Em showed a further reduction in the FACS titre compared to pSFFV-Em (116-fold) (Fig.3-7D).

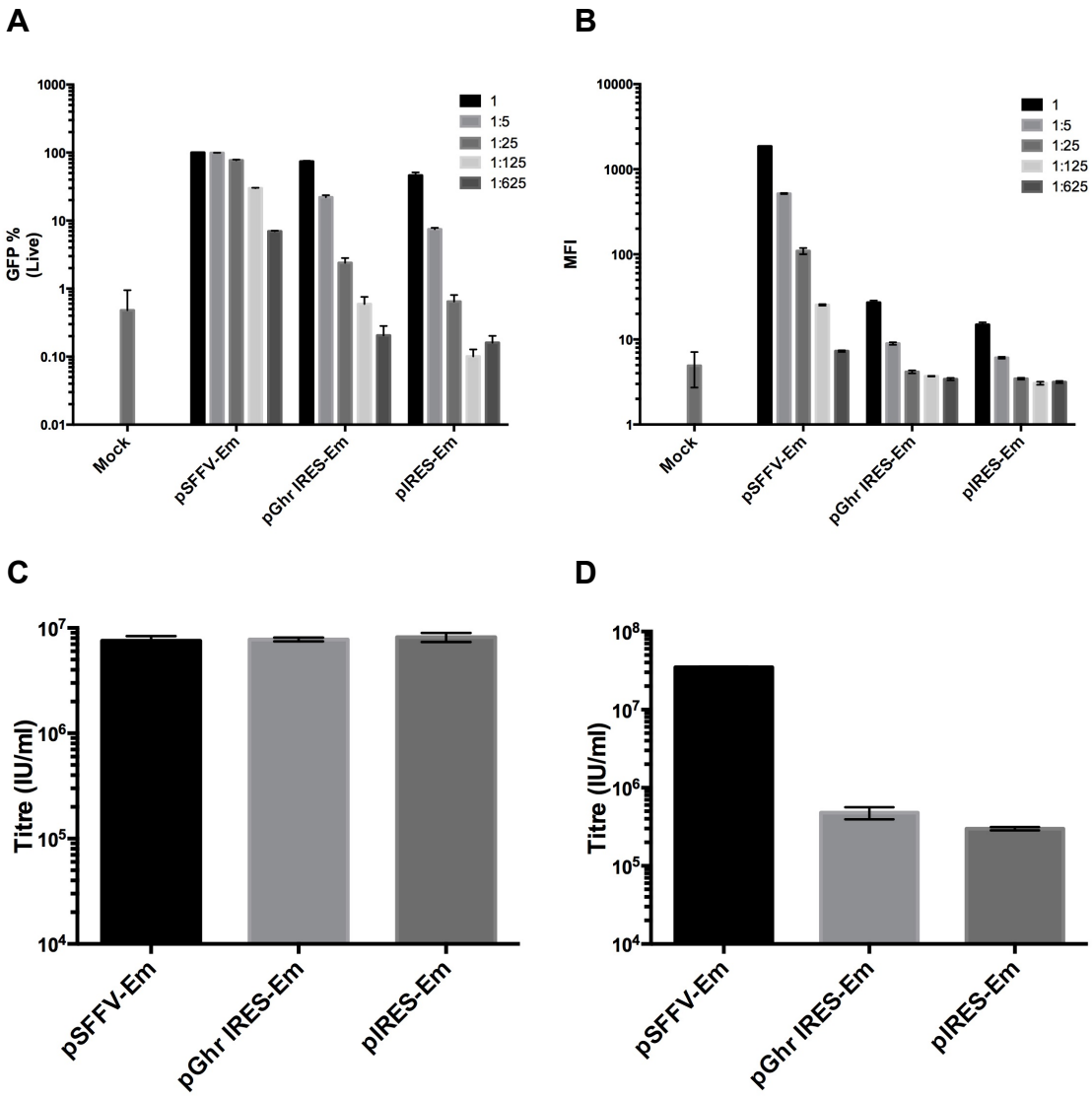


Fig.3-7 Vector transduction in HEK293T cells to estimate virus infectivity

(A) Using a virus harvest of pSFFV/pGhrIRES/pIRES-Em by transient transfection on HEK293T cells, each virus was titrated on HEK293T cells. Em positive cells (%) in live cells in comparison to mock-transduced cells were calculated by FlowJo software (http://www.flowjo.com/). (B) Mean fluorescent intensity (MFI) of Em positive cells in each virus dilution was analysed by FlowJo software (http://www.flowjo.com/). The shift of MFI of each vector was compared. (C) Each vector titre based on vector copy number in genome detected by qPCR was compared. (D) Each vector titre based on Em protein expression detected by FACS was compared. Each experiment was performed in duplicate. Error bar in each bar graph shows the standard deviation of between duplicates (FACS) or triplicate (qPCR) in a single experiment.

3.3.5 Bcl15 showed lower marker gene expression than HEK293T cells, however the same trend of vector DNA transfer and marker gene expression

Bcl15 cells were transduced by the same vector constructs tested in HEK293T cells in 3.3.4 because this cell line was to be used for screening transduced cells with marker gene expression in the future assay described in Chapter 4. This cell line is known for having lower transduction efficiency compared to HEK293T cells (Bokhoven et al., 2009), accordingly, reduced marker gene expression in Bcl15 cells was also expected in our experiment.

The percentage of cells with Em expression was detected by FACS (Fig.3-8A). The percentage increases in a dose-dependent manner in all three tested vectors. When the MFI at highest virus dose among test vectors was compared, pSFFV-Em showed 55-fold higher than pGhr IRES-Em, and 44-fold higher than pIRES-Em (Fig.3-8B).

As it was observed in the Fig.3-6 (bottom), the promoterless vectors showed a high qPCR titre (1.09×10^6 (IU/ml) for pGhr IRES-Em and 9.95×10^5 (IU/ml) for pIRES-Em) similar to that of pSFFV-Em (1.77×10^6 (IU/ml)) (Fig.3-8C). Also, similarly to HEK293T cells, Bcl15 cells transduced by pGhr IRES-Em showed lower Em expression (34-fold reduction) compared to that of pSFFV-Em (Fig.3-8D). In addition, the Em expression by transduction of pIRES-Em was also lower than that of pSFFV-Em (20-fold reduction) (Fig.3-8D). The observation of lower marker gene expression of promoter-less vector compared to the positive is consistent with the results in HEK293T cells. While, a difference of Em expression between pGhr IRES-Em and pIRES-Em was not as distinct as transduction in HEK293T cells.

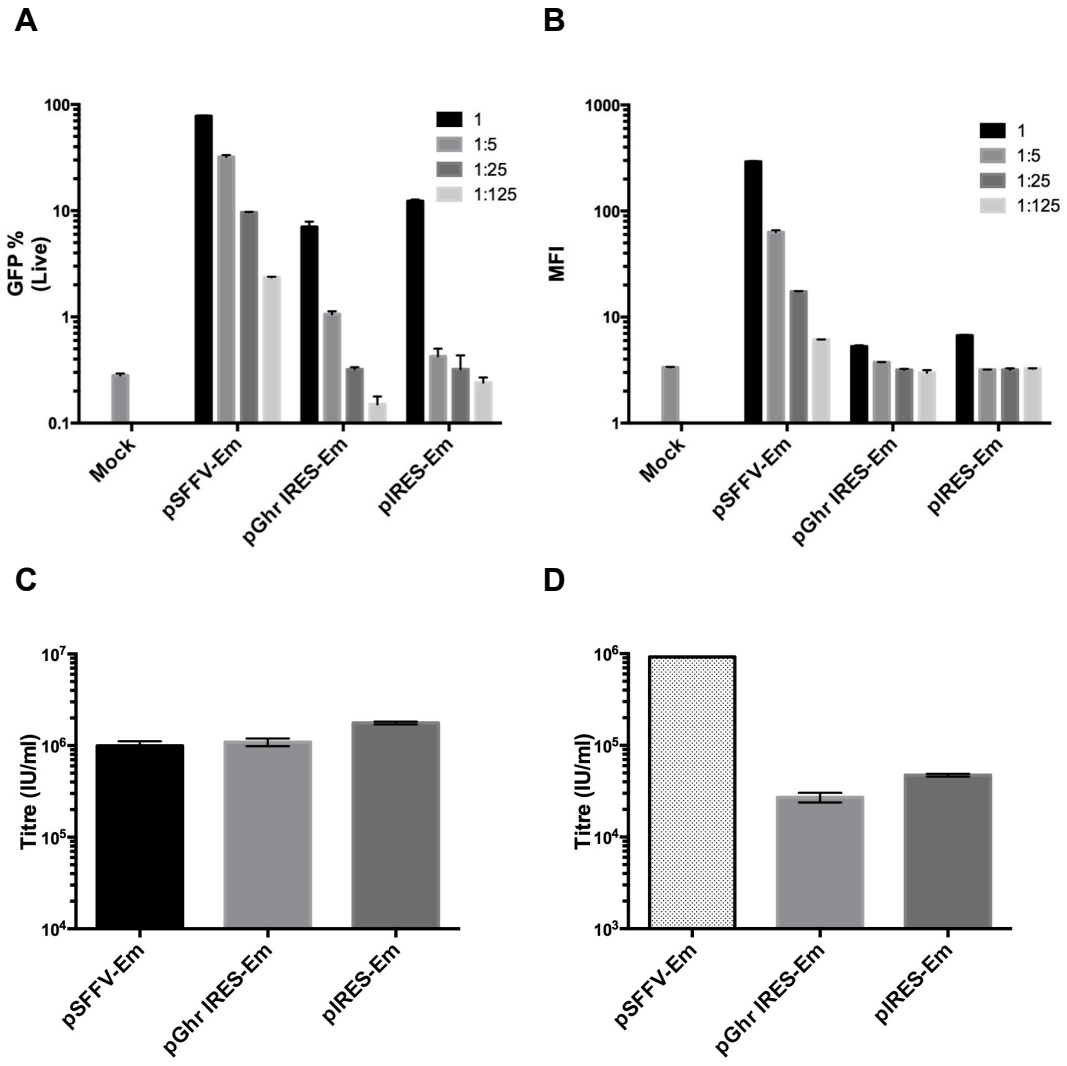


Fig.3-8 Vector titration in Bcl15 cells to estimate virus infectivity

(A) Using the same virus harvested used in HEK293T cells (Fig.3-6), the virus was titrated on Bcl15 cells. Em positive cells (%) in live cells in comparison to mock-transduced cells were calculated by FlowJo software (http://www.flowjo.com/). (B) Mean fluorescent intensity (MFI) of Em positive cells in each virus dilution was analysed by FlowJo software (http://www.flowjo.com/). The shift of MFI of each vector was compared. (C) Each vector titre based on vector copy number in genome detected by qPCR was compared. (D) Each vector titre based on Em protein expression detected by FACS was compared. Each experiment was performed in duplicate. Error bar in each bar graph shows standard deviation between duplicates (FACS) or triplicate (qPCR) in a single experiment.

3.3.6 Puromycin drug concentration optimisation in Bcl15 cells

We proposed screening potential splice-in fusion mRNAs being a cause of mutagenesis in transduced cells. We thought that selection of transduced cells with fusion mRNAs by the addition of drug into the culture medium would be an easier method than cell sorting for Em expression. Therefore, puromycin drug selection was performed on cells transduced by puromycin constructs. An optimal puromycin concentration should be sufficient to eliminate all untransduced parental Bcl15 cells. In addition, this elimination of untransduced background parental cells might reduce the number of spontaneous mutants that were irrelevant to vector integrations (Bokhoven et al., 2009). To decide optimal puromycin concentration, initially, 3.35 x 10^4 parental Bcl15 cells were treated with puromycin at different concentrations (0.1 to 0.8 μg/ml) and cell growth was monitored under a microscope. Live cell number was counted by trypan blue staining (Gibco) on day 7 after the drug treatment was initiated (Fig.3-9). The bar graph showed that 0.5 μg/ml puromycin concentration killed most of the background cells compared to lower puromycin concentrations (0.1 to 0.4 μg/ml).

We then tested the puromycin concentration required for selection of vector transduced clones. Puromycin vectors (pGhr IRES-Puro and pUBIQ-Puro) were titrated by culturing transduced cells under a range of puromycin concentration (0.5 to 1 μg/ml) (Fig.3-10A). The positive control vector pUBIQ-Puro was titrated at three puromycin concentrations, 0.6, 0.8 and 1 μg/ml. At the 0.6 μg/ml puromycin concentration, the number of puroR clones of pGhr IRES-Puro transduction was more than 10 times less than that of pUBIQ-Puro. The number of puroR clones generated by transduction of pUBIQ-Puro did not change largely over the tested puromycin concentrations (0.6, 0.8 and 1 μg/ml) (Fig.3-10A). The number of puroR clones generated by pGhr IRES-Puro showed a decrease in the number of clones as the puromycin concentration increased. qPCR performed on transduced cells without puromycin selection showed that vector DNA transfer was similarly efficient on both pGhr IRES-Puro and pUBIQ-Puro (Fig.3-10B).

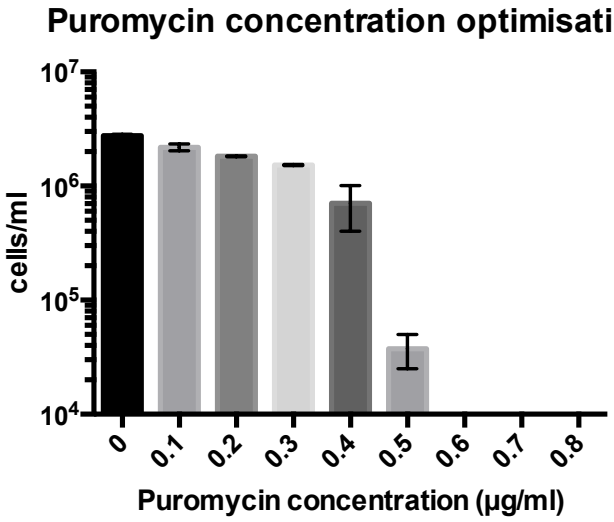


Fig.3-9 Bcl15 cells cultured in different concentrations of puromycin.

3.35 x 10^4 Bcl15 cells were cultured in puromycin at a series of concentrations (0.1-0.8 µg/ml). The live cell numbers were counted on day 7 using trypan blue (Gibco). The X axis of the bar graph shows puromycin concentration and the Y axis shows cell concentration (cells/ml). Cell culture was performed in duplicate per puromycin concentration. Error bar in each bar graph shows standard error of mean (SEM) between two independent experiments. The live cell number below the detection limit (10^4) is not shown in the graph.

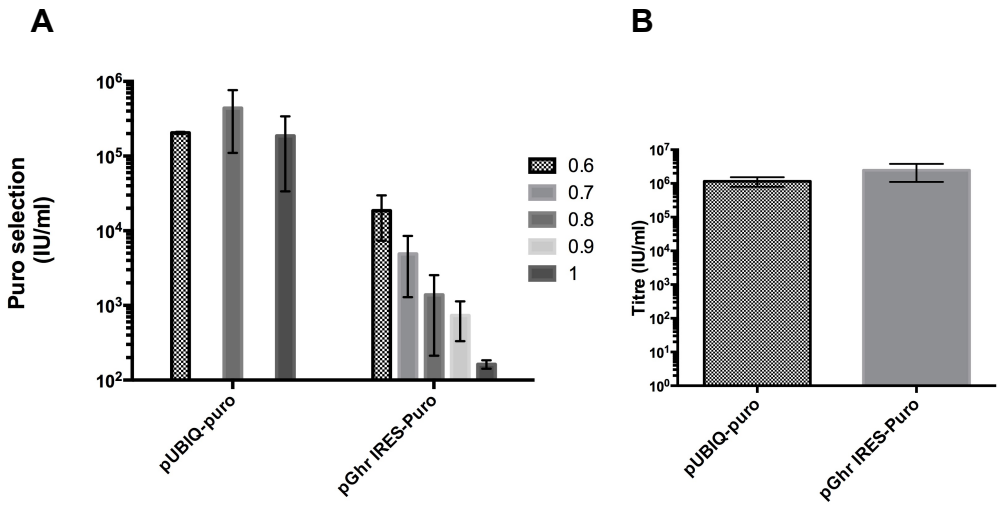


Fig.3-10 Vector titration of puromycin constructs by various puromycin concentrations (0.5-1 mg/ml).

(A) Vector titration based on puromycin resistance was carried out in Bcl15 cells (M.M 2.4.3.3). The X axis of the bar graph shows the puromycin concentrations (μg/ml) and the Y axis shows the titre calculated based on the number of puromycin resistant (puroR) clones that were emerged in each puromycin culture. The positive control (pUBIQ-Puro) was tested in the three puro concentrations (0.6, 0.8, 1 μg/ml). (B) Separately, vector infectivity was estimated by qPCR. Error bar represents standard error of mean (SEM) of duplicate (puro selection) and triplicate (qPCR) between two independent experiments.

3.3.7 Transduction by promoterless puromycin vectors resulted in efficient vector DNA transfer but lower puromycin expression

Puromycin vectors were then titrated on both HEK293T (Fig.3-11A, B) and Bcl15 cells (Fig.3-11C, D) using the vector titration protocol as described in the Fig.3-5. and 1 μg/ml puromycin for selection. Upon transduction of different cell lines, the vector genome titre was very similar for the three vectors, though higher in the HEK293T cells (Fig.3-11A and C, blue bars). HEK293T cells transduced by the promoterless vectors showed the reduced number of puroR clones, which resulted in lower puroR titre compared to that of pUBIQ-Puro: 145-fold reduction in pGhr IRES-Puro, and 545-fold in pIRES-Puro (Fig.3-11B). When those puroR constructs were titrated on Bcl15 cells, the puroR titre of pGhr IRES-Puro

showed 570-fold reduction and the puro titre of pIRES-Puro showed 5412-fold reduction compared to that of pUBIQ-puro (Fig.3-11D). These results are consistent with the result obtained by transduction of Em vector constructs (Fig.3-7C, D).

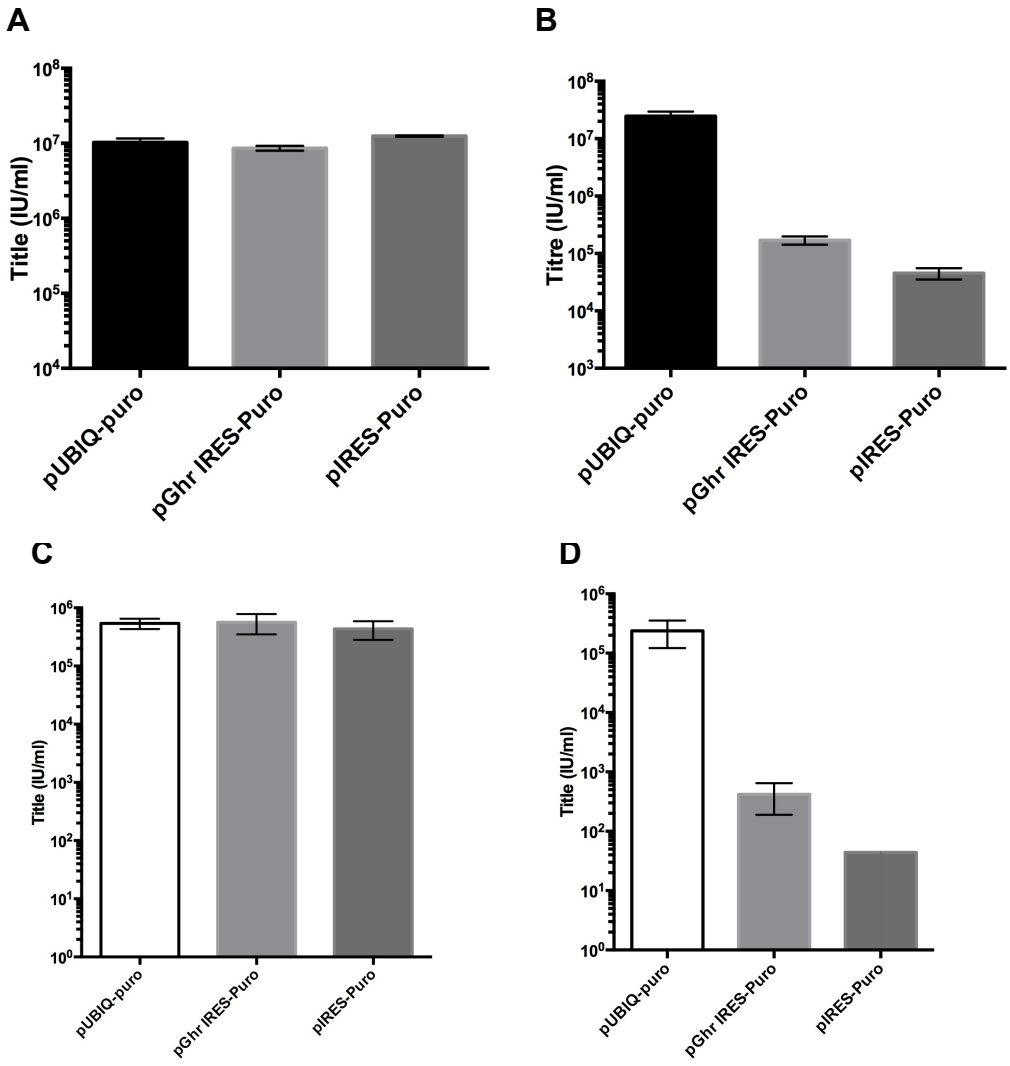


Fig.3-11 Puromycin vector titration on HEK293T or Bcl15 cells to estimate the vector infectivity

Puromycin vectors were titrated in HEK 293T (A, B) cells and Bcl15 (C, D) cells based on puromycin selection (B, D) and qPCR (A, C) to observe the difference in puromycin-resistant protein expression and vector gene transfer into host cells. qPCR was performed in HEK293T and Bcl15 cells, respectively. The error bar on each bar graph represents standard error of mean (SEM) of duplicate (puro selection) and triplicate (qPCR) between three (HEK293T cells) or two (Bcl15 cells) independent experiments. An error bar could not be drawn in pIRES-Puro (D) because one experiment did not generate any puroR cells in the assay protocol.

3.4 Discussions

This chapter tested the virus infectivity and transgene expression of our model vectors to see if they functioned as we expected. Transgene expression in promoterless vectors relied on the presence of host cellular promoter therefore we hypothesised that most transduced cells would be silent for transgene expression, which was observed in transduction experiments (Fig.3-7D, Fig.3-8D, and Fig.3-11B, D). Titration of promoter-less vectors (pGhr IRES-Em/Puro and pIRES-Em/Puro) along with positive controls (pSFFV-Em and pUBIQ-Puro) on HEK293T and Bcl15 cells showed consistent results independent of a marker gene, or cell lines. As a next step, we focused on transduced cells with marker gene expression to investigate the mechanism of potential genotoxic events via splice-in in the future mutagenesis assay. Vector transduction by puroR vector constructs was the most relevant experiment for the next chapter because the selected puroR cells could generate cytokine-independent mutants. We postulated that puroR cells with strong puromycin resistance could potentially transform into cytokine-independent cells. Such selection was achievable under a high concentration of puromycin. The vector titration on the Bcl15 cells with various puromycin concentrations showed that 1 μg/ml was an optimal puromycin concentration to select puroR cells with strong puromycin resistance.

3.4.1 The *Ghr* exon 2 SA was detected as a potential strong SA

The promoter-less vector designed for inducing host-vector chimeric transcripts via splice-in was successfully constructed (Fig.3-2). The initial assessment for the designed promoterless vectors was performed by NetGene2 *in silico* splice site finder. Using this sequence-based splice site prediction, the *Ghr* exon 2 SA derived from Bcl15 cellular sequence was shown as the strongest SA in each promoter-less vector provirus sequence (Fig.3-3B). Other splice sites, either known or cryptic, also includes splice sites that play an important role for structural and accessory gene expression in HIV-1 replication such as HIV-1

major splice donor SD1 (343) and the known SA downstream of RRE sequence SA7 (1507) (see 1.6.6.4, the main introduction). These splice sites can potentially function as a splice site with a host cellular sequence to generate fusion transcripts and alter exogenous gene and even protein expression as observed in the *HMGA2* case (Cavazzana-Calvo et al., 2010). The actual use of the identified splice sites by NetGene2 can be confirmed by examining the sequence of fusion mRNAs expressed in transduced cells. Therefore, the presence of those listed splice sites should be also inspected in isolated host-vector chimeric transcripts in the later study (Chapter 5 and 6). If any of those cryptic splice sites are mainly used in the fusion transcripts, they can be mapped and mutated for instance to test that occurrence of fusion mRNAs will be reduced as reported previously (Cesana et al., 2012) (Moiani et al., 2012). Such sites can be ranked by the number of use on fusion mRNAs. This can indicate the strength of each identified splice site by times of use and be compared with the *in silico* result (Fig.3-3).

3.4.2 Reduced transgene expression in transduced cells by the promoterless model vector

In all transduction results of promoterless vectors reproducibly showed efficient vector DNA delivery on host cells and low marker gene expression because of the promoterless design. One possible explanation for this is the low frequency for the integrated promoterless vector to encounter a host cellular promoter upstream in the same orientation relative to the transcriptional direction (Fig.3-7D, Fig.3-8D, and Fig.3-11B, D). The relatively random distribution of lentiviral vector integration sites (Berry et al., 2006) would also contribute to this effect. In addition, in the host cell nonsense-mediated decay pathway senses premature termination codons and degrades products (Chang et al., 2007). This could be another possible cause to keep the level of host-vector fusion mRNA low, which was shown in the assay using the inhibitor of degradation by the pathway in the context of ß Thalassaemia gene therapy (Moiani et al., 2012).

Although the molecular mechanisms of this transgene expression cannot be concluded from these titration results, it is postulated that the chimeric transcripts expressed in transduced cells may contain both splice-in and read-through forms (Cesana et al., 2012) (Moiani et al., 2012) (Knight et al., 2010) and contribute to marker gene expression. Even splice-in with a host SD at a different SA within the model vector could be observed. This aspect was investigated in the following Chapter 5 and 6 by identifying the fusion transcripts expressed in transduced cells.

3.4.3 The introduced *Ghr* exon 2 SA may enhance transgene expression

Upon vector transduction, the pGhr IRES vector was tested to see whether it can increase marker gene expression in comparison to pIRES vector. Indeed, the SA vector showed enhancement in marker gene expression over the SA-less vector, based on the number of emerged puroR cells (Fig.3-7D and Fig.3-11B, D). While transduction of both pGhr and pIRES vector on Bcl15 cells did not show a clear difference in Em expression (Fig.3-8D). This could be due to less efficient transduction efficiency than HEK293T cells and for instance small difference in experimental condition might affect Em gene expression. However reproducible data about the small reduction of titre by transduction of pIRES vector than pGhr vector was demonstrated in different cell lines and different marker gene constructs. Therefore this is the most likely the trend about marker gene expression. The only difference between pGhr and pIRES vector is the sequence of the *Ghr* locus. Hence, this suggests that the presence of the introduced strong SA of *Ghr* exon 2 affects the formation of fusion transcripts by certain mechanisms, for instance, enhancing the interaction between the *Ghr* exon 2 SA with a host SD and induce splice0in between them.

3.4.4 Transduction of pGhr IRES-Puro generated puroR cells with various levels of puromycin resistance

When puroR titre at 0.5 μg/ml and 1 μg/ml by pGhr IRES-Puro transduction were compared, there was a nearly 100-fold reduction in the number of puroR clones (Fig.3-10A). This shows that the majority of puroR cells have a weak to moderate puromycin resistance. Then what possibly decides the strength of puromycin resistance of transduced cells? One possibility is the strength of a host promoter upstream of vector integration. If the host promoter is highly active in transcription, the expression level of fusion mRNAs could be relatively higher than other fusion RNA species transcribed by relatively lower active promoters. Another possibility is a vector integration site that could affect the molecular function of gene transcripts such as stabilising or enhancing expression as fusion forms. Such chimeric mRNA can persist in the transduced cells and could be a driving force of cell transformation.

However, this experiment lacks a possible negative control for pGhr IRES-Puro and that could be pIRES-Puro. Based on the result shown in Fig.3-11D, slightly lower titre based on puromycin selection by pIRES-Puro may be expected at higher puromycin concentration compared to the tire by pGhr IRES-Puro. A certain concentration of puromycin could separate puroR cells generated specifically by the *Ghr* exon 2 SA in pGhr IRES-Puro by testing pIRES-Puro in one experiment. If such concentration exists, the more optimal puromycin concentration could be selected, which could be beneficial to investigate splice-in fusion mRNAs. Therefore, generated puroR cells by transduction of pGhr IRES-Puro discussed above can contain host-vector fusion mRNAs irrelevant to the effect of the *Ghr* exon 2 SA such as read-through or splice-in with another SA located upstream of the *Ghr* exon 2 SA.

As a next step, we decided to use the promoter-less vector to generate puroR cells at first and then cytokine-independent mutants from the puroR cells. The IM assay performed previously in the literature used a lentiviral vector with intact LTRs with GFP marker gene and mutagenesis was quantitatively measured based on the number of mutants that had been transformed by vector integration and aberrant splicing. However, our study used SIN lentiviral vector with the puroR gene. This difference in the experimental condition made it difficult to

expect the number of mutants from a certain number of puroR cells. Therefore, based on the titre by puromycin selection (Fig.3-11B, D), the experimental conditions of IM assay (multiplicity of infection (MOI) and virus volume (μl)) were decided to obtain more than 10^4 puroR clones and initially focused on the isolation of IL-3 independent mutants.

Chapter 4

Insertional mutagenesis assay

4.1 Introduction

A number of methods have been developed to test the mutagenic potential of integrating viral vectors. Ideally, such a method needs to produce a quantitative measurement of the risk of host gene activation or silencing, so that vectors can be modified to reduce this. An early method involved the transduction of haematopoietic stem cells from tumour-prone *Cdkn2a*$^{-/-}$ mice by GRVs or LVs, followed by bone marrow transplant and quantification of tumour type and frequency (Montini et al., 2006). However, the time and expense of this assay then prompted an *in vitro* assay measuring plating efficiency and proliferation of mouse primary bone marrow cells (Modlich et al., 2006). Our group then developed a third assay using the IL-3 dependent mouse pre-B cell line BAF3 and its derivative Bcl15, which is engineered to express the survival gene Bcl2 (Bokhoven et al., 2009). The principle of this assay is that a vector insertion into a gene that allows survival or proliferation of the cells allows them to grow in the absence of IL-3. Use of a cell line rather than primary cells makes the assay cheaper and more reproducible. Bcl15 cells proved more robust upon IL-3 removal than the parental BAF3 cells and allowed the isolation of more mutants. The mechanism of LV mutagenesis in this assay was an insertion into the growth hormone receptor (*ghr*) gene first intron leading to upregulation of Ghr expression allowing cells to proliferate in growth hormone. The LV produces a fusion transcript encoding ghr using known, cryptic or introduced splice donor sites in the vector. So it was possible to use this assay to test clinical candidate vectors for the presence of splice donor sites and eliminate them when necessary (Montiel-Equihua et al., 2012) (Knight et al., 2012). During these experiments several IL-3 independent clones were isolated from Bcl15 cells that were not transduced with the vector. These “background” mutants all secreted IL-3 (Bokhoven et al., 2009).

The reported clinical trial using a LV to treat ß-thalassaemia found that a LV had caused expansion of a dominant hematopoietic cell clone by activating HMGA2 expression (Cavazzana-Calvo et al., 2010). In this case expression was activated by the production of a truncated *HMG2A* fusion transcript which splices

into a cryptic SA in the mutated U3 of the LV, this removes a 3' regulatory sequence from the *HMG2A* mRNA. My aim was to adapt the Bcl15 cell assay to detect any "splice in" fusion events that could make the cells IL-3 independent. If successful this could provide a quantitative measurement for clinical candidate vectors allowing the detection and elimination of splice acceptor sites, thereby improving their safety.

4.2 Aims

1. To use the vector pGhr IRES-Puro to transduce Bcl15 cells, then starve them of IL-3 followed by rescue with IL-3 re-addition to identify IL-3 independent mutants
2. To assess whether any IL-3 independent clones secrete IL-3 as this was previously identified as a common mechanism of background mutants
3. To identify vector integration sites in any IL-3 independent clones
4. To identify host-vector fusion transcripts in any IL-3 independent clones

4.3 Results

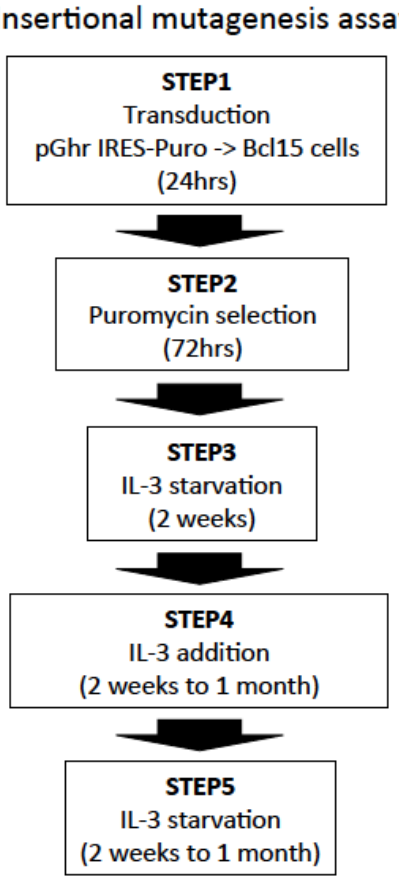
4.3.1 Insertional mutagenesis assay isolated an IL-3 independent mutant from transduced cells with a pGhr IRES Puro vector

To isolate IL-3 independent mutants generated by a “splice-in” from a host gene to the **pGhr IRES Puro vector**, Bcl15 cells were transduced, selected in puromycin then starved of IL-3 followed by an IL-3 rescue (M.M 2.5.1). Any cells that then appeared were once again starved of IL-3 to confirm IL-3 independence (Fig.4-1A). Initially Bcl15 cells transduced by pGhr IRES-Puro were cultured in bulk in 1 μg/ml puromycin for three days to select puromycin resistant (puroR) cell population. Such puromycin resistance should come from “splice in” from a cellular gene. To isolate IL-3 independent mutants from the puroR bulk populations, cells were washed then incubated at a maximum density of 1 to 1.4 x 10^5 cells/well in 24-well plates. This selection method was chosen because an estimated frequency similar to that of our previous “splice-in” assay would ensure that each well contained at most a single IL-3 independent cell (Bokhoven et al., 2009). To rescue potential IL-3 independent mutants WEHI3B supernatant was re-added to the culture medium after 14 days of IL-3 starvation. The rescued potential clones at this step (step 4 survivor) were used for testing up-regulation of IL-3 mRNA and 5’ RACE analysis for fusion transcripts. Once cells were expanded, IL-3 was washed away again and incubated to identify truly IL-3 independent mutants.

Two rounds of IM assay were performed and the results are summarised in Fig.4-1B. During the IM assay, a small aliquot of transduced cells was used for puromycin selection using a limiting dilution experiment in order to calculate the number of puroR clones generated by each transduction. 9.7 x 10^6 cells or 5 x 10^7 cells challenged with the vector yielded 1.13 x 10^4 or 8.3 x 10^4 puroR clones respectively in the two experiments. The number of isolated puroR clones from puromycin titration at the first and the second IM assay was 5 or 6 at the highest cell dilution, respectively. The five puroR clones from the first IM assay was used in Fig.4-2. At the rescue step by IL-3 re-addition, a 24-well plate was used for

each cell dilutions (1, 1:3 and 1:9). More than 50 % of the well became recovered when plated at 1.1 x 10^4 cells per well and the accurate number of IL-3 survivors of each IM assay is unknown. If the number of survivors had been accurately titrated, the number of IL-3 independent mutants would be estimated. However in the IM assays the cell density was not enough to obtain recovered wells in a linear range.

A



B

The summary result of IM assay

	The 1st IM assay	The 2nd IM assay	Isolation step
Initial cell number	9.7 x 10^6	5 x 10^7	STEP1
Virus input (ml)	57	143.8	
qPCR titre (IU/ml)	2.64 x 10^6	1.53 x 10^6	
MOI	15.5	4.4	
PuroR* clones	1.13 x 10^4	8.3 x 10^4	STEP2/3
IL-3I clones	1	0	STEP5

*This number of puroR was subjected to the IL-3 starvation.

Fig.4-1 The overview of the experimental plan and the results of Insertional Mutagenesis (IM) assay.

(A) The flow of IM assay is drawn. Since the transducing vector carried a puroR gene, cells were selected for puromycin resistance (puroR) (STEP 2). Puromycin was then kept in the culture through the following steps for mutant isolation (from STEP 2 to STEP 5). (B) The experimental conditions, the number of puroR clones and IL-3 mutants are summarised by each round of experiments. pGhr IRES-Puro virus harvest with known

infectivity determined by Q-PCR of the genome (2.64 x 10^6 (the first IM), 1.53 x 10^6 (the second IM) (IU/ml)) was used to transduce Bcl15 cells. Virus volume (ml) for transduction was designed to obtain 1 x 10^4 (the first IM) or 5 x 10^4 (the second IM) puromycin resistant (puroR) clones, based on preliminary titrations based on the selection of puroR cells under 1 μg/ml puromycin in the culture medium. At 24 hours post-transduction, cells were subjected to puromycin selection (1 μg/ml) for 3 days to select puroR cells. Under this puromycin selection, untransduced parental Bcl15 and weak puroR cells are susceptible to the puromycin concentration and are eliminated. On the day when the puromycin selection started, aliquots of transduced cells were plated in 96-well plates and cultured in 1 μg/ml puromycin culture medium to obtain titre based on the selection of puroR cells. In addition, an aliquot of transduced cells was cultured for 7 to 10 days for qPCR to measure the efficiency of the vector DNA transfer into cells as presented in titration experiments of Chapter 3. After 3 days of puromycin selection, the isolated bulk puroR cells were washed to remove IL-3. The cells were re-suspended in 1 μg/ml puromycin culture medium and seeded in 24-well plates at 1:3 serial dilutions down to 1:9 with a starting cells number of 1 to 1.4 x 10^5. The cells were kept cultured without IL-3 supernatant (IL-3 starvation) for two weeks. In order to encourage the expansion of potential IL-3 independent mutants, IL-3 (10 % WEHI3B supernatant) was then re-added in the culture. Once cells were expanded, they were then cultured under IL-3 starvation again at least for two weeks.

Before starting this IM assay, the number of possible IL-3 independent clones were unknown, therefore the first trial was performed at a higher MOI (15.5) in which we aimed to obtain 10^4 puroR clones, with a high number of integrants in each cell. The first IM assay generated one IL-3 independent mutant (the IL-3I clone). To obtain more numbers of IL-3 independent mutants and to compare the mechanisms of potential IL-3 independence, another round of IM assay was performed. In the second IM assay, we used a lower MOI (4.4) and a larger number of cells to give roughly the same number of total integrants but reduce the number of integrants per cell. This MOI was chosen so that it would be less complex to identify integrants and fusion transcripts leading to IL-3 independence in any mutants. Although the total number of puroR cells was

greater than the first IM assay, this round did not generate any IL-3 independent mutants.

4.3.2 The *Ghr* exon 2 SA was not mainly used in mRNA expressed in the transduced cells by pGhr IRES-Puro

In Chapter 3, marker gene expressing cells by transduction of pGhr IRES-Em or Puro were not examined to see if splice-in using the *Ghr* exon 2 SA led to marker gene expression. To test if splice-in via the *Ghr* exon 2 SA is actually occurring to express puromycin mRNA in transduced cells, RT-PCR was performed (Fig.4-2A and B) (M.M 2.5.2). Two different forward primers were designed to distinguish splice-in using or not using the *Ghr* exon 2 SA. This was achieved by one forward primer annealing upstream from the *Ghr* exon 2 SA and the other primer annealing the downstream. We used total RNA extracted from the puroR clones in the puromycin titration at the highest dilution, some of the step 4 survivors randomly picked and the IL-3I clone from the first IM assay. A previously obtained puroR clone from pUBIQ-Puro transduction and parental Bcl15 as a negative control were run in the same reaction in parallel. Both pairs of primers showed amplification of DNA bands at the predicted size (the upstream of the *Ghr* exon 2 SA: 1k bp, the downstream of the *Ghr* exon 2 SA: 800 bp). While tested, transduced cells including puroR cells, survivors and IL-3I and Ubiq showed similar puroR amplification by primers annealing to the puroR gene (Fig.4-2C, 629 bp). Ideally a housekeeping gene such as ß-actin should be used as a standard, but at least this puroR amplification could be used as a control. All tested puroR cells were selected by 1 µg/ml puromycin and this threshold regulates the number of puroR cells therefore its puromycin expression would be estimated at the similar level as demonstrated in Fig.4-2C.

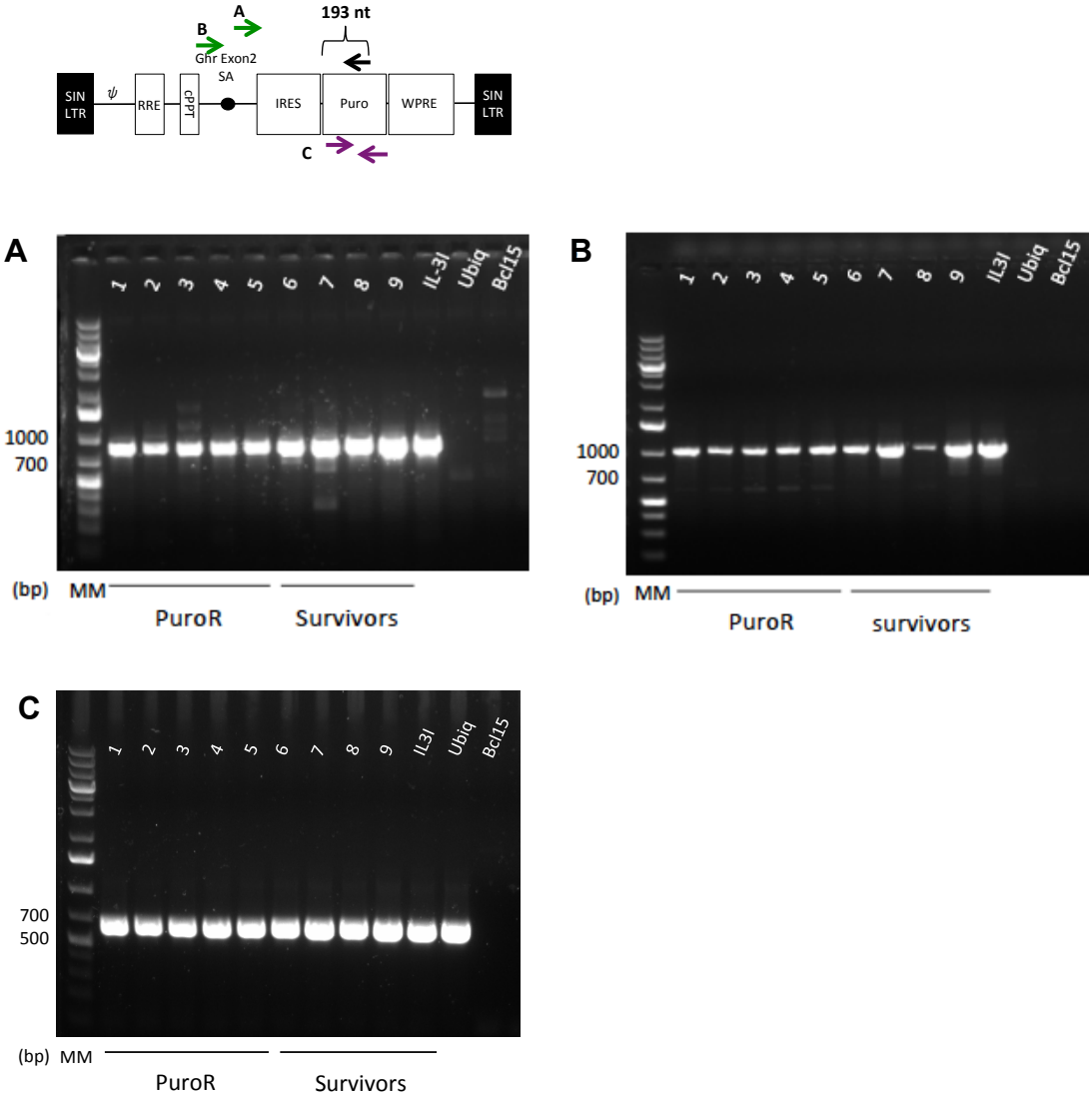


Fig.4-2 RT-PCR on the isolated puro-resistant (puroR) clones and the IL-3I clone from the first round IM assay.

The extracted total RNA was reverse transcribed using optimised blend of oligodT and random primers provided by QuantiTect Reverse Transcription Kit (Qiagen) (M.M 2.5.2). PCR was then performed on this cDNA template using two different forward primers as indicated in the provirus picture on the top. (A) The primer pair to test the presence of puromycin mRNA. A forward primer that anneals to the downstream of the *Ghr* exon 2 SA was used with a reverse primer that anneals to the puromycin sequence. (B) In order to test if the introduced *Ghr* exon 2 SA is exclusively used for splicing in the potential fusion puro mRNAs, a forward primer that anneals to upstream of the *Ghr* exon 2 SA was used with the same puro reverse primer used in (A). (C) Puromycin amplification

(629 bp) was tested as a control for puroR gene-coding vector (pGhr IRES-Puro and pUBIQ-Puro). PuroR: puromycin resistant clones; Survivors: step 4 survivors from the first IM assay); Ubiq: cells transduced by pUBIQ-Puro; Bcl15: untransduced cells; MM: molecular marker.

4.3.3 Phenotypic characterisation of the IL-3I clone

In order to investigate the mechanism of IL-3 independence in the IL-3I clone, the supernatant in the culture of the IL-3I clone was tested on parental Bcl15 cells for autocrine factor secretion (M.M 2.1.2). As a control the WEHI3B supernatant that was harvested in the same condition with the IL-3I clone supernatant was tested at different concentrations. Cell proliferation was measured by trypan blue (Gibco) (Fig.4-3A) or a colorimetric MTT assay (Cell Counting Kit-8, Dojindo) (Fig.4-3B). Both proliferation assays showed that the Bcl15 cells cultured in the 100 % IL-3I mutant supernatant showed cell growth similar to those cultured in 0.5 to 1.0 % WEHI3B a concentration that is less than 1:10 of routine use. Accordingly, the cell growth rate of the 20 % IL-3I clone supernatant was not as robust as the 100 %. This demonstrates that IL-3I clone do secrete a growth factor, for example IL-3 at a lower level, or a less potent factor.

From this data it was possible that the IL-3I clone might have evolved in a culture that contained a small sub-population of cells secreting a high level of growth factor, which then supported the remaining cells. To test this hypothesis the growth rate of the IL-3I clone at a range of initial cell numbers was measured. The lowest cell number (5.6×10^4 per well) is the condition at which parental Bcl15 cells expand efficiently when 10 % WEHI3B supernatant is supplied in the culture. Fig.4-3C shows that IL-3I clone could grow at each cell density, albeit with a lag at the lowest cell density.

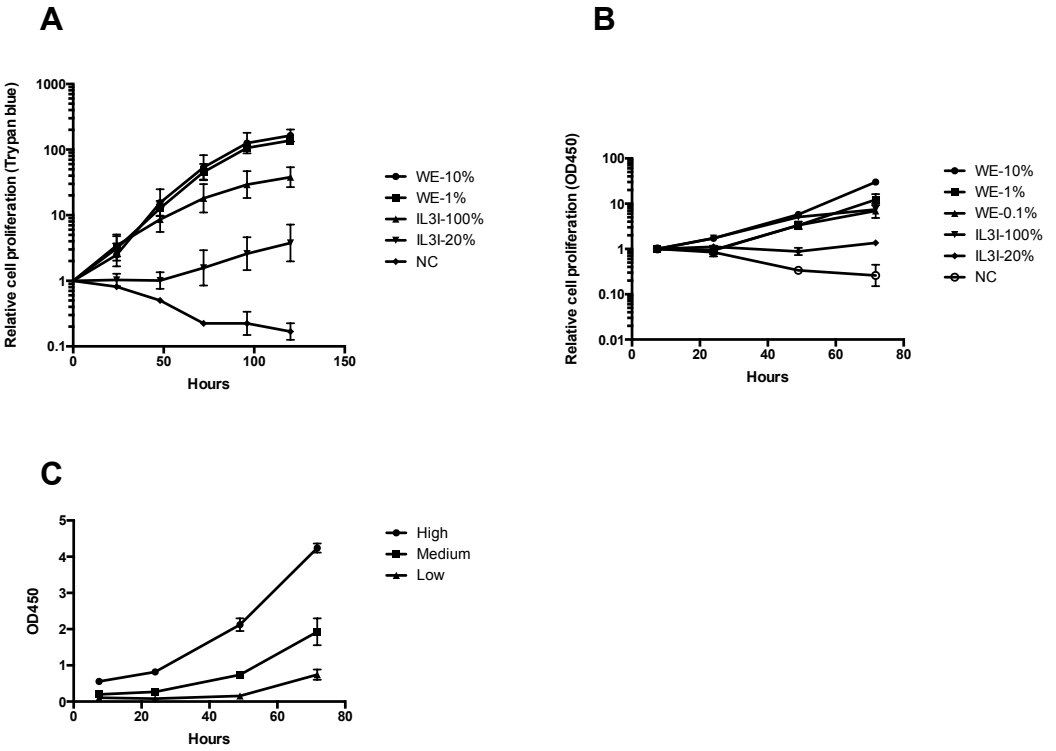


Fig.4-3 Autocrine factors secreted by the IL-3I clone (from the first round of IM assay) supported the cell growth of parental Bcl15 cells.

(A) Parental Bcl15 cells were cultured in the supernatant of the IL-3I clone (20 and 100 %) in comparison to positive control WEHI3B cell supernatant (0.01, 0.05, 0.1 (the concentration used only in B), 0.5, 1 and 10 %) for 6 days (M.M 2.1.2). The cell number was counted every 24 hours by trypan blue (Gibco). WEHI3B cells and the IL-3I clone supernatants were harvested in the same condition (post 24-hour harvest after seeding cell density at 5 x 10^5 cells/ml). (B) The same cell viability assay was repeated using CTK-8 (Dojindo). (C) The cell growth rate of the IL-3I clone at different cell density was measured by CTK-8 (Dojindo). Three initial cell numbers were used; 5.6 x 10^4 (low), 1.7 x 10^5 (medium), 5 x 10^5 (high). Cell proliferation was measured every 24 hours up to three days.

Those experiments were performed in duplicates. The error bar in each graph indicates the SEM between two experiments.

The previous work of Bokhoven et al (Bokhoven et al., 2009) showed that up-regulation of the IL-3 expression was a common mechanism of factor independence in “background” cell mutants that had become IL-3 independent in the absence of vector integration. Therefore, we initially tested for the presence of IL-3 mRNA by RT-PCR (Fig.4-4A) (M.M 2.5.3). WEHI3B cells that secrete IL-3 were used as a positive control to compare the strength of IL-3 expression. IL-3 mRNA amplification by RT-PCR was only detectable in the WEHI3B cells, but not in the IL-3I clone. To confirm if this undetectable IL-3 mRNA expression is due to the absence of the expression in the IL-3I clone or not, RT-qPCR was performed (Fig.4-4B) (M.M 2.5.3). RT-qPCR showed that a slight up-regulation of IL-3 mRNA in the IL-3I clone but was not as robust as the WEHI3B cells. IL-3 mRNA amplification was not detectable in other tested samples as in Fig.4-4A such as two step 4 survivors from the first IM assay and untransduced Bcl15 cells as expected.

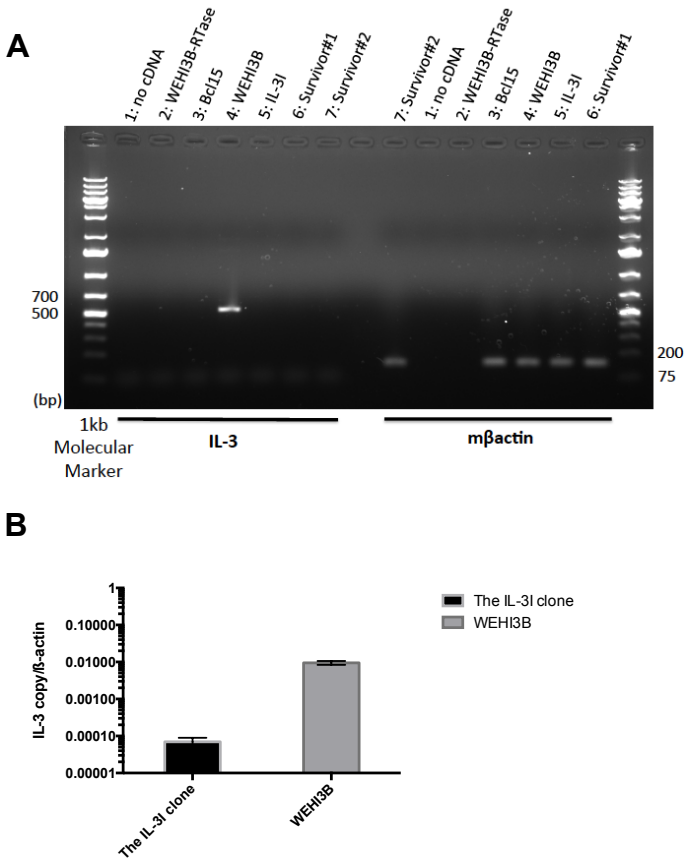


Fig.4-4 IL-3 mRNA detection

(A) Total RNA was reverse transcribed using an optimised blend of oligodT and random primers by QuantiTect Reverse Transcription Kit (Qiagen) and IL-3 mRNA expression was tested. The expected band size of IL-3 and mouse beta-actin control is 510 bp and 140 bp, respectively. Total RNA from WEHI3B cell and untransduced Bcl15 cell were used as a positive control and a negative control, respectively. Survivior 1 and 2 on the gel stands for step 4 survivors. The tested samples were loaded on a 1 % gel and electrophoresed. (B) The same sample series in (A) was tested for mRNA copy number of IL-3 and mouse beta-actin by SYBR green qPCR. Actin copy number was used to normalise the IL-3 copy number. The other samples tested in (A) were run in the same qPCR, however had no amplification of IL-3 mRNA, therefore they did not appear in the graph. Experiments were performed in triplicates. The error bar indicates the SEM between two experiments.

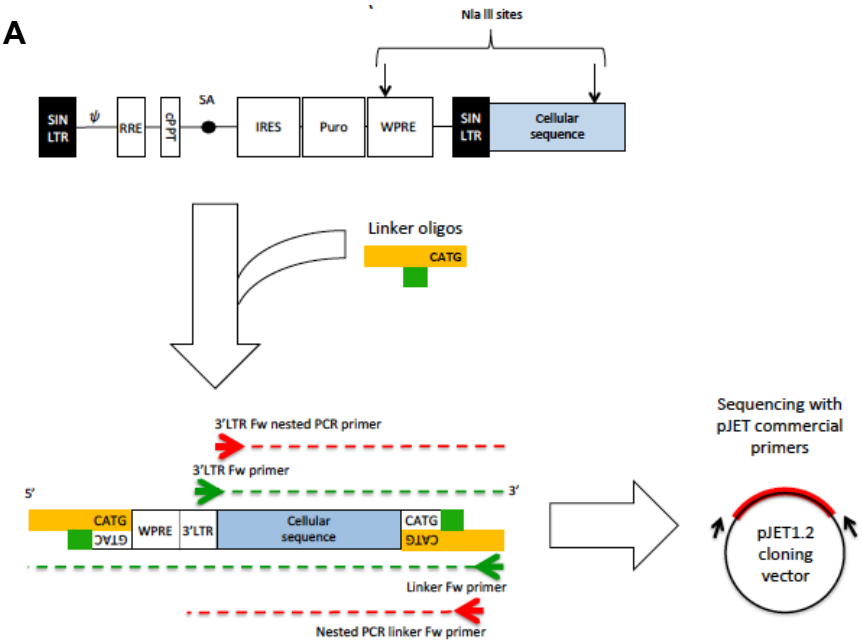
4.3.4 Integration site identification in the IL-3I clone by LM-PCR

LV integration sites in the IL-3I clones that link to the IL-3 independence were then interrogated. If integration sites near oncogenes as listed in RTCGD or cell survival-related genes could be identified, they can be the target for fusion mRNA expression by the integrated vector and the cause of IL-3 independence.

Integration sites were identified by ligation-mediated PCR (LM-PCR) (Fig.4-5A) (M.M 2.5.4) (Wu et al., 2003). The digested DNA fragments by a 4-cutter enzyme, NlaIII were ligated with synthetic oligonucleotide sequences (linkers) thereby flanked at both DNA ends. Using a vector primer that anneals in the U5 of the 3' LTR and a linker primer vector-host sequences were amplified. Amplicon DNAs were then cloned into a pJET1.2 cloning vector for sequencing and analysed by UCSC BLAT and Ensembl to find hits of host cellular gene. The two rounds of LM-PCR are summarised in Fig.4-5B. Four genes (highlighted in red) were found having vector integrations in the same orientation to the transcription direction of the host cellular genes.

At this stage I performed a preliminary analysis to see if any of the integration sites had been identified in the RTCGD (see Chapter 5, Introduction). Of the thirteen integration sites identified by LM-PCR I found that three of them appeared more than once in the RTCGD: *Sorcs2*, *Mgmt* and *Angpt1*. In general a single hit in this database may easily represent a passenger integrant in a multiply infected tumour, but sites appearing more than once have the possibility of significance. None of these three genes is classified as "common integration sites" (CIS), the gene loci with a proven functional significance in tumour formation.

A



B

Insertion site (Chr:integration site)	Genes with insertions or the nearest gene	RTCGD hits	Ref
1:93781343	*Atg4b* (intron 8-9)	0	(Marino et al., 2003)
1:9688240	*Mybl1* (intron 4-5)	1	(Mettus et al., 1994)
2:98667291	*Gm10800* (exon 1)	0	(Church et al., 2009)
3:145597269	*Znhit6* (intron 8-9)	0	(Francis et al., 2009)
4:7689360	*Car8* (downstream)	1	(Kato, 1990)
5:36325218	*Sorcs2* (Intron 1-2)	3	(Rezgaoui et al., 2001)
7:54159361	-	-	-
7:136570830	*Mgmt* (upstream)	2	(Shiota et al., 1992)
10:25508756	*Epb41l2* (intron 18-19)	0	(Takeuchi et al., 1994)
11:102559236	*Gpatch8* (upstream)	0	(Kai et al., 1997)
14:14781133	*Slc4a7* (intron 20-21)	1	(Okazaki et al., 2002)
15:42434109	*Angpt1* (intron 8-9)	2	(Davis et al., 1996)
Y:1158993	*Uty* (intron 14-15)	0	(Mazeyrat et al., 1998)

Fig.4-5 Vector integration site identification in the IL-3I clone by ligation-mediated PCR (LM-PCR).

(A) The scheme of LM-PCR (Wu et al., 2003) is explained. Synthesised linker oligo nucleotides were then ligated to the digested gDNA. In order to amplify target DNA fragments that have host cellular sequence at the 3' side of vector sequence, linker oligonucleotides were ligated. Primers binding to the 3' U5 and to the linker sequence amplify the vector-host genome sequences. The amplified products were electrophoresed on a 2 % agarose gel and the DNA bands were excised and column-purified. The DNA amplicons were cloned into a pJET cloning vector and sequenced. (B) The list of vector integration sites and genes within or near the integration sites from two rounds of LM-PCR are summarised. The vector integration site is described with the chromosomal number and location (the left and right side of the colon). When the integration site is a gene, the location is described either in an exon or an intron (between exon numbers). The gene name highlighted in red stands that both vector and the gene direction are the same. The number of RTCGD hits was extracted from http://variation.osu.edu/rtcgd/.

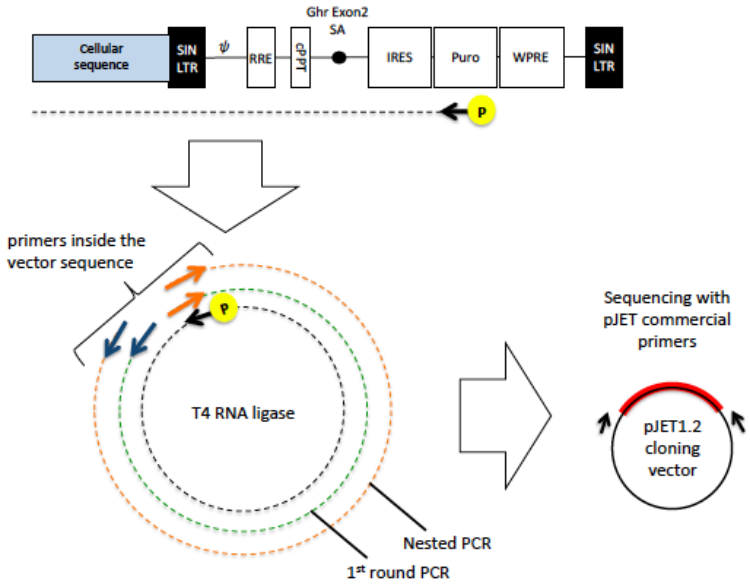
4.3.5 Identification of host cellular sequences in host-vector chimeric transcripts by 5' RACE

LM-PCR identified 13 integration sites and targeted genes respectively, however, we need to confirm if they are present on chimeric transcripts expressed in the IL-3I clone. We then employed 5' RACE to identify host cellular genes adjacent to 5' side of vector sequence (Fig.4-6) (M.M 2.5.5). In the 5' RACE, a reverse transcription (RT) primer annealing at the puromycin sequence (193 nt downstream of the start site of the puroR gene) was designed to investigate host cellular sequences in the chimeric puromycin mRNAs. Theoretically, circularised reverse transcribed single-stranded cDNA by T4 RNA ligase can contain a host cellular sequence between vector sequences. The DNA amplicons were cloned into a pJET1.2 cloning vector and sequenced. I also tested RNA from some of the step 4 survivors in this assay, although these were not truly IL-3 independent so any detected transcripts would be of unknown significance.

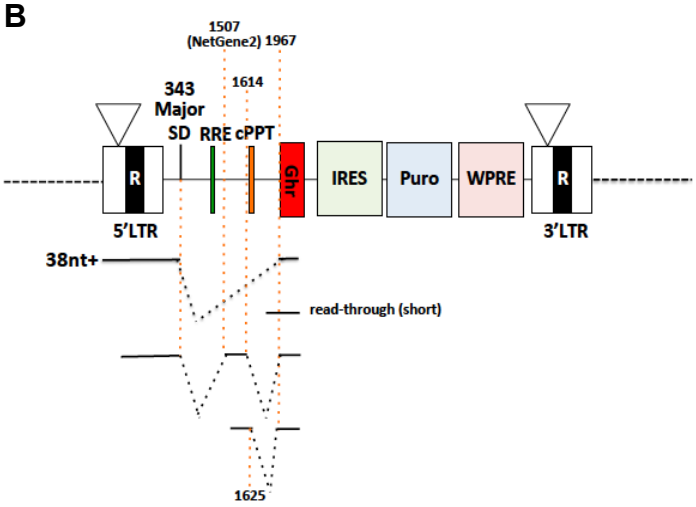
As to the IL-3I clone, one identified transcript presented 38 nt of potential host cellular sequence, however this did not match any sites on the GRCm38.p4 mouse reference sequence. In the identified transcripts, various splicing patterns were observed using the known SD (343: the HIV MSD), the known SAs (1507, 1967: the *Ghr* exon 2 SA)) and the cryptic SDs (1614, 1625). The known SD 343 and the SA 1507 were also detected in the *in silico* assay but not the cryptic SDs (Fig.3-3B).

One of the step 4 survivors showed splice-in fusion transcripts with myosin heavy polypeptide 9, non-muscle (*Myh9*) on chromosome 15 (Fig.4-6C, (right)). *Myh9* is also conserved in human and plays an important role in cytokinesis, cell motility and maintenance of cell shape (Lalwani et al., 2000). The isolated splice-in chimeric transcripts presented two different forms: splice-in from the exon 2 SD of *Myh9* to the *Ghr* exon 2 SA or to the HIV-1 major SD (MSD), followed by splicing to the *Ghr* exon 2 SA. Intriguingly, the SA site within 5' LTR (235) is cryptic since it was not detected by NetGene2 *in silico* analysis (Fig.3-3B, Chapter 3), but this SD was also observed in another study (Cesana et al., 2012). In both the IL-3I clone and a step 4 survivor, short sequence reads were also detected not including host cellular sequences but with different splicing site use within the vector (Fig.4-6C,D (left)). Some of the other transcripts detected from 5' LTR had various splicing patterns as observed in the IL-3I clones such as spliced-in between the HIV MSD with the *Ghr* exon 2 SA, or between the HIV MSD and the known SA 1507. The likelihood of 1507 being used is weaker (confidence: 0.26) than the *Ghr* exon 2 SA (confidence: 0.69) according to NetGene2. Other cryptic SD sites at position 1614 and 1625 were not predicted by NetGene2 (Fig.3-3B, Chapter 3).

A


Cellular sequence
SIN LTR
ψ
RRE
cPPT
Ghr Exon2 SA
IRES
Puro
WPRE
SIN LTR
P
primers inside the vector sequence
T4 RNA ligase
Nested PCR
1st round PCR
Sequencing with pJET commercial primers
pJET1.2 cloning vector

B


1507 (NetGene2) 1967
1614
343 Major SD
RRE
cPPT
R
Ghr
IRES
Puro
WPRE
R
5'LTR
3'LTR
38nt+
read-through (short)
1625

C

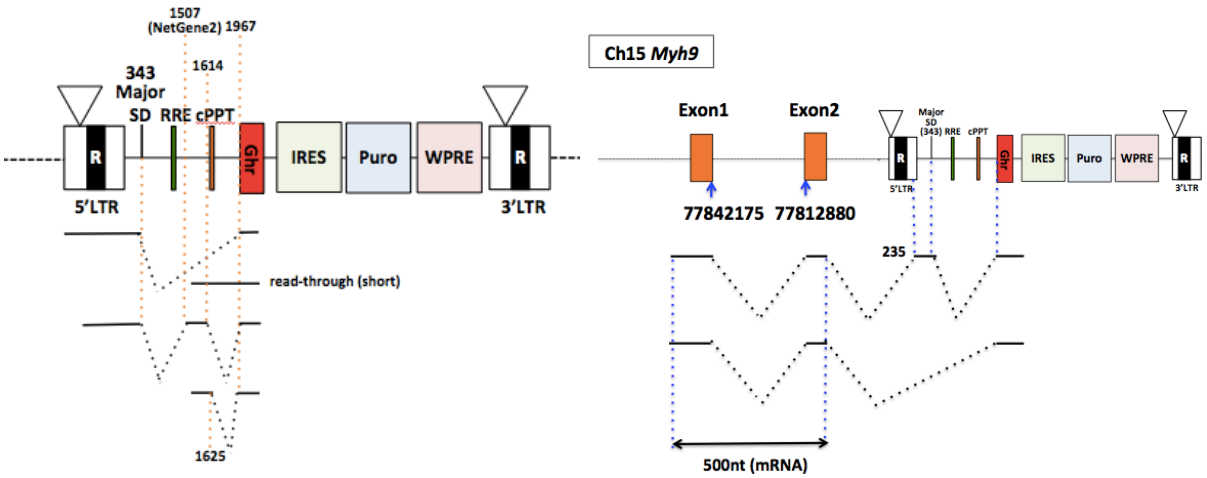

1507 (NetGene2) 1967
1614
343 Major SD
RRE
cPPT
R
Ghr
IRES
Puro
WPRE
R
5'LTR
3'LTR
read-through (short)
1625
Ch15 Myh9
Exon1
Exon2
77842175
77812880
Major SD (343)
RRE
cPPT
235
500nt (mRNA)

Fig.4-6 Detection of host-virus fusion mRNA of the IL-3I clone and a step 4 survivor by 5' RACE.

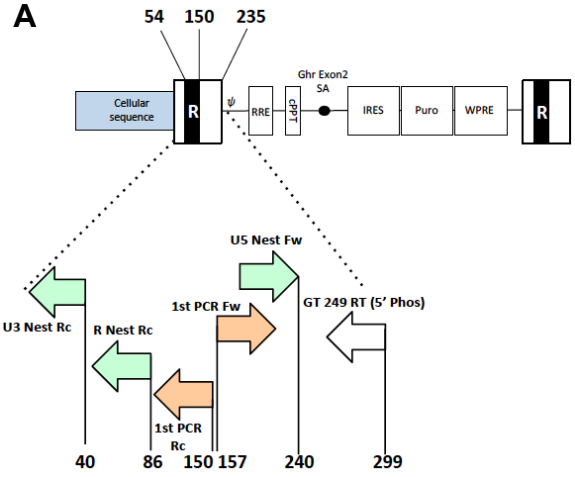
(A) The scheme of 5' RACE is visualised (M.M 2.5.5). The extracted total RNA was reverse-transcribed using a puroR gene-specific RT primer by QuantiTect Reverse Transcription Kit (Qiagen), followed by ligation of the single-stranded cDNA by T4 RNA ligase. The first round and nested PCR on the ligated products were performed to enrich target sequences. The final products were column-purified and cloned into a pJET cloning vector. The cloned fragments were sequenced and analysed. (B) An additional 38 nt at the 5' end of the vector was detected in the IL-3I clone, which is assumed to be host a cellular sequence. However this could not be mapped to the GRCm38.p4 mouse reference genome sequence. (C) The step 4 survivor showed various splicing patterns (left) and a fusion mRNA with *Myh9* (right). Splice-in from the *Myh9* exon 2 to the HIV major SD (MSD) or directly to the *Ghr* exon 2 SA were detected.

4.3.6 Further investigation of host cellular sequences in host-vector chimeric transcripts by designing the RT primer at further 5' side of the vector sequence

Because the first 5' RACE failed to obtain host cellular sequence in splice-in fusion form (Fig.4-6B), as a second attempt the RT primer was newly designed to prime upstream of HIV MSD to detect host cellular sequences in fusion transcripts expressed in the IL-3I clone more easily (Fig.4-7A) (M.M 2.5.5). Using the same protocol as in Fig.4-6 and the cloned final DNA bands were sequenced. As a result three host-cellular genomic sequences were identified on chromosomes 4, 7 and 17 (Fig.4-7B). The same integration sites on the chromosome 4 and 7 were detected by LM-PCR (Fig.4-5). To verify if they are genuine integration sites, PCR on gDNA of the IL-3I clone was carried out. PCR primers to amplify the identified genomic regions were designed approximately 200 nt of downstream and upstream of each integration site (Fig.4-7D, a table on the left). Firstly the primer function was tested if they can amplify the target genomic sequence (Fig.4-7C (green arrows) and D). Except for the PCR water control, all samples presented PCR amplification with the expected DNA length.

Next, using the genomic primers (Fig.4-7C (green arrows)) paired with vector primers annealing to 5' and 3' side in a vector (Fig.4-7C (orange arrows)) those three candidate sites were confirmed by PCR on gDNA. Only the integration site on chromosome 4 showed amplification, which is in good agreement with the identified vector integration site by LM-PCR (Fig.4-7E). The other two candidate sites on chromosome 7 or 17 showed non-specific amplification although the integration site was identified on chromosome 7 by LM-PCR (Fig.4-5).

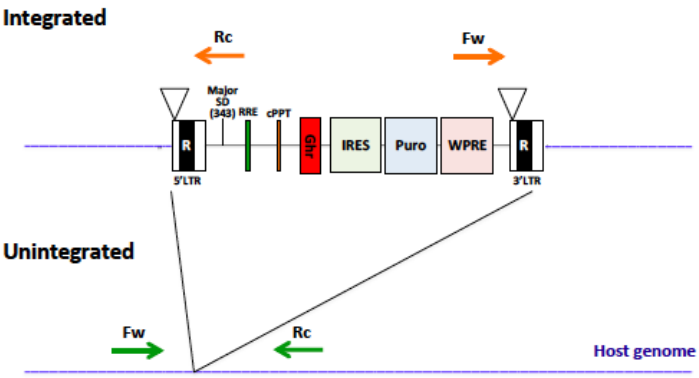
A



B

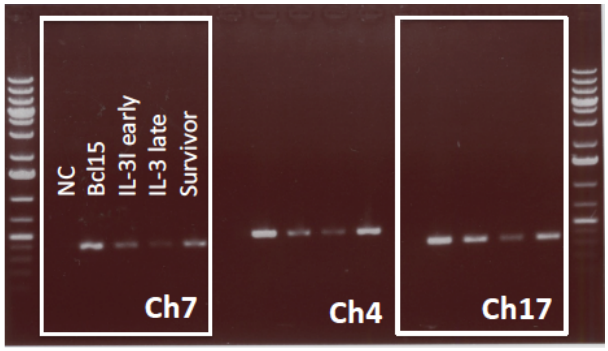
Chromosome number	Matched region	Sequence length (nt)
4	7689181-7689364	184
7	54159344-54159365	22
17	94384228-94384253	26

C



D

Chromosome number	Sequence length(nt)
4	479
7	387
17	335



E

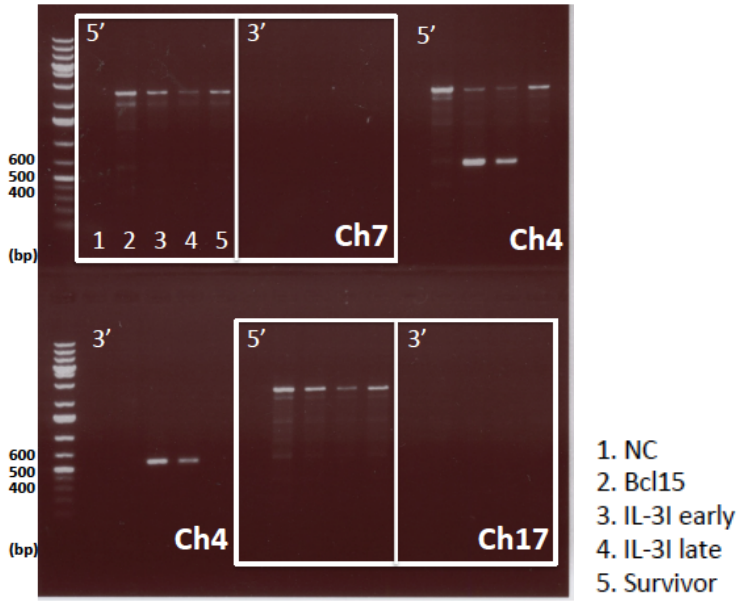


Fig.4-7 Another round of 5' RACE identified the same integration site on chromosome 4 that was found in LM-PCR

(A) The primer binding sites in the next 5' RACE (M.M 2.5.5) is indicated by arrows; RT primer (a white arrow), first PCR (an orange arrow) and nested PCR (a green arrow). Two Nested PCR Rc primers were tested on the same amplified sample by the first PCR. (B) Three candidate host cellular genomic sequences (on Ch4, 7 and 17) were identified and summarised in a table. (C) The primers annealing to the 5' or 3' side of vector integration site (green arrows) were designed on host genome to confirm the identified integration site. When vector integration site was confirmed, the boundary between host and vector sequence was amplified by primer pairs (green Fw and orange Rc, and green Rc and orange Fw). PCR was performed on gDNA extracted from the IL-3I clone, a step 4 survivor and parental Bcl15 cells. (D) To confirm the primer function, the paired chromosomal primers were tested on gDNA. The PCR-amplified products were run on a 1 % agarose gel. The same sample order shown in Ch7 is also applied to Ch4 and 17. The expected DNA length amplified by each pair of chromosomal primers is summarised in the table on the right side. (E) Host genome on each 5' or 3' side of the integrated vector sequence (indicated 5' or 3' on the gel picture) was amplified to confirm vector integration sites (M.M 2.5.5). The amplified PCR products were run on a 1 % agarose gel. Five samples were tested in each 5' or 3' reaction for three sites and the sample order is shown on the right side of the gel picture.

4.4 Discussion

This project was designed for the identification of host genes that could be regulated by “splicing in” to a SA present in a LV. I hypothesised that this assay would detect events such as that present in the ß-thalassaemia clinical trial where splicing into a LV removed the 3’ end of the *HMGA2* transcript (Cavazzana-Calvo et al., 2010). The use of the Bcl15 cell line presupposed that there would be mRNAs whose expression could be upregulated in this manner which then made the cells IL-3 independent. It would also be possible that the LV insertion could disrupt expression of one allele of a gene, for example that inhibits cell death, and that this could render cells IL-3 independent, however I considered this option less likely as such mutants have not been identified in our previous work.

By analogy with the previous work in my lab I reasoned that even a single host gene that responded in such a manner would be of great value. The study of Bokhoven et al showed LV integration into *Ghr* could upregulate this gene by splicing out from the vector (Bokhoven et al., 2009). During the work the assay was adapted and an extra growth hormone was added to the assay so that such mutants grew more readily. This single locus could then be used to screen and eliminate SD sites in clinical candidate vectors (Knight et al., 2012) (Cesana et al., 2012). So if I could find a “splicing in” locus and determine its mechanism of making Bcl15 cells IL-3 independent, I could potentially adapt the assay to this mechanism.

4.4.1 pGhr IRES-Puro transduction in IM assay generated one IL-3 independent clone

Using the functionally characterised pGhr IRES-Puro, I performed two mutagenesis assays and isolated one mutant, the IL-3I clone. This occurred at a frequency of 1 mutant per 1.5 x 10^8 integrants. It is not too surprising that the frequency of such mutants was so low. The Ghr mutants in the “splice out” assay

were more frequent (1 in 4.1 x 10^7 integrants) but in this case the LV could integrate into the 48.3 kb upstream of the second intron of the *Ghr* gene. If the "splice in" targets required integration into a more limited region or regions, far fewer mutants would be detected. It remains unclear whether the higher MOI used in the first assay was important for isolating the single mutant or whether this occurrence in one assay but not the other is a stochastic difference.

Another possible reason contributing to the small mutant number might be a strong puromycin selection at the early stage of the IM assay protocol. The initial purpose of this selection was to isolate puromycin resistant (puroR) clones which have been potentially transformed to IL-3 independent by splice-in events. However, this early selection cuts off any other puroR clones with a lower level of puroR expression that could be IL-3 independent mutants. A future assay could be performed with a lower level of puromycin in the selection medium, or with no puromycin selection.

4.4.2 The IL-3 independent clone secreted autocrine factor to support cell growth

The use of the introduced *Ghr* exon 2 SA was interrogated by PCR (Fig.4-2). If the puromycin resistance of transduced cells were gained exclusively by splice-in using the SA of *Ghr* exon 2, the PCR amplification would be detectable using the downstream primer but not the upstream primer. The positive amplification by different forward primers can explain read-through transcripts without using the *Ghr* exon 2 SA also contribute to puromycin resistance but the frequency of this event is unknown from this experiment. This suggests that host-virus chimeric transcripts expressed in these tested samples could be generated using any SA sites upstream of the *Ghr* exon 2 SA or read-through from host cellular sequences not mediated by any splicing events.

The IL-3 independent clone showed secretion of a growth factor demonstrated by the expansion of parental Bcl15 cell in supernatant from the IL-3I mutant

(Fig.4-3A and B). Some level of IL-3 transcript was found in this clone (Fig.4-4B). However, the work of Bokhoven et al would have classified this level of IL-3 mRNA as “background” and did not find significant growth factor secretion by clones with this level of IL-3 mRNA (unpublished data). In the work of Bokhoven et al, very low levels of IL-3 mRNA (< $10^3/10^9$ 18S rRNA transcripts) were scored as negative. This would equate to approximately <10^3/actin transcript (Bas et al., 2004), so the clone IL-3I falls within this very low level (Fig. 4-3A and B). It is possible that the growth factor secreted by the IL-3I clone is a very low level of IL-3, however it also remains possible that another survival factor is secreted, which may even act in synergy with a low level of IL-3. In addition, the proliferation of the IL-3I clone is enhanced by higher cell density (Fig.4-3C). This suggests that the whole population is able to secrete a growth factor and proliferate in an autocrine manner. The lag at low density could be explained by the cell cycle inhibitory effect of Bcl2 when IL-3 is removed (Marvel et al., 1994).

4.4.3 A few integration sites identified had RTCGD hits

In order to identify the mechanism of IL-3 independence in the clone IL-3l I first cloned LV integration sites by LM-PCR (Fig.4-5B). As the cells were transduced at high MOI it was not surprising to find thirteen sites, and it is not clear whether this represents a full list. I performed a preliminary screen of these sites in the Retrovirus Tagged Cancer Gene Database (RTCGD) and the full relevance of the RTCGD to the Bcl15 assay is described in the introduction to Chapter 5. Of the 13 sites 3 appear more than once in the RTCGD; *Sorcs2*, *Mgmt* and *Angpt1*. *Sorcs2* is a receptor with roles in intracellular vesicle trafficking and neurotrophin binding in the nervous system (Lane et al., 2012), *Mgmt* is a DNA repair enzyme, often upregulated as a mechanism of cancer drug resistance (Erickson, 1991), and *Angpt1* is a growth factor involved in vascularisation, including in cancer (Huang et al., 2010). However, integration into these loci does not indicate that host gene expression is affected, so I then moved on to attempt to identify fusion transcripts in the IL-3I clone by 5’ RACE. Not every integrant will give rise to such a transcript that can perturb host gene expression, and those that do are

more likely to be associated with IL-3 independence.

4.4.4 5' RACE identified known and cryptic splice sites within a vector sequence possibly in fusion mRNA, but failed to identify the host sequence itself in the IL-3I clone

This method failed to identify any host gene fusion transcripts in clone IL-3I. Use of RNA-Seq for this and the implications of the results are discussed in Chapter 5. The 5' RACE did identify splicing patterns within the vector, both in a fusion transcript from one of the step 4 survivors with an integrant in *Myh9* and from an unknown integrant or integrants in the clone IL-3I. As the step 4 survivor is not truly IL-3 independent I do not believe that this *Myh9* integrant is necessarily of any functional significance in IL-3 independence. My *in silico* analysis showed a list of predicted splicing site candidates in the vector provirus sequence (Fig.3-3A and B, Chapter 3). I found that the HIV-1 major SD (MSD, 343) or the known HIV-1 SA downstream of RRE (1507) were predicted sites and also used in the detected transcripts. In addition, novel cryptic SD sites were also observed (1614 and 1625). These unpredicted SDs could be mutagenic depending on the vector integration site and host chromosomal environment and should perhaps be mutated. Short read-through sequence fragments were also detected not using the *Ghr* exon 2 SA. This is in good agreement with the presence of the read-through transcripts that resulted in band amplification with the upstream primer (Fig.4-2B).

The second 5' RACE (Fig.4-7), we identified two genomic sequences that match with vector integration site detected by LM-PCR (Fig.4-5). One site on chromosome 4 was verified as genuine integration site by gDNA PCR amplifying host-vector sequence boundary at 5' and 3' side of the vector. In LM-PCR analysis, a neighbouring gene *car8* was found 550 kb downstream of the vector integration site on the chromosome 4. Vector integration is in the anti-sense relative to the *car8* transcription direction and the vector has self-inactivating LTR, therefore the possible *car8* fusion transcript does not contribute to puroR

gene expression in the IL-3I clones. *Car8* and the relation to tumourigenesis are not known.

In addition, the second 5' RACE aimed to read a further 5' host cellular sequence adjacent to vector sequence. The RT primer at the 5' side of the vector was designed based on identified spliced transcripts within vector sequence and the use of HIV MSD (Fig.4-6B). This design of the RT primer could miss fusion transcripts having the direct splice-in between and a host SD and the *Ghr* exon 2 SA or any cryptic SAs between the MSD. This may not lead to the recovery of identified vector integration sites by LM-PCR (Fig.4-5) and some possible host sequences in fusion transcripts.

Overall, the 5' RACE results gave some information on splice sites possibly used in fusion transcripts that could give rise to puroR expression. One integration site on chromosome 4 was confirmed using different techniques (Fig.4-5 and 4-6). Because 5' RACE could not identify and complete a host cellular sequence in splice-in fusion form, further mRNA analysis by RNA-Seq was required to the identify fusion transcripts in clone IL-3I. We expected that such host sequence might be related to IL-3 independence and the results of this analysis are presented in Chapter 5. We also considered that step 4 survivors could give us a hint about which host cellular sequences could be affected by vector integration and possibly lead mutant generation. At step 4 of mutagenesis assay, cultured transduced cells are starved from growth factor and certain selection of gene expression related to cell survival is expected. Pooled step 4 survivors were also analysed by RNA-Seq in the next chapter.

Chapter 5

Identification of host-vector fusion mRNAs and assessing their relevance for cell survival

5.1 Introduction

The purpose of the experiments in this Chapter was firstly to the identify host-vector fusion transcripts present in the clone IL-3I. Due to the fact 5' RACE had failed to identify such transcripts, I decided to employ RNA-Seq. In this method cDNA is synthesized from mRNA, and because I was looking for host-vector fusion transcripts I optimised a primer within the LV sequence for this first strand cDNA synthesis. After successful identification of two fusion transcripts in the clone IL-3I, I reasoned that I could apply this method to examine fusion transcripts in bulk populations of LV transduced Bcl15 cells (Fig.5-1). I therefore also analysed host-vector fusion transcripts in an IL-3 starved population and unstarved control cells.

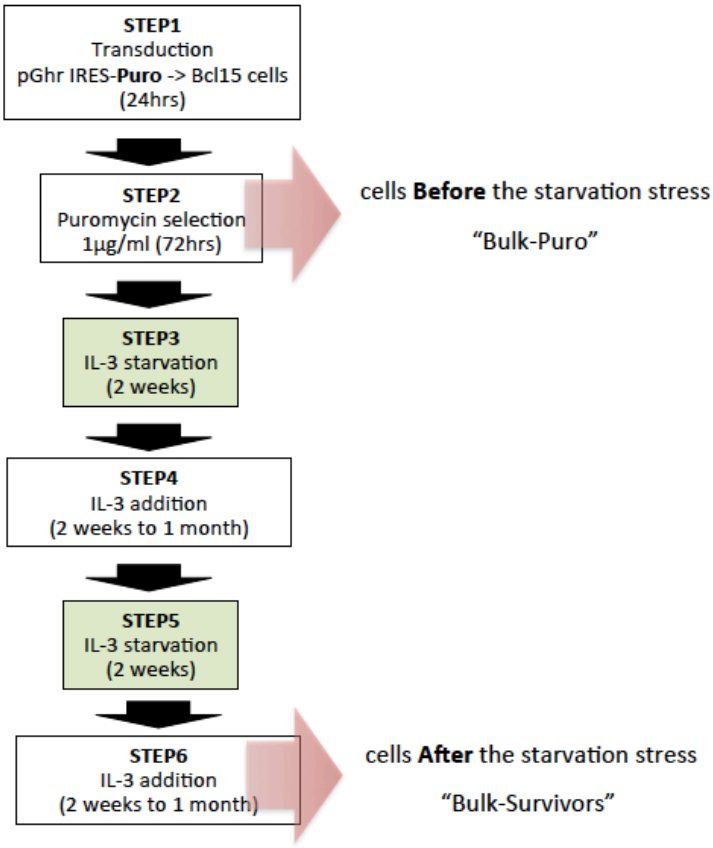


Fig.5-1 The harvest of transduced bulk populations

Two bulk populations were harvested at the different steps in the second IM assay for Illumina RNA sequencing to compare the expression profile of fusion mRNAs.

In order to estimate the possible relevance of the detected transcripts to cell survival or proliferation I planned to employ two methods. Firstly, BAF3 cells (the parent of Bcl15) have been extensively studied as a model for haematopoietic cell survival and growth. I therefore prepared a table of genetic modifications of BAF3 cells that have been reported to enhance survival or proliferation (Table 5-1). Since Collins et al demonstrated that BAF3 cells could be used to express heterologous receptors making them responsive to new ligands EGFR and IL2R (Collins et al., 1988) (Collins et al., 1990), a large number of groups have used these cells for receptor function analysis (Table 5-1, Group 1). Secondly, oncogenic proteins identified in human tumours have been shown to render BAF3 cells tumourigenic in mice (Table 5-1, Group 2). Thirdly, activated signaling molecules (sometimes forms found in human tumours) or inhibitors of signaling pathways have been expressed in the cells to examine which pathways lead to survival or proliferation (Table 5-1, Group 3). Finally, three retroviral or LV mutagenesis screens have been carried out in these cells to identify IL-3 independent mutants. The screen by Marvel and colleagues identified Bcl-X as a common retroviral insertion site in four cell clones that were resistant to multiple apoptotic stimuli (Thomas et al., 1998). The screen by Koh and colleagues, using transduction of a cDNA library in a retroviral vector, also identified Bcl-X, a number of kinases and two transcription factors (Koh et al., 2004). The screen of BAF3 and Bcl15 cells from our group by Bokhoven et al showed that many LV insertional mutants selected by IL-3 removal had activated expression of the growth hormone receptor (*ghr*) and those selected after gammaretroviral mutagenesis had insertion into IL-3 itself and a number of other genes (Bokhoven et al., 2009). My assays were performed in Bcl15 cells, already expressing a high level of Bcl2, so upregulation of Bcl2 or its homologue Bcl-X was unlikely to be detected. However, I decided that one useful analysis of any fusion transcripts that I detected would be to see if their cellular partner had previously been reported to have an effect on survival or proliferation of BAF3 cells in the literature summarised in Table 1.

Secondly, Bokhoven et al noted that almost all mutants isolated in their RV or LV screens, but not control clones, had at least one insertion into a site identified in

the Retrovirus Tagged Cancer Gene Database (RTCGD) (Akagi et al., 2004). This database provides a summary of all genes with insertions of retroviruses or transposons, where these agents have been used to induce tumours in a variety of animal models. I updated the analysis of genes identified in Bokhoven et al, then analysed the genes identified in the BAF3 cell screens by Marvel et al and Koh et al (Table 5-2) and found that 48 out of 54 independent clones showed insertion in (or expression of in case of Koh) genes featured in the RTCGD. Thus, genes that render BAF3 cells able to survive or proliferate in the absence of IL-3 largely overlap with genes that are activated by retroviral or transposon integration in animal tumour models. Therefore, I decided that I would also examine the cellular partners of fusion transcripts to see if they were represented in RTCGD. A combination of these two approaches should indicate any candidate fusion transcripts that might be expected to influence cell survival or proliferation.

Table 5-1 Identified genes related to cell proliferation and survival using BAF3 cells

Group 1

Gene	Gene name	Reference
EGFR	Epidermal growth factor receptor	(Collins et al., 1988)
IL2R	Interleukin 2 receptor	(Collins et al., 1990)
GMCSFR	Granulocyte macrophage colony-stimulating factor receptor	(Park et al., 1986)
IL5R	Interleukin 5 receptor	(Sakamaki et al., 1992)
EPOR	Erythropoietin receptor	(Frederiks et al., 1978)
PDGFR	Platelet-derived growth factor receptor	(Satoh et al., 1993)
GCSFR	Granulocyte colony-stimulating factor receptor	(Dong et al., 1993)
MPL/thrombopoietin	Myeloproliferative leukemia virus oncogene/thrombopoietin	(Drachman et al., 1995)
FGFR	Fibroblast growth factor receptor	(Wang et al., 1994)
PRLR	Prolactin receptor	(O'Neal and Yu-Lee, 1994)
CNTFR	Ciliary neurotrophic factor receptor	(Kruttgen et al., 1995)
IL11R	Interleukin-11 receptor	(Nandurkar et al., 1996)
c-Met	Hepatocyte growth factor receptor	(Schwall et al., 1996)
IL7R	Interleukin-7 receptor	(van der Plas et al., 1996)
IL12R	Interleukin-12 receptor	(Zou et al., 1997)
LEPR	Leptin receptor	(Ghilardi and Skoda, 1997)
PAR2	Proteinase-activated receptor 2	(Mirza et al., 1997)
Flt3	Fms-related tyrosine kinase 3	(Shibayama et al., 1998)
IL9R	Interleukin 9 receptor	(Louahed et al., 1999)
IL21R	Interleukin 21 receptor	(Parrish-Novak et al., 2000)
IL6R	Interleukin 6 receptor	(Jostock et al., 2001)
IL22R	Interleukin 22 receptor	(Rutherford et al., 2001)
c-Kit	Mast/stem cell growth factor receptor Kit	(Ueda et al., 2002)
IL27R	Interleukin 27 receptor	(Pradhan et al., 2007)

GHR	Growth hormone receptor	(Barclay et al., 2007)
TFRC	Transferrin receptor	(Shi et al., 1997)

Group 2

Gene	Gene name	Reference
BCL-2	Apoptosis regulator Bcl-2	(Collins et al., 1992)
Bcr/Abl	Breakpoint cluster region /ABL proto-oncogene 1, non-receptor tyrosine kinase	(Bazzoni et al., 1996)
Tel/PDGF	ETS variant 6/platelet-derived growth factor receptor beta	(Sjoblom et al., 1999)
ALK/ATIC	ALK tyrosine kinase receptor/5-Aminoimidazole-4-Carboxamide Ribonucleotide Formyltransferase/IMP Cyclohydrolase	(Ma et al., 2000)
Tel/Syk	ETS variant 6/Spleen tyrosine kinase	(Kuno et al., 2001)
E2a/Pbx1	Transcription factor E2-alpha/Pre-B-cell leukemia transcription factor 1	(Rutherford et al., 2001)
BCR/FGFR1	Breakpoint cluster region/Fibroblast growth factor receptor 1	(Demiroglu et al., 2001)
NPM/ALK	Nucleophosmin/Anaplastic lymphoma kinase	(Kasprzycka et al., 2006)

Group 3

Gene	Gene name	Reference
MKK1	MAP kinase kinase 1	(Perkins et al., 1996)
Akt	RAC-alpha serine/threonine-protein kinase	(Ahmed et al., 1997)

Ras	Related RAS Viral (R-Ras) Oncogene Homolog	(Suzuki et al., 1997)
JNK	c-Jun N-terminal kinase	(Smith et al., 1997)
mSos	Son of sevenless homolog 1	(Tago et al., 1998)
Btk	Tyrosine-protein kinase BTK	(Deng et al., 1998)
Rac	RAC-alpha serine/threonine-protein kinase	(Nishida et al., 1999)
STAT5	Signal transducer and activator of transcription 5A	(Zhang et al., 2000b)
SHP-1	Src homology phosphatase-1	(Paling and Welham, 2002)
Lyn	LYN Proto-Oncogene, Src Family Tyrosine Kinase	(Lannutti and Drachman, 2004)
Pim1	Pim-1 Proto-Oncogene, Serine/Threonine Kinase	(Kim et al., 2005)
ALK	Anaplastic Lymphoma Receptor Tyrosine Kinase	(Wan et al., 2006)
Jak2	Janus Kinase 2	(Walz et al., 2006)
Lck	LCK Proto-Oncogene, Src Family Tyrosine Kinase	(Shi et al., 2006)
Daxx	Death-Domain Associated Protein	(Muromoto et al., 2010)

Table 5-2 Genes related to oncogenesis derived from BAF3 IM assays

Referene	Surviving clones	Gene	Number of RTCGD hits
Bokhoven (Bokhoven et al., 2009)	C10	Dusp6	4
	C40	Osbpl3	4
	C57	Sema4b	5
	C96	Zfp36/Plekhg2	9
	G18	Il3	0
	10 LV clones	Ghr	6
	HV48	Senp1	2
	HV72	Primpol	0
Marvel (Marvel et al., 1994)	4 clones	Bcl2l1 (Bcl-X)	7
Koh (Koh et al., 2004)	5 clones	Fms	6
	4 clones	Fes	1
	2 clones	Tyk2	0
	2 clones	Araf	1
	2 clones	Map3k8	16
	2 clones	Map2k3	0
	1 clone	Hhex	41
	1 clone	Hlx	0
	1 clone	Bcl2l1	7

5.2 Aims

1. To identify host-vector fusion transcripts in the clone IL-3I by developing a vector-primed RNA-Seq protocol
2. To use this protocol to identify host-vector fusion transcripts in an IL-3-starved population of Bcl15 cells, compared to unstarved controls
3. To assess the possible relevance of such host-vector fusion transcripts in promoting cell survival with reference to the BAF3 cell literature and the RTCGD

5.3 Results

5.3.1 RNA deep sequencing to analyse the fusion mRNAs in the clone IL-3I

Theoretically, a host-vector fusion transcript can be generated when the model vector is integrated downstream of a cellular promoter in the same orientation as the host transcription, allowing the *Ghr* exon 2 SA to be used with the host gene SD (Fig.5-2A). To detect host sequences and splicing patterns in fusion transcripts, a specific reverse transcription (RT) primer, SA1, was designed to anneal 47 nt downstream from the beginning of the *Ghr* exon 2. This RT primer contains a region designed to anneal to the vector (red) and an overhang sequence (green) (Fig.5-2B). After single-stranded cDNA is generated, double-stranded DNA is synthesised by using a random octamer with a different overhang sequence (Fig.5-2C) (M.M 2.6.5).

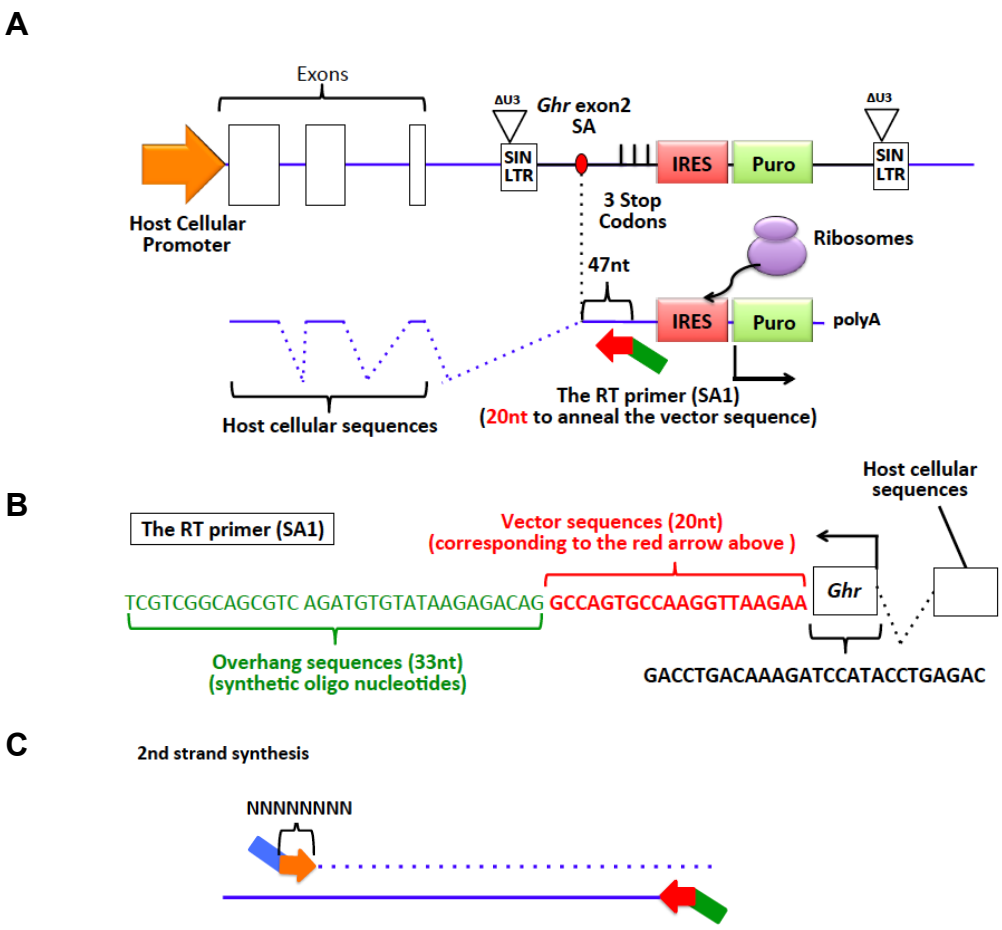


Fig.5-2 The scheme of cDNA preparation for MiSeq RNA sequencing

(A) Host-virus splice-in fusion mRNAs can be expressed by integration of our model vector in the same orientation with host transcriptional direction. In order to detect the splice-in chimeric transcripts, a reverse transcription (RT) primer was designed, SA1. A red arrow with a green tail in the RT primer indicates a complementary sequence to the vector-specific region and an overhang sequence (tag), respectively. (B) The RT primer sequence. (C) The second strand of the cDNA is synthesised using a random primer (octamer) with overhang sequences (tag). Using the primer pair that anneals to the tag sequences at both ends of transcripts, PCR can be performed to enrich target DNA fragments.

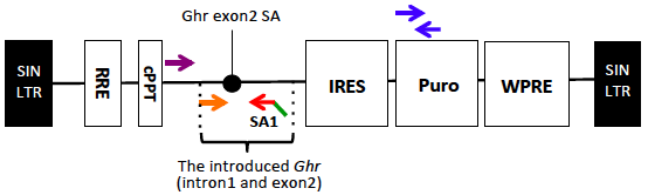
5.3.2 Testing the SA1 primer in DNA amplification and RT priming

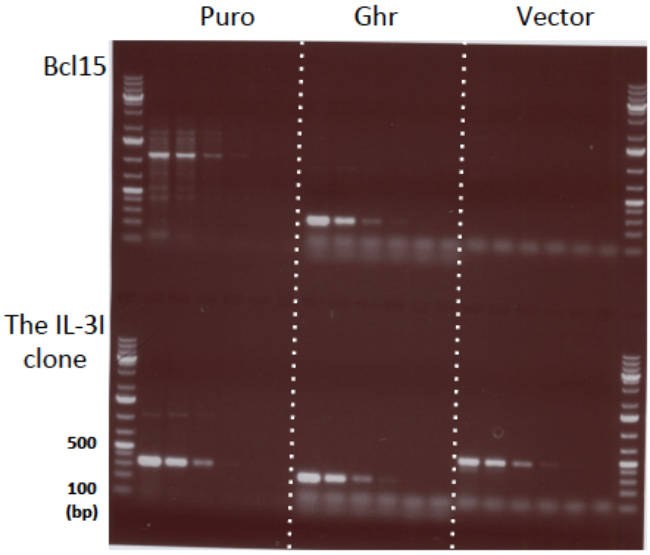
Initially, one RT primer candidate (SA1) was tested for the general amplification from gDNA extracted from untransduced Bcl15 cells and the IL-3I clone (M.M 2.6.4). DNA amplification was tested semi-quantitatively on 1:5 dilutions of gDNA from a starting amount of 100 ng. This RT primer anneals to *Ghr* exon 2 and a forward primer anneals to the introduced *Ghr* region. Therefore, the amplification is detected in both cell populations (Fig.5-3A, the middle part in the gel picture). When the SA1 primer was used with the forward primer annealing to the vector backbone (upstream of the introduced *Ghr*), only the IL-3I clone amplified (Fig.5-3, the right side of the gel picture). Puro amplification was tested as a positive control for the IL-3I clone; amplification was observed in the IL-3I clone, not in the Bcl15 as expected. Therefore, SA1 can prime DNA synthesis by annealing to the expected site on the vector sequence as shown by amplification products of the expected size.

Three different RT primers were then tested by SYBR green qPCR: SA1, SA2 and IRES, in order to choose an RT primer with optimal amplification (M.M 2.6.4). I decided to use a bulk transduced cell population at this point as a test of RT priming; this might be more stringent as many transcripts could be present at a lower level than those in the IL-3I clone. Therefore, mRNA from Bcl15 and the transduced Bcl15 cells with pGhr IRES-Puro cultured in puromycin (Bulk-Puro, see Fig.5-1) was reverse transcribed with one of three candidate RT primers and qPCR was performed with three primer pairs: Ghr, puromycin (Puro) and mouse beta-actin (Fig.5-3B). When the cycle number detected in parental Bcl15 cells would be regarded as negative, the lowest cycle number 32.44 with RT primed by SA1 and detected by Ghr primers could be a threshold in this reaction. With this regard, positive PCR amplification by the Ghr primer (the orange Fw primer and a green tag sequence of each RT primer) was only observed with SA1, but not with two other RT primers (the upper part of the table in Fig.5-3B). As controls, PCR amplification of puromycin and mouse beta-actin was also tested in the same qPCR reaction. Since only transduced cells selected by puromycin have puroR mRNA expression, Bulk-Puro mRNA had PCR amplification with puro primers but Bcl15 cells did not (the middle part of the table in Fig.5-3B).

While both Bcl15 cells and Bulk-Puro showed PCR amplification of mouse beta-actin. Across three PCR primers amplification was not observed in negative control (No RTase-oligodT). PCR amplification with puro primers is similar among three candidate RT primers, suggesting the presence of host-vector fusion mRNAs. However, the fact that PCR amplification with Ghr primers was only detectable with SA1 implies that SA2 and IRES RT primers probably did not efficiently prime reverse transcription. Therefore, SA1 was selected as an optimal RT primer for mRNA library preparation for RNA sequencing. Ideally this experiment should have been duplicated but it was not.

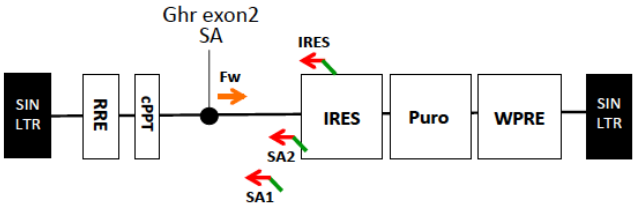
A





Primer	Band size (bp)
Puro	330
Ghr	199
Vector	437

B



qPCR primers	RT primers	Bcl15	Bulk-Puro
Ghr	oligodT	ND	ND
	SA1	32.44	21.34
	SA2	ND	ND
	IRES	37.14	ND
	No RTase-oligodT	ND	38.38
Puro	oligodT	ND	19.03
	SA1	ND	19.12
	SA2	ND	19.40
	IRES	ND	19.61
	No RTase-oligodT	ND	37.73

Mbeta-actin	oligodT	19.38	17.84
	SA1	19.89	18.86
	SA2	19.29	18.27
	IRES	19.85	19.09
	No RTase-oligodT	33.05	33.93

Fig.5-3 Design of an RT primer

(A) PCR was performed on gDNA extracted from parental Bcl15 cells and the IL-3I clone to test the specific amplification by three different pairs of primers including a RT primer (SA1); a pair of puromycin primers (puro, blue arrows), Ghr (orange arrow) and vector (purple arrow). The Ghr primer (orange) and the vector primer (purple) were paired with primer SA1. gDNA samples were serially diluted (1:5, 5 points) from the starting amount of 100 ng. The furthest left lane of three reactions (divided by dashed line) is negative control (no gDNA). PCR was performed by GoTaq G2 DNA polymerase (Promega). (B) In order to choose an optimal RT primer, target amplification by two other RT primers SA2 and IRES were compared with by SA1. The annealing site of each RT primer is indicated on the vector provirus schematic by the red arrows with the green tails. cDNA was synthesised using mRNA extracted from Bcl15 cells and Bulk-Puro. As a control RT primer, oligo dT was tested alongside. qPCR was then performed by QuantiTect SYBR green (Qiagen) using three pairs of primers: Puro primers used in (Fig. 5-3A), the green tag sequence of each RT primer (indicated with green) with a forward primer of *Ghr* exon 2 (an orange arrow in the schematic) and mouse actin primers. Amplification by those primer pairs was summarised by a cycle number in the table. Ghr: primer pairs for growth hormone receptor; Puro: primer pairs for puromycin resistant gene; mActin: primer pairs for mouse beta actin; ND: non-detectable

5.3.3 mRNA prepared by direct lysis showed nonspecific amplification of DNA in the untransduced negative control

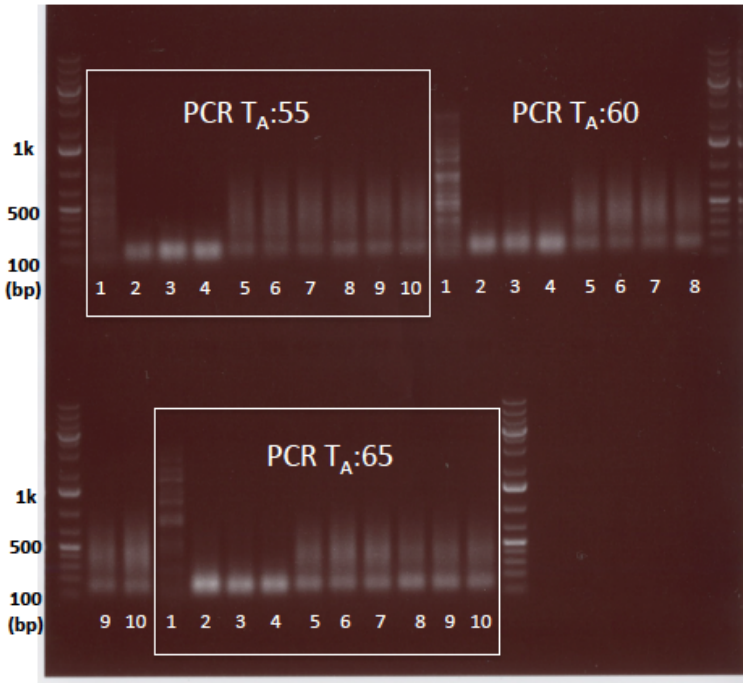
Initially, mRNA was prepared by direct lysis and then poly(A)-tailed mRNA was selected by oligotex beads (Qiagen) (M.M 2.6.2). Then, the mRNA of Bcl15 and Bulk-Puro cells was reverse transcribed with SA1 primer at three incubation

conditions of 42, 45 and 50 ^{o}C. Double-stranded DNA was synthesised from a single-stranded cDNA template using a forward primer with a 3' random octamer. Using a pair of primers which anneal to the overhang sequence on the RT and random octamer primer, PCR was performed to amplify DNA in the range of 100 bp to 1 kb. However, a smear of DNA bands was observed in both Bcl15 and Bulk-Puro cells. In addition, non-specific bands were also observed in the water control prepared for this PCR analysis (Fig. 5-4A).

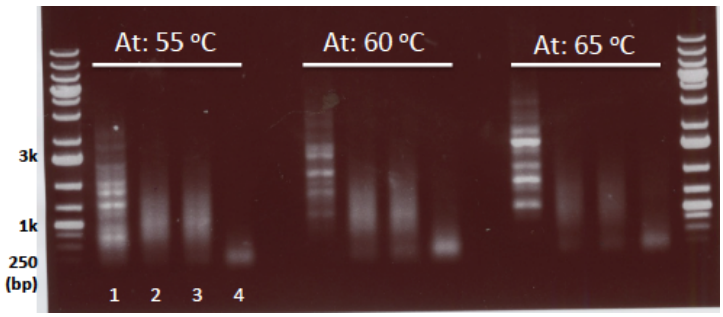
There are some steps in this pipeline that could be improved. For instance, the protocol for second strand DNA synthesis could be modified. As the primer for the second strand synthesis is a random octamer that requires low annealing temperature (Wong et al., 1996). Therefore, an incubation step at a lower temperature was added before the main incubation at 37 ^{o}C. However, this did not improve the final PCR outcome (Fig.5-4B). Hence, NGS sample preparation further upstream in the protocol, the method of mRNA extraction, was changed.

A

Lane #	Sample	RT temp
1	PCR NC (H_2O)	-
2	RT NC (-RTase)	50
3	RT NC (-RTase)	45
4	RT NC (-RTase)	42
5	Bcl15	50
6	Bcl15	45
7	Bcl15	42
8	Bulk-Puro	50
9	Bulk-Puro	45
10	Bulk-Puro	42



B



Lane#	Samples
1	PCR NC (no template)
2	Bcl15
3	Bulk-Puro
4	Bulk-Puro (-RTase)

Fig.5-4 Non specific amplification by KOD polymerase PCR after second strand DNA synthesis was not improved by optimising the incubation temperature of RT, double-stranded DNA synthesis and annealing temperature of the KOD polymerase-PCR.

(A) Specific amplification of DNA bands with the range of length from 100 to 1k bp was tested by KOD polymerase PCR. Prior to this PCR, mRNA from untransduced Bcl15 cells and Bulk-Puro was extracted by the direct cell lysis and selection of polyA tailed RNAs by oligotex beads (Qiagen) (M.M 2.6.2). 10 ng mRNA was reverse transcribed at three different RT temperatures (42, 45 and 50 ^{o}C), followed by double-stranded DNA synthesis. The double-stranded DNA has a tag sequence at both ends that were introduced by the RT primer and the primer to synthesise double-stranded DNA. PCR

amplification using tag sequences was performed by KOD polymerase (Novagen) at three different annealing temperatures (T_A: 55, 60 and 65 ^{o}C). As a negative control for the PCR, RT control (without RTase) and PCR water control were used. The electrophoresed gel (1 %) picture is shown. The lane number in the chart corresponds to that indicated in the gel. (B) We assumed that additional incubation steps at the synthesis of second strand cDNA could reduce unwanted background amplification. In order to increase the binding efficiency of random octamer, a 10-minute incubation at three different temperatures (15, 20 and 25 ^{o}C) were performed prior to the general incubation protocol (37 ^{o}C for 60 minutes followed by 75 ^{o}C for 20 minutes). KOD PCR was then tested at three different annealing temperatures (T_A: 55, 60 and 65 ^{o}C). The electrophoresed gel (1 %) picture is shown. The lane number in the chart corresponds to that indicated in the gel.

5.3.4 Optimisation of specific RNA amplification

We started from total RNA to extract mRNA by polyA selection (M.M 2.6.2). Reverse transcription was then performed, followed by the second strand DNA synthesis. In these experiments, I also included a bulk transduced population starved of IL-3 (Fig.5-1, Bulk-Survivors) as this might give further host-vector fusion transcripts that promote cell survival. The PCR showed a clear difference in amplification, showing minimal amplification in untransduced Bcl15 cells but smears of DNA bands in the expected size range in the other samples (Bulk-Puro, Bulk-Survivors and the IL-3I clone) (Fig. 5-5A). To keep a broad variety of DNA species, different low numbers of PCR cycles were tested, and 22 cycles proved to be adequate (Fig. 5-5B) (M.M 2.6.6). The final PCR samples were prepared in triplicate for each sample using 22 cycles, and Fig. 5-5C shows an aliquot of each of these final products. The samples were indexed using the Nextera XT Index Kit v2 set C (Illlumina) to identify products from each sample replicate so that samples could be included in a single sequencing run. The DNA fragments were sequenced using the 500-cycle paired-end sequencing kit (Illumina). The indexing and MiSeq run were kindly performed by Dr. Edward T. Mee (NIBSC) (M.M 2.6.6).

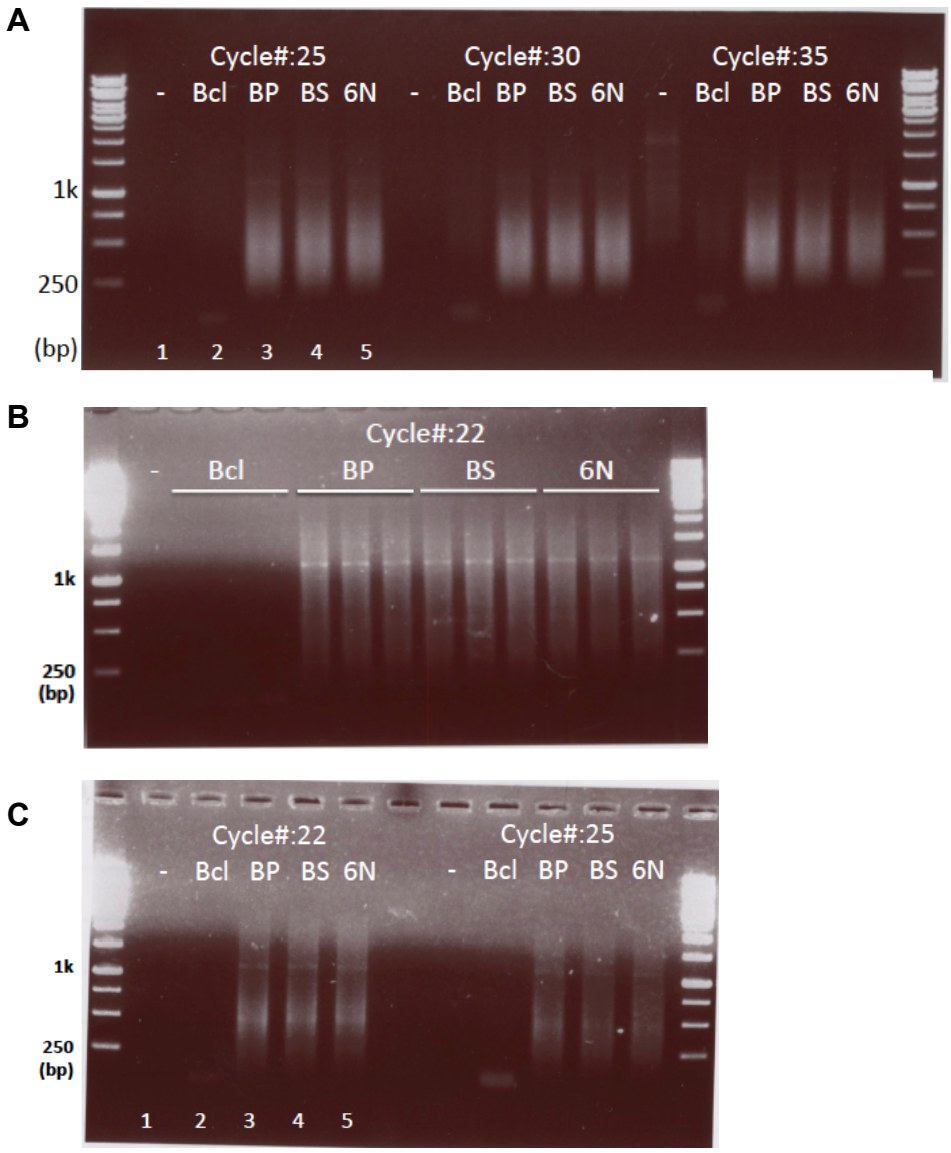


Fig.5-5 The amplification by KOD polymerase-PCR

(A) Four samples were tested: parental Bcl15 cells (Bcl), Bulk-Puro (BP), Bulk-Survivors (BS) and the IL-3I clone (6N). mRNA was prepared from total RNA employing poly(A) mRNA selection by oligotex (Qiagen). Reverse transcription on 100 ng mRNA of each sample was performed at 50 ^{o}C, followed by the second strand cDNA synthesis after pre-incubation at three different temperatures (15, 20 and 25 ^{o}C). KOD polymerase-PCR was performed at the annealing temperature of 52 ^{o}C with three different cycle numbers (25, 30 and 35). The PCR samples were run on a 1 % agarose gel. (B) In order to keep diverse species of PCR products, a lower PCR cycle number was tested (22) and compared to the amplification pattern with the cycle number 25. (C) The PCR amplification was performed at a cycle number 22 before indexing and performing RNA

sequencing (Dr. Edward T. Mee at NIBSC). Triplicate samples were prepared from each cell sample and aliquots of these PCR samples were run on a 1 % agarose gel.

5.3.5 Successful MiSeq run showed a small number of reads with vector specific sequences

The sequencing run generated a total of 14,867,601 reads passing filter, with 62.11 % of total bases having a Phred quality score of 30 (Q30) or greater, which is the benchmark of sequencing quality. Q30 shows the error rate of sequencing of 1 in 1000 chance. This number of reads and high quality showed the MiSeq run to have been successful. The raw data of RNA-Seq was obtained using the sequencing analysis tool, trimmomatic (http://www.usadellab.org/cms/?page=trimmomatic), which was kindly performed by Dr. Mark Preston (NIBSC). The average number of three replicates of each sample is described (Table 5-3 the left-hand column). The raw sequencing reads were filtered by a Phred quality score of 30 to obtain sequencing with a high accuracy (Ewing and Green, 1998). The number of raw reads and reads after the Q30 filtering are summarised as the total number of forward and reverse reads of each sample in Table 5.3 the middle column.

To investigate host-vector fusion mRNAs, the forward reads were firstly analysed. I chose 47 nt sequence including the additional 27 nt vector-specific sequence at 5' side of SA1. The 47 nt should be present on fusion mRNA whether the *Ghr* exon 2 SA that is immediate upstream of the 47 nt sequence is used or not because the sequence was read 3' downstream of the 47 nt. No cryptic splice sites were predicted in the 47 nt when analysed by Netgene2 in the Chapter 3. The forward sequence was filtered by the 47 nt sequence, kindly performed by Dr. Mark Preston (NIBSC) (Table 5-4 the left-hand column). The number of reads passing the filter was unexpectedly small compared to the total number of reads (Table 5-3, the left-hand column).

Ssample name	Reads Total	Reads Q30	Read Length (nt)
IL-3I cone	576218.7	258234.3	176.5
Bcl15	173222	22289.3	160.3
Bulk-Puro	493918	221837.3	176.5
Bulk-Survivors	570168.7	254106.7	174.7

Table 5-3 The summary of RNA-Seq reads

The sequence read number of four tested samples (the IL-3I clone, Bcl15 cell, Bulk-Puro and Bulk-Survivors) are summarised in three categories; total number of reads generated by RNA-Seq (reads total, the left-hand column), filtered reads passing Phred quality score of 30 (reads Q30, the middle column) and the average read length of filtered reads (read length, the right-hand column). Each number represents the average of three replicates of samples. The analysis of the reads were performed by a flexible read trimming tool for Illumina NGS data trimmomatic (http://www.usadellab.org/cms/?page=trimmomatic), which was kindly performed by Dr. Mark Preston (NIBSC).

Sample name	Vector reads
The IL-3I clone	147.3
Bcl15	5
Bulk-puro	60
Bulk-survors	117

Table 5-4 The number of forward sequence reads with the 47 nt vector sequence

The forward sequence reads with the 47 nt vector sequence were extracted from total forward sequence reads, which was kindly performed by Dr. Mark Preston (NIBSC) using trimmomatic (http://www. usadellab.org/cms/?page= trimmomatic) as in the Table 5-3. The number in this table is the average read number of three replicates of each sample.

All the filtered reads with the 47 nt were then clustered by Clustal X (Jeanmougin

et al., 1998) for fast screen of possible fusion mRNAs (M.M 2.6.6). The two typical alignments are shown in Fig.5-6. In the majority represented by Fig.5-6 (blue), only vector sequences that multiply spliced within the vector were detected. In minority sequences represented by Fig.5-6 (orange), the vector sequence at the 5' side joined the 3' region of sequence that did not align to the vector sequence. One representative sequence was then selected in each clustered read. The FASTA format of the accumulated representative sequence reads was blasted to host mouse genomic sequence (GRCm38.p4) using Ensembl (http://www.ensembl.org/ index. html) to see if a mouse sequence is present on the reads. In addition, to detect splicing patterns within vectors and the boundary between the host and vector sequence, the FASTA sequence was blasted to mouse genomic sequence (GRCm38.p4) and pGhr IRES-Puro provirus sequence by NCBI nucleotide blast (https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome). Because the read is paired-end, the corresponding reverse reads to the minority sequences that had host sequences on were blasted to host mouse genomic sequence (GRCm38.p4) using Ensembl (http://www.ensembl.org/ index. html) to see whether the same gene locus appears as in the forward reads. Once this process is done, separately, all reverse reads were blasted to the mouse genomic sequence (GRCm38.p4) using NCBI nucleotide blast (https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome) to search any new host mouse sequence because the reverse reads were primed by a random octamer (Fig.5-2C). However, we could obtain more numbers of reverse reads that had the already identified host sequences but not find such new gene locus by this approach. The details of the identified host sequences and the splicing patterns are described in the next section 5.4.7.

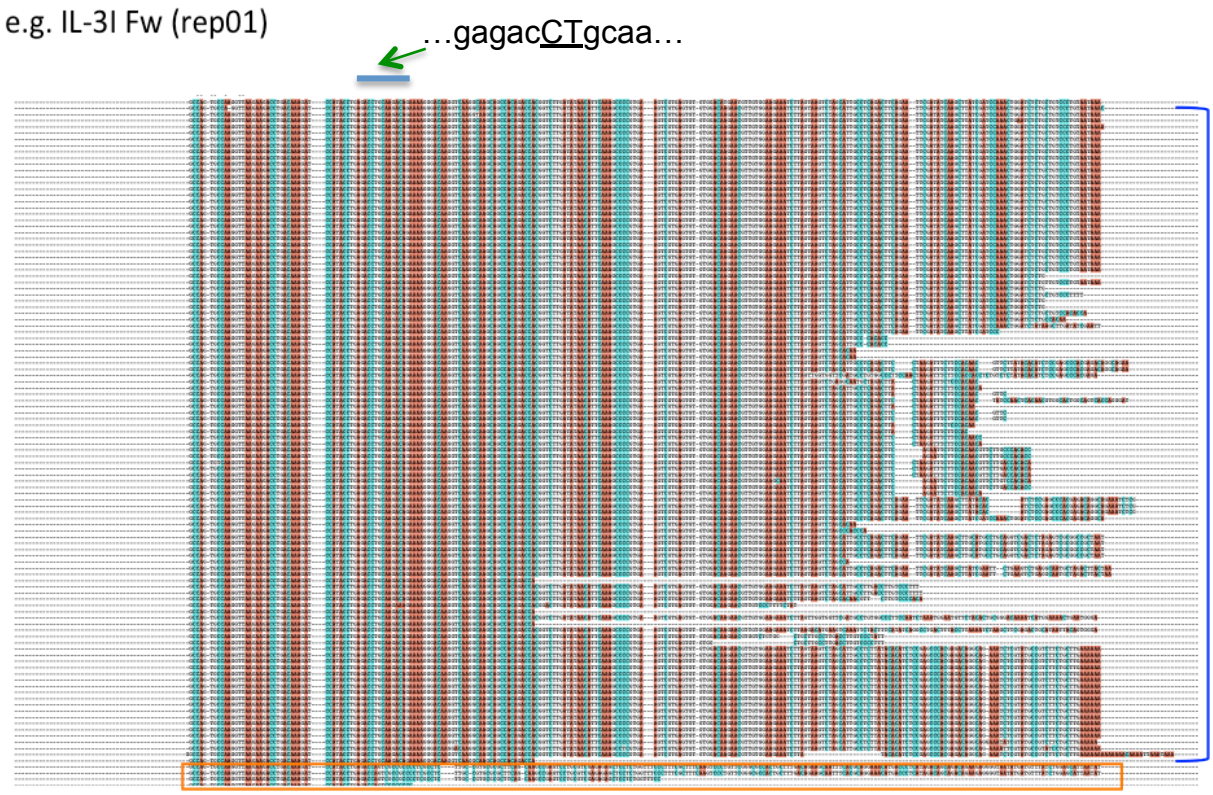


Fig.5-6 The overview of sequence reads clustered by ClustalX (e.g one of three replicates from the IL-3I clone)

Forward (Fw) sequence reads of the IL-3I clone were clustered using ClustalX with a basic default setting (M.M 2.6.6). The majority of sequences with the vector sequence (blue) and the minor sequences having host mouse sequences (orange) are clustered as shown. The output of Fw sequence reads is reverse reading. The sequence in the *Ghr* exon 2 SA (CT in the sequence) and its position are highlighted by the blue bar.

5.3.6 Identity of host sequences in fusion transcripts

A summary of the host mRNAs found in fusion transcripts is presented in Table 5-5. I separately reported the number of sequences with each host sequence in three categories as a result of each blasting step described in 5.3.5. The forward sequence was firstly analysed by clustering and blasting, and, the number of sequence hits were counted. Then we investigated the reverse reads paired to the forward sequence that had host cellular sequence. The number of reverse reads that hit the same gene locus were obtained. Separately, by blasting all reverse sequences we found more reverse reads with the already identified host cellular sequences.

To obtain the known gene functions, the gene ontology data was extracted from the UniProt online database (*www.uniprot.org/*) (Table 5-6). This database provides protein sequences and their functional annotations (Consortium, 2010). Biological and molecular functions of each gene as extracted are described. I also examined whether any of these genes had been reported to serve a survival or proliferation function in the BAF3 cells (summarized in Table 5-1) and whether any appeared in the RTCGD and CIS. This very limited set of data does suggest that the genes detected in the Bulk survivor population either have a known role in BAF3 cell survival, like transferrin receptor (Shi et al., 1997), or appear in the RTCGD or CIS, whereas those in the control group (Bulk-Puro) do not. For the clone IL-3I the detection of an *Angpt1* fusion transcript confirms that this insertion site detected by LM-PCR (Chapter 4) does give rise to a fusion transcript.

Sample	Gene symbol	Gene name	Read hits Fw	Read hits Rv	Read hits Rv new	Potential total read number	Chromosome number	Activity in Baf3 assay	RTCGD hits	CIS hits
The IL-3I clone	Angpt1	Angiopoietin 1	12	4	2	14	15	-	2	-
	Mtpn	Myotrophin	5	1	4	9	6	-	-	-
Bulk-Puro	Ctsc	Cathepsin C	11	2	3	14	7	-	2	-
	Pabpc1	Poly(A) Binding Protein, Cytoplasmic 1	4	3	1	5	15	-	-	-
	Nudt5	Nudix (Nucleoside Diphosphate Linked Moiety X)-Type Motif 5	1	0	0	1	2	-	-	-
	Mtpn	Myotrophin	1	0	0	1	6	-	-	-
	Arid2	AT Rich Interactive Domain 2	1	1	0	1	15	-	-	-
Bulk-Survivors	Hsp90ab1	Heat Shock Protein 90kDa Alpha (Cytosolic), Class B Member 1	2	1	9	11	17	-	-	-
	Actb	Actin, Beta	3	1	0	3	5	-	9	9
	Tfrc	Transferrin Receptor	2	2	0	2	16	+	-	-
	Eif4a2	Eukaryotic Translation Initiation Factor 4A2	0	0	1	1	16	-	2	2

Table 5-5 The summary of identified genes in fusion transcripts and its analysis of sequence reads

The number of sequence reads with the listed host cellular sequences was obtained as a result of analysis as in Fig.5-6 and blasting (M.M 2.6.6). The first column on the right side of “Gene name” shows the number of forward sequence reads. The second one shows reverse sequence reads that was identified from the corresponding sequence of the forward reads in the first column. The third one is the reverse reads obtained by blasting all reverse reads against mouse genomic sequence (GRCm38.p4) by NCBI nucleotide blast. The fourth one is the sum of the first and the third columns, which presents the total number of sequence reads of each identified host sequence.

Sample	Gene symbol	Biological function	Molecular funtion	Ref*
The IL-3I clone	Angpt1	negative regulation of apoptotic process regulation of tumor necrosis factor production positive regulation of ERK1 and ERK2 cascade regulation of I-kappaB kinase/NF-kappaB signaling	receptor binding receptor tyrosine kinase binding vascular endothelial growth factor receptor binding	1
	Mtpn	positive regulation of cell growth positive regulation of NF-kappaB transcription factor activity positive regulation of cardiac muscle hypertrophy	sequence-specific DNA binding	2
Bulk-Puro	Ctsc	aging positive regulation of proteolysis involved in cellular protein catabolic process response to organic substance	chloride ion binding cysteine-type peptidase activity	3
	Pabpc1	nuclear-transcribed mRNA catabolic process, nonsense-mediated decay positive regulation of viral genome replication RNA splicing	mRNA binding poly(A) binding protein C-terminus binding	4
	Nudt5	nucleobase-containing compound metabolic process ribonucleoside diphosphate catabolic process	ADP-ribose diphosphatase activity nucleoside-diphosphatase activity snoRNA binding	5
	Mtpn	positive regulation of cell growth positive regulation of NF-kappaB transcription factor activity positive regulation of cardiac muscle hypertrophy	sequence-specific DNA binding	2
	Arid2	nucleosome disassembly positive regulation of transcription from RNA polymerase II promoter	DNA binding metal ion binding transcription factor binding	6
Bulk-Survivors	Hsp90ab1	negative regulation of apoptotic process positive regulation of protein serine/threonine kinase activity regulation of interferon-gamma-mediated signaling pathway regulation of type I interferon-mediated signaling pathway virion attachment to host cell	ATP binding poly(A) RNA binding GTP binding	7
	Actb	ATP-dependent chromatin remodeling platelet aggregation	ATP binding kinesin binding Tat protein binding	8
	Tfrc	cellular response to extracellular stimulus acute-phase response aging	poly(A) RNA binding double-stranded RNA binding iron ion transmembrane transporter activity	9
	Eif4a2	negative regulation of RNA-directed RNA polymerase activity regulation of translational initiation	ATPase activity poly(A) RNA binding translation initiation factor activity	10

Table 5-6 The summary of functional features of identified genes in fusion transcripts

Biological and molecular function were extracted UniProt (http://www.uniprot.org/).

Table 5-6 (continued)

Ref*: reference; 1: (Davis et al., 1996); 2: (Pennica et al., 1995); 3: (McGuire et al., 1997) ; 4: (Wang et al., 1992); 5: (Yang et al., 2000); 6: (Wilsker et al., 2005); 7: (Mazzarella and Green, 1987); 8: (Alonso et al., 1986); 9: (Stearne et al., 1985) ;10: (Nielsen and Trachsel, 1988)

5.3.7 Use of splice sites in the fusion transcripts.

I examined the use of vector splice sites in the fusion transcripts by aligning the sequencing reads with vector and host sequences, these alignments are shown in Fig.5-7.

The *Angpt1* and *Nudt5* loci were found to have a splice-out fusion mRNA. The vector insertion at these loci was found in the reverse orientation relative to the host gene transcriptional direction. In the reverse sequence of the *Ghr* intron 1, a cryptic splice donor (SD) at 1847 of vector provirus sequence was detected which formed a fusion transcript with the splice acceptor (SA) of exon 9 of *Angpt1* or exon 2 of Nudt5. This cryptic SD at the position of 1847 was not predicted by Netgene2 *in silico* analysis (Chapter 3). Since the reverse transcription primer was designed to anneal downstream of the *Ghr* exon 2, the full sequence of these transcripts – including the transcriptional starting site – is unknown.

Many of the other loci showed splice-in between a SD in the host gene and cryptic SAs within (232) or near the 5' LTR (294). The Netgene2 analysis in Chapter 3, only site 294 appeared as a predicted SD with 0.16 confidence and others were even predicted. The 232 and 294 SAs were detected in a genome wide screen for LV fusion transcripts previously (Cesana et al., 2012). Following this splice-in, in 6 transcripts the HIV major SD generated 2 further introns between the SD and the known HIV-derived SA at 1507 (Chapter 3 and (Cesana et al., 2012)), and between the cryptic SD at 1614 (Cesana et al., 2012) and the *Ghr* exon 2 SA. In a further six transcripts, the HIV major SD spliced

directly to the *Ghr* exon 2 SA. A direct splice from a cellular SD to the *Ghr* exon 2 SA was only found in three transcripts. In many gene loci, several of the different splicing patterns were observed.

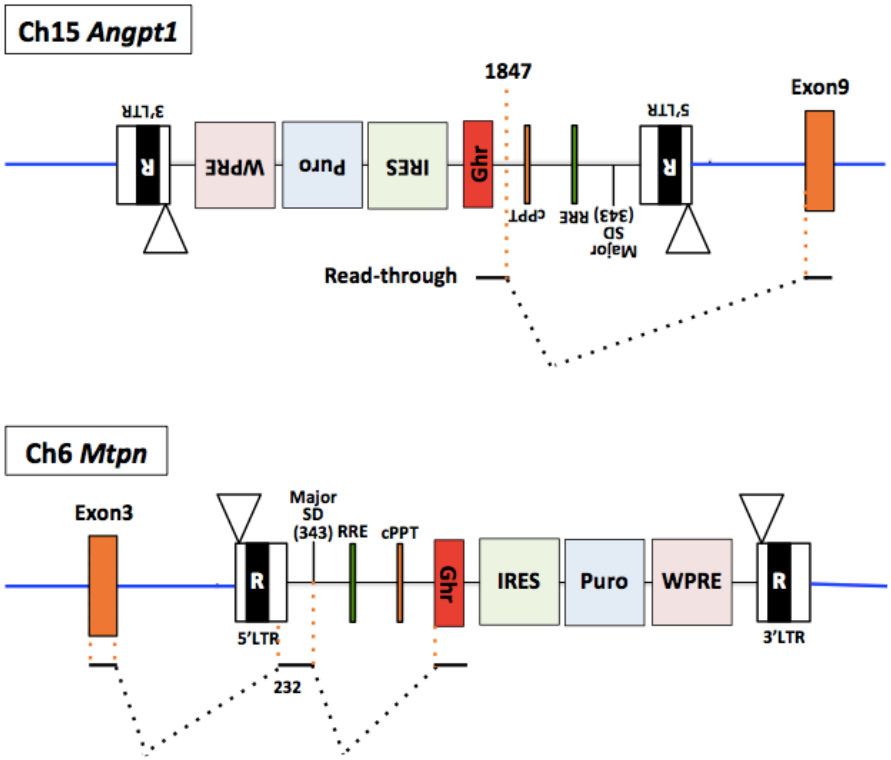
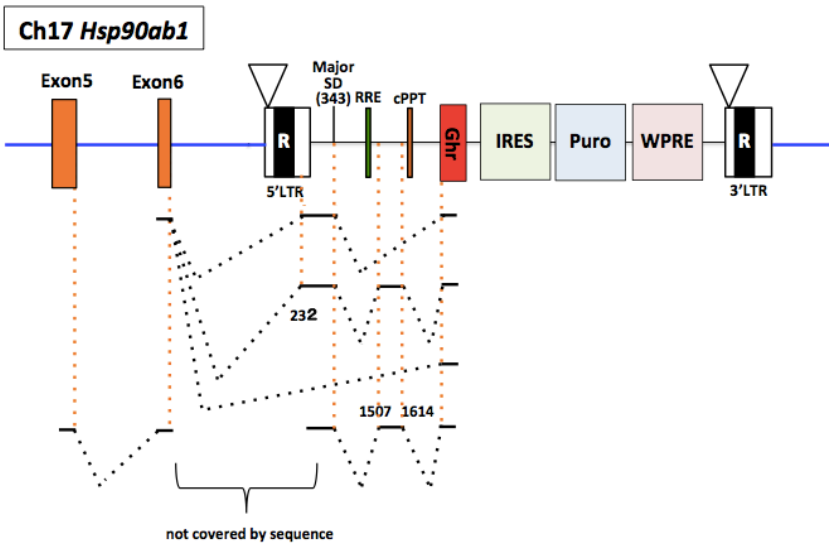
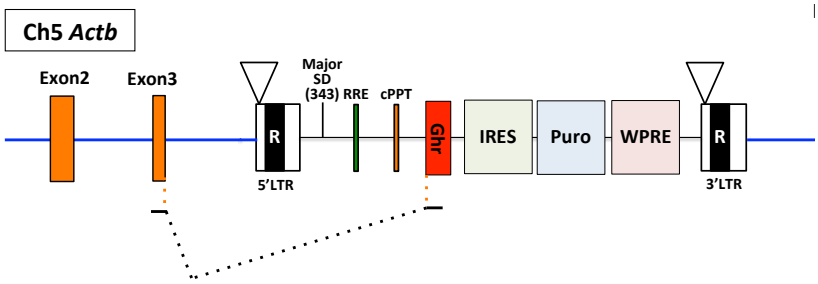
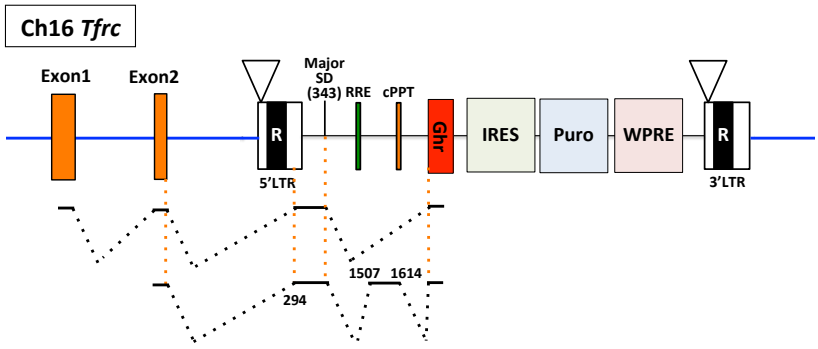


Fig.5-7 *Angpt1* and *Mtpn* were present in the host-virus fusion mRNAs expressed in the IL-3I clone.

The blast result showed that *Angpt1* and *Mtpn* were present in fusion mRNA expressed in the IL-3 clone. The vector provirus integration is shown in the each gene locus. Splicing pattern was estimated by obtained sequence. Thick black line shows the region presented on the detected fusion mRNA. The cryptic SA 232, or cryptic SD 1847 in the vector sequence is indicated. Chromosomal number of each gene locus (Ch) is shown next to the gene name.


Ch17 Hsp90ab1
Exon5
Exon6
Major SD
(343) RRE
cPPT
R
5'LTR
Ghr
IRES
Puro
WPRE
R
3'LTR
232
1507
1614
not covered by sequence


Ch5 Actb
Exon2
Exon3
Major SD
(343) RRE
cPPT
R
5'LTR
Ghr
IRES
Puro
WPRE
R
3'LTR


Ch16 Tfrc
Exon1
Exon2
Major SD
(343) RRE
cPPT
R
5'LTR
Ghr
IRES
Puro
WPRE
R
3'LTR
1507
1614
294

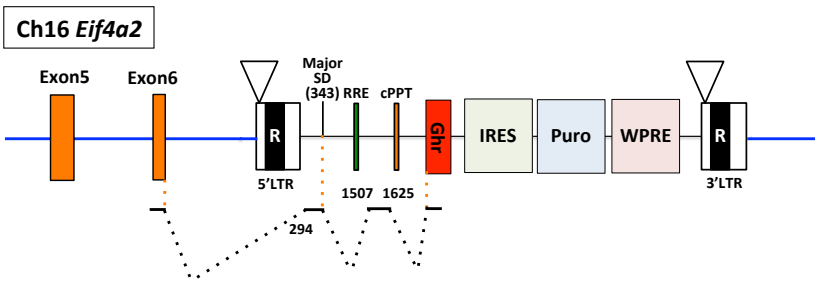
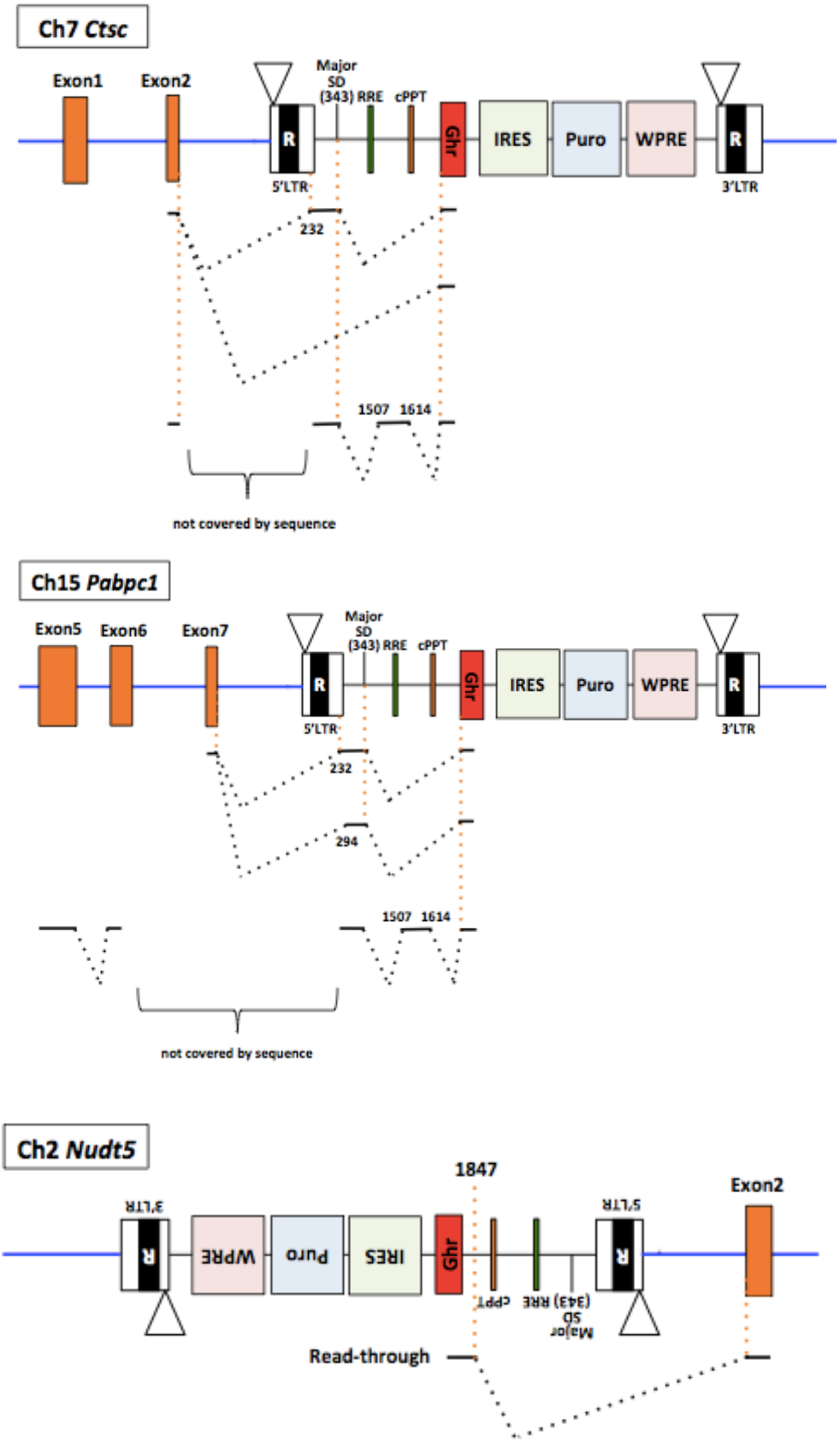

Ch16 Eif4a2
Exon5
Exon6
Major SD
(343) RRE
cPPT
R
5'LTR
Ghr
IRES
Puro
WPRE
R
3'LTR
1507
1625
294

Fig.5-8 Bulk-Survivors expressed fusion mRNA in *Hsp90ab1*, *Actb*, *Tfrc* and *Eif4a2* locus.

The vector provirus integration is shown in each gene locus. The splicing pattern was estimated by an obtained sequence. The thick black line shows the region presented on the detected fusion mRNA. Multiple splicing patterns were observed in the identified fusion mRNAs. In addition to the splice acceptor (SA) that is boundary of host-vector sequence, other splice sites on vector sequence were also indicated. Chromosomal number of each gene locus (Ch) is indicated next to the gene name.


Ch7 Ctsc
Exon1
Exon2
Major SD (343)
RRE
cPPT
Ghr
IRES
Puro
WPRE
R
5'LTR
3'LTR
232
1507
1614
not covered by sequence
Ch15 Pabpc1
Exon5
Exon6
Exon7
Major SD (343)
RRE
cPPT
Ghr
IRES
Puro
WPRE
R
5'LTR
3'LTR
232
294
1507
1614
not covered by sequence
Ch2 Nudt5
1847
Exon2
3'LTR
WPRE
Puro
IRES
Ghr
cPPT
RRE
Major SD (343)
5'LTR
R
Read-through

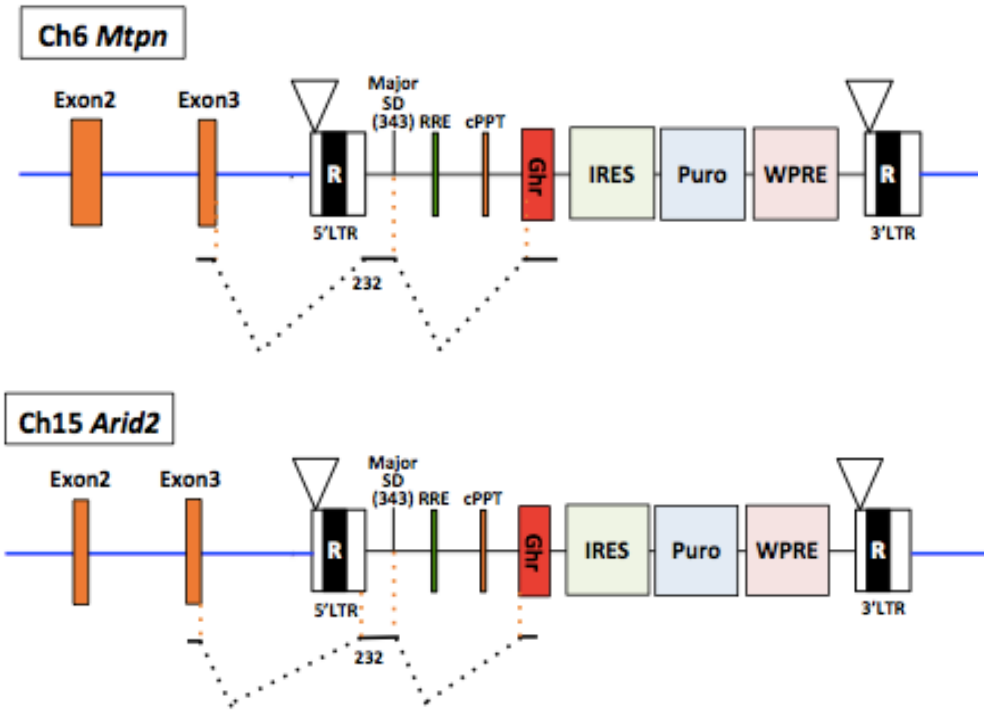


Fig.5-9 Bulk-Puro expressed fusion mRNA in *Ctst*, *Pubpc1*, *Nudt5*, *Mtpn* and *Arid2* locus.

The vector provirus integration is shown in each gene locus. The splicing pattern was estimated by an obtained sequence. The thick black line shows the region presented on the detected fusion mRNA. Multiple splicing patterns were observed in the identified fusion mRNAs. In addition to the splice acceptor (SA) that is boundary of host-vector sequence, other splice sites on vector sequence were also indicated. The chromosomal number of each gene locus (Ch) is indicated next to the gene name.

5.4 Discussion

The identification of two fusion transcripts in the clone IL-3l brings the total number of detected integrants in this clone to 15. *Angpt1* was the only gene identified by more than one method, in this case LM-PCR for integration site (Fig.4-5, Chapter 4) and also vector-primed RNA-Seq (Fig.5-7). It is clearly likely that many integration sites may not generate fusion transcripts. Also the LM-PCR dataset is incomplete as *Mtpn* does not appear there. The time limit for my PhD meant that I was unable to perform further experiments after the RNA-Seq run. So at this point I can only propose candidates that may be involved in clone IL-3l survival.

5.4.1 *Angpt1* can be the potential candidate gene related to IL-3 independence of the clone IL-3l

Angpt1 is clearly one good candidate for promoting survival or proliferation of the clone IL-3l and this gene loci was found by LM-PCR (Fig.4-5B). Firstly, Angpt1 can act as a survival factor for haematopoietic progenitor cells (Arai et al., 2004). Secondly *Angpt1* has two insertion sites and its receptors *Tie1/Tek* have three insertion sites in the RTCGD. So while the most reported effect of Angpt1 are on vascularisation (reviewed in (Yancopoulos et al., 2000)), an effect on a haematopoietic progenitor cell line is not unlikely. It is hard to predict the effect of the LV insertion in *Angpt1* on the expression of the protein without further information. It is possible that the insertion reduces protein expression from this allele, however it is also possible that the fusion transcript is more stable and increases protein expression. Time limitation prevented me from confirming this hypothesis during my PhD, but this work will now be completed in the lab. Key experiments will be to examine whether *Angpt1* mRNA level differs between Bcl15 and the clone IL-3l cells, whether the clone IL-3l produces Angpt1 and whether recombinant Angpt1 can promote survival or growth of Bcl15 cells.

It should be noted that the *Angpt1* fusion mRNA does not encode puroR

therefore this must come from a puroR expressing integration site, likely *Mtpn* or *Myh9* in the IL-3I clone. It is also possible that other integration sites in the clone IL-3I have fusion transcripts, not detected because the RNA-Seq coverage is partial, or that LV insertion in other genes regulates gene expression in the absence of fusion transcripts. A future plan would be to screen expression of mRNA from each of the 15 integration sites, perhaps starting with those in the RTCGD. It is also clearly possible that more than one of these integration sites contributes to survival or proliferation of the clone IL-3I. I designed the original experiment to transduce Bcl15 cells at high copy number so that a large number of integration sites would be generated. The isolation of only one clone in these experiments demonstrates that this model LV is not highly mutagenic.

5.4.2 Host genes in analysed fusion transcripts by RNA-Seq were related to Bcl15 cell survival that was previously reported

The Bulk-Survivors are derived from Bulk-Puro cells which then survived IL-3 starvation. Again the low number of total sequence reads for fusion mRNA analysis will have prevented the full repertoire of host sequences being identified. However, as an example of shared host sequences, a *Mtpn*-vector fusion mRNA was detected in the same splice-in form in both the IL-3I clone and Bulk-Puro cells. Thus the *Mtpn* locus may be a 'hot spot' for efficient splice-in fusion mRNA expression. Intriguingly, the genes identified in Bulk-Survivors had all been identified as relevant to Bcl15 cell survival or tagged in the RTCGD, whereas those in Bulk-Puro have not. Firstly, Tfrc has been shown to be a potent survival factor for BAF3 cells (Shi et al., 1997), so up-regulation of its receptor would provide a simple survival enhancement mechanism. Secondly, HSP90 is a molecular chaperone that controls the stability and function of a number of proteins in various signaling cascades, therefore its upregulation is related to cancer growth and survival (Neckers, 2007). Breast cancer is one example. The cytoplasmic isoform of HSP90, HSP90AB1, is implicated with poor prognosis in different subtypes of breast cancer including the aggressive triple negative breast cancer (TNBC) subtype (Cheng et al., 2012). Eif4a2 is overexpressed in

some tumour types including gliomas (Oblinger et al., 2016) and lymphomas (Gupta et al., 2002). It is a factor required for efficient initiation of protein synthesis and Eif4a2 inhibitors have been proposed as anti-cancer agents (Raza et al., 2015). Finally, beta actin is a common cytoskeletal component playing a key role in cell motility and migration, to my knowledge it is not known to be regulated at the transcriptional level in cancer. However it does appear 10 times as an upregulated gene in the RTCGD, with 7 of these studies involving B cell tumours (Suzuki et al., 2002) (Suzuki et al., 2006). To come to any conclusion about the relevance of these sites it would be necessary to repeat the RNA-Seq experiment to achieve greater numbers of Bulk-Survivor and Bulk-Puro sequences so that I could estimate whether there was a significant difference between appearance of RTCGD hits in the Bulk-Survivor and Bulk-Puro population. I could also overexpress or knock-down expression of candidate genes in Bcl15 cells to see if this affected cell survival.

5.4.3 Variety of fusion mRNAs in the same gene locus were observed

The majority of fusion mRNAs had splice-in forms using a SA site within or near the 5' LTR in the vector sequence. The use of a different SA within the vector in the same gene locus suggests that one gene locus affords variety of alternative splicing. It is also possible that the different transcripts come from different cell clones that harbor different integration sites. To distinguish between these possibilities I would need to perform sensitive PCR, with vector and integration site primers to determine whether multiple LV integrations are present in the identified gene loci. Although the most forward sequence reads showed read-through (Fig.5-6), the isolation of fusion mRNAs spliced between a host SD the *Ghr* exon 2 SA suggests the *Ghr* exon 2 SA contributes to interacting with a host SD as designed. Such fusion transcripts could contribute the increased transgene expression of pGhr IRES-Puro compared with pIRES-Puro that was reported in Chapter 3 (Fig.3-10D), but another RNA-Seq on the both transduced construct could explain this possibility by quantitating such fusion reads.

5.4.4 The identification of cryptic splice sites; the antisense integration raised the cryptic SD in the introduced intron sequence of the *Ghr* exon 2

Vector integration direction relative to the host gene seems to affect splicing patterns, either splice-in or splice-out. The marker gene expression of our promoterless vector is detected when splice-in mRNA generation occurs and the transcript is translated from a host promoter. By analysing host-vector fusion transcripts in this chapter we found cryptic splice sites within the vector sequence in the sense integration to cellular transcript. The cryptic SA within (232) 5' LTR was identified in this study, which was previously found in (Cesana et al., 2012).

Interestingly, vector integration in the reverse orientation such as *Angpt1* and *Nudt5* showed a splice-out fusion mRNA using a cryptic SD (1847) on the negative strand of the vector sequence. The SD was present within the cloned *Ghr* locus fragment. Therefore it was artificially generated but not predicted by NetGene2. The SD would be possibly used in other locus and a completed RNA-Seq run could answer this possibility because the identified host genes and splicing pattern by RNA-Seq in this study are partial. This observation showed that exogenous sequence in a vector can be cryptic splice sites, therefore all these elements need to be assessed for genotoxic potential in respect to aberrant splicing (Knight et al., 2010) (Knight et al., 2012). In this regard the biological assay is more valuable than a computer prediction and should be carried out to find any cryptic splice sites.

6 General discussions

6.1 Using these results to generate safer LV

My results have identified a number of cryptic splice sites in the LV pGhr IRES-Puro, detailed in the results of Chapter 4 and 5. All but one of these was detected in the LV inserted in the same orientation as gene transcription, and likely gave rise to puromycin resistance. Some of these were previously described by Cesana et al (Cesana et al., 2012), but those SDs at 1614 and 1625 were not previously reported. The two assays differ in that Cesana et al analysed fusion transcripts from all integrants, whereas I have analysed those that give rise to puromycin expression. All of these, apart from the cryptic SDs, are in regions of the LV that are required for vector function, namely the LTR (232, 235 and 294), HIV-1 major SD (MSD) (343), and SA within Env (1507). So, to produce safer vectors all these cryptic splice sites should be mutated in pGhr IRES-Puro. In the study of Cesana et al (Cesana et al., 2012), a mutation in the HIV-1 MSD critically reduced background transcripts although virus infectivity also declined greatly. This is not surprising as the MSD is known to be critical for inhibition of polyadenylation at the 5' LTR (Furger et al., 2001) and more subtle mutations in this site should be tested. Mutations of multiple splice sites, including the SA within Env (1507), were also performed but the transcript expression affected by that specific SA was not investigated. In our vector case, the titre of the novel LV would need to be checked to make sure that the combination of mutations did not affect vector function and another splicing assay should be performed to check that *Ghr* exon 2 SA is then used.

I also detected one cryptic SD in the reverse orientation LV insertion into *Angpt1* in the clone IL-3l, reported in Chapter 5, which did not lead to puromycin resistance (5.4.4). In this case the SD was in the intron 1 of *Ghr* that was cloned with the exon 2 SA and is not found in other LVs so is not relevant for future mutation. However, it would be valuable to perform a similar assay with the promoterless puromycin gene inserted into the reverse orientation of a LV, the vector pGhr IRES-Puro reverse as described in Chapter 3. This would allow

puromycin selection to then identify cryptic SAs in the reverse orientation. It is also possible that removal of the *Ghr* exon 2 SA from the test vectors would reveal further cryptic SA sites.

Finally, my results, and those of Cesana et al (Cesana et al., 2012), were obtained in different cell lines; in their case a human lymphoblastoid cell line and primary haematopoietic cells, in my case a mouse haematopoietic cell line. It is still unclear why a particular splice acceptor is chosen in a particular integration locus. Also, as many of my fusion transcripts were cloned from a bulk population, I don't know whether a single integration locus may choose more than one SA. Therefore, to optimise vector safety a more extensive screen of candidate clinical LV should probably be carried out in the relevant gene therapy target cell population using extensive RNA-seq analysis.

6.2 Improvement of the mutagenesis assay

I initially hoped to identify a locus where LV insertion led to IL-3 independence via a splice-in mechanism. This would have allowed a quantitative assay for splice-in mutation to be developed in the same way as our group has used the *Ghr* locus as a quantitative assay for LV splice-out mutation (Bokhoven et al., 2009) (Knight et al., 2010) (Knight et al., 2012). However, I only isolated one IL-3 independent clone (IL-3I). In this clone I have identified 13 integrants and two species of a fusion transcript, in the *Angpt1* and *Mtpn* loci. The *Mtpn*-vector fusion transcripts result from a splice-in event and are likely to be capable of giving rise to puromycin expression. In contrast, the *Angpt1*-fusion transcripts result from a cryptic SD in the reverse orientation, and thus cannot give rise to puromycin expression.

My hypothesis at present is that these two fusion transcripts contribute to puromycin resistance and IL-3 independence, respectively. The fact that two hits are needed for this clone to emerge may explain why I did not isolate a similar clone when I repeated the assay at a lower multiplicity of transduction. In any case I did not identify a candidate locus that could render cells IL-3 independent

by a splice-in mechanism. In future assays, if more IL-3 independent clones could be isolated, the mechanism of the IL-3 independence could be compared. I consider that the puromycin selection step could be improved for this purpose. Throughout our IM assay, 1 μg/ml puromycin was kept in the culture. This puromycin concentration indeed selected puroR cells over a certain threshold of puromycin expression (Fig.3-10A). This may select puroR cells with strong puroR gene expression, but this would not necessarily affect cell transformation. The LV integration and perturbation of the neighboring gene related to cell survival is more important than transduced cells with strong puromycin resistance. In this regard, puromycin drug selection could be omitted.

6.3 The implication of these results for IL-3 signalling

Angpt1 regulation remains a good candidate for rendering Bcl15 cells IL-3 independent. As explained in Chapter 5, Angpt1 is a growth factor which supports survival and proliferation of haematopoietic progenitor cells (Huang et al., 2010) (Haemmerle et al., 2014). The IL-3l clone secretes a growth factor, making this possible up-regulation the most likely mechanism. However, *Angpt1* also appears in the RTCGD as a down-regulated gene linked to tumorigenesis in a mouse model (Haemmerle et al., 2014). Sadly the time of my PhD funding expired before I could follow up these ideas. The obvious next step would be to determine whether *Angpt1* is up-regulated or down-regulated in IL-3l cells compared to Bcl15 cells. This would be done by qRT-PCR compared to a house-keeping gene. Depending on the result, the next step would be to either over-express *Angpt1* in Bcl15 cells or inhibit its expression and then examine cell survival and proliferation. It is hard to predict which possibility is the most likely result of the LV anti-sense insertion in the *Angpt1* locus. In either case this would be a novel result, bringing an intersection between the angiopoietin signalling pathway and the IL-3 signalling pathway.

In our assay we use IL-3 dependence to screen genes related to cell proliferation and survival. The identified genes isolated in our IM assays might be limited by gene expression in murine pro-B Bcl15 cells. To expand our insight

about splice-in IM by LV integrations, we can apply our vector model to other cell types. For instance, recently established murine IL-7 dependent cell line MOHITO (mouse hematopoietic interleukin-dependent cell line) is a candidate because splice-in IM can be investigated in the T-cell context using cytokine deprivation (Kleppe et al., 2011). In addition, investigating an effect of vector integration in human primary cells, including embryonal and induced pluripotent stem cells (ES and iPS cells), will give an insight about vector genotoxicity in the closer setting to clinical gene therapy.

7 Bibliography

Agostini, I., Popov, S., Li, J., Dubrovsky, L., Hao, T., Bukrinsky, M., 2000. Heat-shock protein 70 can replace viral protein R of HIV-1 during nuclear import of the viral preintegration complex. Exp Cell Res 259, 398-403.

Ahmed, N.N., Grimes, H.L., Bellacosa, A., Chan, T.O., Tsichlis, P.N., 1997. Transduction of interleukin-2 antiapoptotic and proliferative signals via Akt protein kinase. Proc Natl Acad Sci U S A 94, 3627-3632.

Aiuti, A., Biasco, L., Scaramuzza, S., Ferrua, F., Cicalese, M.P., Baricordi, C., . . . Naldini, L., 2013. Lentiviral Hematopoietic Stem Cell Gene Therapy in Patients with Wiskott-Aldrich Syndrome. Science 341.

Aiuti, A., Slavin, S., Aker, M., Ficara, F., Deola, S., Mortellaro, A., . . . Bordignon, C., 2002. Correction of ADA-SCID by stem cell gene therapy combined with nonmyeloablative conditioning. Science 296, 2410-2413.

Akagi, K., Suzuki, T., Stephens, R.M., Jenkins, N.A., Copeland, N.G., 2004. RTCGD: retroviral tagged cancer gene database. Nucleic acids research 32, D523-527.

Alekseyenko, A.V., Kim, N., Lee, C.J., 2007. Global analysis of exon creation versus loss and the role of alternative splicing in 17 vertebrate genomes. Rna 13, 661-670.

Alonso, S., Minty, A., Bourlet, Y., Buckingham, M., 1986. Comparison of three actin-coding sequences in the mouse; evolutionary relationships between the actin genes of warm-blooded vertebrates. Journal of molecular evolution 23, 11-22.

Ambrose, Z., Aiken, C., 2014. HIV-1 uncoating: connection to nuclear entry and regulation by host proteins. Virology 454-455, 371-379.

Antoniou, M., Harland, L., Mustoe, T., Williams, S., Holdstock, J., Yague, E., . . . Crombie, R., 2003. Transgenes encompassing dual-promoter CpG islands from the human TBP and HNRPA2B1 loci are resistant to heterochromatin-mediated silencing. Genomics 82, 269-279.

Arai, F., Hirao, A., Ohmura, M., Sato, H., Matsuoka, S., Takubo, K., . . . Suda, T., 2004. Tie2/angiopoietin-1 signaling regulates hematopoietic stem cell quiescence in the bone marrow niche. Cell 118, 149-161.

Arechavala-Gomeza, V., Khoo, B., Aartsma-Rus, A., 2014. Splicing modulation therapy in the treatment of genetic diseases. The application of clinical genetics 7, 245-252.

Arhel, N., 2010. Revisiting HIV-1 uncoating. Retrovirology 7, 96.

Arhel, N.J., Souquere-Besse, S., Munier, S., Souque, P., Guadagnini, S., Rutherford, S., . . . Charneau, P., 2007. HIV-1 DNA Flap formation promotes uncoating of the pre-integration complex at the nuclear pore. The EMBO journal 26, 3025-3037.

Artamonova, I.I., Gelfand, M.S., 2007. Comparative Genomics and Evolution of Alternative Splicing: The Pessimists' Science. Chemical Reviews 107, 3407-3430.

Ast, G., 2004. How did alternative splicing evolve? Nature reviews. Genetics 5, 773-782.

Astrakhan, A., Sather, B.D., Ryu, B.Y., Khim, S., Singh, S., Humblet-Baron, S., . . . Rawlings, D.J., 2012. Ubiquitous high-level gene expression in hematopoietic lineages provides effective lentiviral gene therapy of murine Wiskott-Aldrich syndrome. Blood 119, 4395-4407.

Aubourg , P., Blanche , S., Jambaqué , I., Rocchiccioli , F., Kalifa , G., Naud-Saudreau , C., . . . Bougnères , P.-F., 1990. Reversal of Early Neurologic and Neuroradiologic Manifestations of X-Linked Adrenoleukodystrophy by Bone Marrow Transplantation. New England Journal of Medicine 322, 1860-1866.

Azzouz, M., Martin-Rendon, E., Barber, R.D., Mitrophanous, K.A., Carter, E.E., Rohll, J.B., . . . Mazarakis, N.D., 2002. Multicistronic lentiviral vector-mediated striatal gene transfer of aromatic L-amino acid decarboxylase, tyrosine hydroxylase, and GTP cyclohydrolase I induces sustained transgene expression, dopamine production, and functional improvement in a rat model of Parkinson's disease. The Journal of neuroscience : the official journal of the Society for Neuroscience 22, 10302-10312.

Bachenheimer, S., Darnell, J.E., 1975. Adenovirus-2 mRNA is transcribed as part of a high-molecular-weight precursor RNA. Proceedings of the National Academy of Sciences 72, 4445-4449.

Bachmann, M., Moroy, T., 2005. The serine/threonine kinase Pim-1. The international journal of biochemistry & cell biology 37, 726-730.

Baldauf, H.M., Pan, X., Erikson, E., Schmidt, S., Daddacha, W., Burggraf, M., . . . Keppler, O.T., 2012. SAMHD1 restricts HIV-1 infection in resting CD4(+) T cells. Nature medicine 18, 1682-1687.

Baltimore, D., 1970. RNA-dependent DNA polymerase in virions of RNA tumour viruses. Nature 226, 1209-1211.

Barclay, J.L., Anderson, S.T., Waters, M.J., Curlewis, J.D., 2007. Regulation of

suppressor of cytokine signaling 3 (SOC3) by growth hormone in pro-B cells. Mol Endocrinol 21, 2503-2515.

Barre-Sinoussi, F., Chermann, J.C., Rey, F., Nugeyre, M.T., Chamaret, S., Gruest, J., . . . Montagnier, L., 1983. Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS). Science 220, 868-871.

Bas, A., Forsberg, G., Hammarstrom, S., Hammarstrom, M.L., 2004. Utility of the housekeeping genes 18S rRNA, beta-actin and glyceraldehyde-3-phosphate-dehydrogenase for normalization in real-time quantitative reverse transcriptase-polymerase chain reaction analysis of gene expression in human T lymphocytes. Scand J Immunol 59, 566-573.

Baumgartel, V., Ivanchenko, S., Dupont, A., Sergeev, M., Wiseman, P.W., Krausslich, H.G., . . . Lamb, D.C., 2011. Live-cell visualization of dynamics of HIV budding site interactions with an ESCRT component. Nature cell biology 13, 469-474.

Bazzoni, G., Carlesso, N., Griffin, J.D., Hemler, M.E., 1996. Bcr/Abl expression stimulates integrin function in hematopoietic cell lines. The Journal of clinical investigation 98, 521-528.

Bedard, K., Krause, K.H., 2007. The NOX family of ROS-generating NADPH oxidases: physiology and pathophysiology. Physiol Rev 87, 245-313.

Berget, S.M., 1995. Exon recognition in vertebrate splicing. The Journal of biological chemistry 270, 2411-2414.

Berget, S.M., Moore, C., Sharp, P.A., 1977. Spliced segments at the 5' terminus of adenovirus 2 late mRNA. Proc Natl Acad Sci U S A 74, 3171-3175.

Berk, A.J., 2016. Discovery of RNA splicing and genes in pieces. Proc Natl Acad Sci U S A 113, 801-805.

Berry, C., Hannenhalli, S., Leipzig, J., Bushman, F.D., 2006. Selection of target sites for mobile DNA integration in the human genome. PLoS computational biology 2, e157.

Biffi, A., Aubourg, P., Cartier, N., 2011. Gene therapy for leukodystrophies. Human molecular genetics 20, R42-53.

Biffi, A., Capotondo, A., Fasano, S., del Carro, U., Marchesini, S., Azuma, H., . . . Naldini, L., 2006. Gene therapy of metachromatic leukodystrophy reverses neurological damage and deficits in mice. The Journal of clinical investigation 116, 3070-3082.

Biffi, A., De Palma, M., Quattrini, A., Del Carro, U., Amadio, S., Visigalli, I., . . . Naldini, L., 2004. Correction of metachromatic leukodystrophy in the mouse model by transplantation of genetically modified hematopoietic stem cells. The Journal of clinical investigation 113, 1118-1129.

Biffi, A., Montini, E., Lorioli, L., Cesani, M., Fumagalli, F., Plati, T., . . . Naldini, L., 2013. Lentiviral hematopoietic stem cell gene therapy benefits metachromatic leukodystrophy. Science 341, 1233158.

Bittner, J.J., 1942. The Milk-Influence of Breast Tumors in Mice. Science 95, 462-463.

Blaese, R.M., Culver, K.W., Miller, A.D., Carter, C.S., Fleisher, T., Clerici, M., . . . Anderson, W.F., 1995. T lymphocyte-directed gene therapy for ADA- SCID: initial trial results after 4 years. Science 270, 475-480.

Bokhoven, M., Stephen, S.L., Knight, S., Gevers, E.F., Robinson, I.C., Takeuchi, Y., Collins, M.K., 2009. Insertional gene activation by lentiviral and gammaretroviral vectors. Journal of virology 83, 283-294.

Booth, D.S., Cheng, Y., Frankel, A.D., 2014. The export receptor Crm1 forms a dimer to promote nuclear export of HIV RNA. eLife 3, e04121.

Boztug, K., Schmidt, M., Schwarzer, A., Banerjee, P.P., Diez, I.A., Dewey, R.A., . . . Klein, C., 2010. Stem-cell gene therapy for the Wiskott-Aldrich syndrome. The New England journal of medicine 363, 1918-1927.

Braun, C.J., Boztug, K., Paruzynski, A., Witzel, M., Schwarzer, A., Rothe, M., . . . Klein, C., 2014. Gene therapy for Wiskott-Aldrich syndrome--long-term efficacy and genotoxicity. Science translational medicine 6, 227ra233.

Brow, D.A., 2002. Allosteric cascade of spliceosome activation. Annual review of genetics 36, 333-360.

Brown, B.D., Venneri, M.A., Zingale, A., Sergi Sergi, L., Naldini, L., 2006. Endogenous microRNA regulation suppresses transgene expression in hematopoietic lineages and enables stable gene transfer. Nature medicine 12, 585-591.

Brunak, S., Engelbrecht, J., Knudsen, S., 1991. Prediction of human mRNA donor and acceptor sites from the DNA sequence. Journal of molecular biology 220, 49-65.

Bukrinsky, M.I., Haggerty, S., Dempsey, M.P., Sharova, N., Adzhubel, A., Spitz, L., . . . Stevenson, M., 1993. A nuclear localization signal within HIV-1 matrix protein that governs infection of non-dividing cells. Nature 365, 666-669.

Burke, B.P., Levin, B.R., Zhang, J., Sahakyan, A., Boyer, J., Carroll, M.V., . . . Symonds, G.P., 2015. Engineering Cellular Resistance to HIV-1 Infection In Vivo Using a Dual Therapeutic Lentiviral Vector. Molecular therapy. Nucleic acids 4, e236.

Burns, J.C., Friedmann, T., Driever, W., Burrascano, M., Yee, J.K., 1993. Vesicular stomatitis virus G glycoprotein pseudotyped retroviral vectors: concentration to very high titer and efficient gene transfer into mammalian and nonmammalian cells. Proc Natl Acad Sci U S A 90, 8033-8037.

Busslinger, M., Moschonas, N., Flavell, R.A., 1981. Beta + thalassemia: aberrant splicing results from a single point mutation in an intron. Cell 27, 289-298.

Caldenhoven, E., van Dijk, T.B., Solari, R., Armstrong, J., Raaijmakers, J.A.M., Lammers, J.-W.J., . . . de Groot, R.P., 1996. STAT3β, a Splice Variant of Transcription Factor STAT3, Is a Dominant Negative Regulator of Transcription. Journal of Biological Chemistry 271, 13221-13227.

Campbell, E.M., Hope, T.J., 2015. HIV-1 capsid: the multifaceted key player in HIV-1 infection. Nature reviews. Microbiology 13, 471-483.

Carteau, S., Hoffmann, C., Bushman, F., Chromosome Structure and Human Immunodeficiency Virus Type 1 cDNA Integration: Centromeric Alphoid Repeats Are a Disfavored Target. J Virol. 1998 May;72(5):4005-14.

Cartier, N., Hacein-Bey-Abina, S., Bartholomae, C.C., Veres, G., Schmidt, M., Kutschera, I., . . . Aubourg, P., 2009. Hematopoietic stem cell gene therapy with a lentiviral vector in X-linked adrenoleukodystrophy. Science 326, 818-823.

Cattoglio, C., Facchini, G., Sartori, D., Antonelli, A., Miccio, A., Cassani, B., . . . Mavilio, F., 2007. Hot spots of retroviral integration in human CD34+ hematopoietic cells. Blood 110, 1770-1778.

Catucci, M., Castiello, M.C., Pala, F., Bosticardo, M., Villa, A., 2012. Autoimmunity in wiskott-Aldrich syndrome: an unsolved enigma. Front Immunol 3.

Cavalli, G., Misteli, T., 2013. Functional implications of genome topology. Nature structural & molecular biology 20, 290-299.

Cavazza, A., Moiani, A., Mavilio, F., 2013. Mechanisms of retroviral integration and mutagenesis. Human gene therapy 24, 119-131.

Cavazzana-Calvo, M., Payen, E., Negre, O., Wang, G., Hehir, K., Fusil, F., . . . Leboulch, P., 2010. Transfusion independence and HMGA2 activation after gene therapy of human beta-thalassaemia. Nature 467, 318-322.

Cesana, D., Sgualdino, J., Rudilosso, L., Merella, S., Naldini, L., Montini, E., 2012. Whole transcriptome characterization of aberrant splicing events induced by lentiviral vector integrations. The Journal of clinical investigation 122, 1667-1676.

Challita, P.M., Skelton, D., el-Khoueiry, A., Yu, X.J., Weinberg, K., Kohn, D.B., 1995. Multiple modifications in cis elements of the long terminal repeat of retroviral vectors lead to increased expression and decreased DNA methylation in embryonic carcinoma cells. Journal of virology 69, 748-755.

Chan, B., Wara, D., Bastian, J., Hershfield, M.S., Bohnsack, J., Azen, C.G., . . . Kohn, D.B., 2005. Long-term efficacy of enzyme replacement therapy for adenosine deaminase (ADA)-deficient severe combined immunodeficiency (SCID). Clin Immunol 117, 133-143.

Chan, D.C., Fass, D., Berger, J.M., Kim, P.S., 1997. Core structure of gp41 from the HIV envelope glycoprotein. Cell 89, 263-273.

Chang, Y.F., Imam, J.S., Wilkinson, M.F., 2007. The nonsense-mediated decay RNA surveillance pathway. Annual review of biochemistry 76, 51-74.

Charrier, S., Dupre, L., Scaramuzza, S., Jeanson-Leh, L., Blundell, M.P., Danos, O., . . . Galy, A., 2007. Lentiviral vectors targeting WASp expression to hematopoietic cells, efficiently transduce and correct cells from WAS patients. Gene therapy 14, 415-428.

Checkley, M.A., Luttge, B.G., Freed, E.O., 2011. HIV-1 envelope glycoprotein biosynthesis, trafficking, and incorporation. Journal of molecular biology 410, 582-608.

Cheng, Q., Chang, J.T., Geradts, J., Neckers, L.M., Haystead, T., Spector, N.L., Lyerly, H.K., 2012. Amplification and high-level expression of heat shock protein 90 marks aggressive phenotypes of human epidermal growth factor receptor 2 negative breast cancer. Breast cancer research : BCR 14, R62.

Chiriaco, M., Farinelli, G., Capo, V., Zonari, E., Scaramuzza, S., Di Matteo, G., . . . Aiuti, A., 2014. Dual-regulated lentiviral vector for gene therapy of X-linked chronic granulomatosis. Mol Ther 22, 1472-1483.

Choe, H., Farzan, M., Sun, Y., Sullivan, N., Rollins, B., Ponath, P.D., . . . Sodroski, J., 1996. The beta-chemokine receptors CCR3 and CCR5 facilitate infection by primary HIV-1 isolates. Cell 85, 1135-1148.

Chow, L.T., Broker, T.R., 1978. The spliced structures of adenovirus 2 fiber message and the other late mRNAs. Cell 15, 497-510.

Chung, J.H., Bell, A.C., Felsenfeld, G., 1997. Characterization of the

chicken β-globin insulator. Proceedings of the National Academy of Sciences 94, 575-580.

Churbanov, A., Rogozin, I.B., Deogun, J.S., Ali, H., 2006. Method of predicting splice sites based on signal interactions. Biol Direct 1, 10.

Church, D.M., Goodstadt, L., Hillier, L.W., Zody, M.C., Goldstein, S., She, X., . . . Ponting, C.P., 2009. Lineage-specific biology revealed by a finished genome assembly of the mouse. PLoS biology 7, 26.

Clancy, S., 2008. RNA Splicing: Introns, Exons and Spliceosome. Nature Education 1, 31.

Clavel, F., Hance, A.J., 2004. HIV drug resistance. The New England journal of medicine 350, 1023-1035.

Collins, M.K., Downward, J., Miyajima, A., Maruyama, K., Arai, K., Mulligan, R.C., 1988. Transfer of functional EGF receptors to an IL3-dependent cell line. Journal of cellular physiology 137, 293-298.

Collins, M.K., Malde, P., Miyajima, A., Arai, K., Smith, K.A., Mulligan, R.C., 1990. Evidence that the level of the p55 component of the interleukin (IL) 2 receptor can control IL 2 responsiveness in a murine IL 3-dependent cell. European journal of immunology 20, 573-578.

Collins, M.K., Marvel, J., Malde, P., Lopez-Rivas, A., 1992. Interleukin 3 protects murine bone marrow cells from apoptosis induced by DNA damaging agents. The Journal of experimental medicine 176, 1043-1051.

Consortium, T.U., 2010. The Universal Protein Resource (UniProt) in 2010. Nucleic acids research 38, D142-148.

Cooper, T.A., Wan, L., Dreyfuss, G., 2009. RNA and disease. Cell 136, 777-793.

Coune, P.G., Schneider, B.L., Aebischer, P., Parkinson's Disease: Gene Therapies. Cold Spring Harb Perspect Med. 2012 Apr;2(4):a009431. doi:10.1101/cshperspect.a009431.

Craigie, R., Bushman, F.D., 2012. HIV DNA integration. Cold Spring Harbor perspectives in medicine 2, a006890.

Craigie, R., Bushman, F.D., 2014. Host factors in retroviral integration and the selection of integration target sites. Microbiology spectrum 2, 10.1128/microbiolspec.MDNA1123-0026-2014.

Cullen, B.R., 1992. Mechanism of action of regulatory proteins encoded by complex retroviruses. Microbiological reviews 56, 375-394.

Cullen, B.R., 2003. Nuclear mRNA export: insights from virology. Trends in biochemical sciences 28, 419-424.

Danos, O., Mulligan, R.C., 1988. Safe and efficient generation of recombinant retroviruses with amphotropic and ecotropic host ranges. Proc Natl Acad Sci U S A 85, 6460-6464.

Darnell, J.E., Jr., 2013. Reflections on the history of pre-mRNA processing and highlights of current knowledge: a unified picture. Rna 19, 443-460.

Darnell, J.E., Wall, R., Tushinski, R.J., 1971. An Adenylic Acid-Rich Sequence in Messenger RNA of HeLa Cells and Its Possible Relationship to Reiterated Sites in DNA. Proceedings of the National Academy of Sciences 68, 1321-1325.

Daugherty, M.D., Booth, D.S., Jayaraman, B., Cheng, Y., Frankel, A.D., 2010a. HIV Rev response element (RRE) directs assembly of the Rev homooligomer into discrete asymmetric complexes. Proc Natl Acad Sci U S A 107, 12481-12486.

Daugherty, M.D., Liu, B., Frankel, A.D., 2010b. Structural basis for cooperative RNA binding and export complex assembly by HIV Rev. Nature structural & molecular biology 17, 1337-1342.

Davis, S., Aldrich, T.H., Jones, P.F., Acheson, A., Compton, D.L., Jain, V., . . . Yancopoulos, G.D., 1996. Isolation of angiopoietin-1, a ligand for the TIE2 receptor, by secretion-trap expression cloning. Cell 87, 1161-1169.

de Lau, L.M., Breteler, M.M., 2006. Epidemiology of Parkinson's disease. The Lancet. Neurology 5, 525-535.

De Rijck, J., de Kogel, C., Demeulemeester, J., Vets, S., El Ashkar, S., Malani, N., . . . Debyser, Z., 2013. The BET family of proteins targets moloney murine leukemia virus integration near transcription start sites. Cell Rep 5, 886-894.

Debyser, Z., Christ, F., De Rijck, J., Gijsbers, R., 2015. Host factors for retroviral integration site selection. Trends in biochemical sciences 40, 108-116.

Demaison, C., Parsley, K., Brouns, G., Scherr, M., Battmer, K., Kinnon, C., . . . Thrasher, A.J., 2002. High-level transduction and gene expression in hematopoietic repopulating cells using a human immunodeficiency [correction of imunodeficiency] virus type 1-based lentiviral vector containing an internal spleen focus forming virus promoter. Human gene therapy 13, 803-813.

Demiroglu, A., Steer, E.J., Heath, C., Taylor, K., Bentley, M., Allen, S.L., . . . Cross, N.C., 2001. The t(8;22) in chronic myeloid leukemia fuses BCR to FGFR1: transforming activity and specific inhibition of FGFR1 fusion proteins. Blood 98, 3778-3783.

Deng, J., Kawakami, Y., Hartman, S.E., Satoh, T., Kawakami, T., 1998. Involvement of Ras in Bruton's tyrosine kinase-mediated JNK activation. The Journal of biological chemistry 273, 16787-16791.

Dominski, Z., Kole, R., 1993. Restoration of correct splicing in thalassemic pre-mRNA by antisense oligonucleotides. Proc Natl Acad Sci U S A 90, 8673-8677.

Dong, F., van Buitenen, C., Pouwels, K., Hoefsloot, L.H., Lowenberg, B., Touw, I.P., 1993. Distinct cytoplasmic regions of the human granulocyte colony-stimulating factor receptor involved in induction of proliferation and maturation. Molecular and cellular biology 13, 7774-7781.

Drachman, J.G., Griffin, J.D., Kaushansky, K., 1995. The c-Mpl ligand (thrombopoietin) stimulates tyrosine phosphorylation of Jak2, Shc, and c-Mpl. The Journal of biological chemistry 270, 4979-4982.

Dvorin, J.D., Bell, P., Maul, G.G., Yamashita, M., Emerman, M., Malim, M.H., 2002. Reassessment of the roles of integrase and the central DNA flap in human immunodeficiency virus type 1 nuclear import. Journal of virology 76, 12087-12096.

Early, P., Rogers, J., Davis, M., Calame, K., Bond, M., Wall, R., Hood, L., 1980. Two mRNAs can be produced from a single immunoglobulin mu gene by alternative RNA processing pathways. Cell 20, 313-319.

Edmonds, M., Abrams, R., 1960. Polynucleotide Biosynthesis: Formation of a Sequence of Adenylate Units from Adenosine Triphosphate by an Enzyme from Thymus Nuclei. Journal of Biological Chemistry 235, 1142-1149.

Ellis, J., Pannell, D., 2001. The beta-globin locus control region versus gene therapy vectors: a struggle for expression. Clin Genet 59, 17-24.

Emery, D.W., Chen, H., Li, Q., Stamatoyannopoulos, G., 1998. Development of a Condensed Locus Control Region Cassette and Testing in Retrovirus Vectors forAγ-Globin. Blood Cells, Molecules, and Diseases 24, 322-339.

Engelman, A., Mizuuchi, K., Craigie, R., 1991. HIV-1 DNA integration: mechanism of viral DNA cleavage and DNA strand transfer. Cell 67, 1211-1221.

Erickson, L.C., 1991. The role of O-6 methylguanine DNA methyltransferase (MGMT) in drug resistance and strategies for its inhibition. Semin Cancer Biol 2, 257-265.

Escors, D., Lopes, L., Lin, R., Hiscott, J., Akira, S., Davis, R.J., Collins, M.K., 2008. Targeting dendritic cell signaling to regulate the response to immunization. Blood 111, 3050-3061.

Ewing, B., Green, P., 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome research 8, 186-194.

Faschinger, A., Rouault, F., Sollner, J., Lukas, A., Salmons, B., Gunzburg, W.H., Indik, S., 2008. Mouse mammary tumor virus integration site selection in human and mouse genomes. Journal of virology 82, 1360-1367.

Fassati, A., Görlich, D., Harrison, I., Zaytseva, L., Mingot, J.-M., 2003. Nuclear import of HIV-1 intracellular reverse transcription complexes is mediated by importin 7. The EMBO journal 22, 3675-3685.

Feng, Y., Broder, C.C., Kennedy, P.E., Berger, E.A., 1996. HIV-1 entry cofactor: functional cDNA cloning of a seven-transmembrane, G protein-coupled receptor. Science 272, 872-877.

Finkelshtein, D., Werman, A., Novick, D., Barak, S., Rubinstein, M., 2013. LDL receptor and its family members serve as the cellular receptors for vesicular stomatitis virus. Proc Natl Acad Sci U S A 110, 7306-7311.

Finotti, A., Breda, L., Lederer, C.W., Bianchi, N., Zuccato, C., Kleanthous, M., . . . Gambari, R., 2015. Recent trends in the gene therapy of beta-thalassemia. Journal of blood medicine 6, 69-85.

Fluharty, A.L., 2014. Arylsulfatase A deficiency.

Forshey, B.M., von Schwedler, U., Sundquist, W.I., Aiken, C., 2002. Formation of a human immunodeficiency virus type 1 core of optimal stability is crucial for viral replication. Journal of virology 76, 5667-5677.

Francis, S.M., Larsen, J.E., Pavey, S.J., Bowman, R.V., Hayward, N.K., Fong, K.M., Yang, I.A., 2009. Expression profiling identifies genes involved in emphysema severity. Respiratory research 10, 1465-9921.

Frederiks, W.M., James, J., Arnouts, C., Broekhoven, S., Morreau, J., 1978. The influence of Triton X-100 on the nuclear envelope of the isolated liver cell nuclei. Cytobiologie 18, 254-271.

Freed, E.O., 2015. HIV-1 assembly, release and maturation. Nature reviews. Microbiology 13, 484-496.

Freed, E.O., Englund, G., Martin, M.A., 1995. Role of the basic domain of human immunodeficiency virus type 1 matrix in macrophage infection. Journal of virology 69, 3949-3954.

Frohman, M.A., Dush, M.K., Martin, G.R., 1988. Rapid production of full-length cDNAs from rare transcripts: amplification using a single gene-specific oligonucleotide primer. Proc Natl Acad Sci U S A 85, 8998-9002.

Fuentes, G.M., Rodriguez-Rodriguez, L., Palaniappan, C., Fay, P.J., Bambara, R.A., 1996. Strand displacement synthesis of the long terminal repeats by HIV reverse transcriptase. The Journal of biological chemistry 271, 1966-1971.

Fujiwara, T., Mizuuchi, K., 1988. Retroviral DNA integration: structure of an integration intermediate. Cell 54, 497-504.

Furger, A., Monks, J., Proudfoot, N.J., 2001. The retroviruses human immunodeficiency virus type 1 and Moloney murine leukemia virus adopt radically different strategies to regulate promoter-proximal polyadenylation. Journal of virology 75, 11735-11746.

Gallay, P., Hope, T., Chin, D., Trono, D., 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, 9825-9830.

Gallo, R.C., 2005. History of the discoveries of the first human retroviruses: HTLV-1 and HTLV-2. Oncogene 24, 5926-5930.

Galy, A., Thrasher, A.J., 2011. Gene therapy for the Wiskott-Aldrich syndrome. Curr Opin Allergy Clin Immunol 11, 545-550.

Gaspar, H.B., Cooray, S., Gilmour, K.C., Parsley, K.L., Adams, S., Howe, S.J., . . . Thrasher, A.J., 2011. Long-term persistence of a polyclonal T cell repertoire after gene therapy for X-linked severe combined immunodeficiency. Science translational medicine 3, 3002715.

Gatignol, A., 2007. Transcription of HIV: Tat and cellular chromatin. Adv Pharmacol 55, 137-159.

Gennery, A.R., Slatter, M.A., Grandin, L., Taupin, P., Cant, A.J., Veys, P., . . . Landais, P., 2010. Transplantation of hematopoietic stem cells and long-term survival for primary immunodeficiencies in Europe: entering a new century, do we do better? The Journal of allergy and clinical immunology 126, 602-610 e601-611.

Ghigna, C., Giordano, S., Shen, H., Benvenuto, F., Castiglioni, F., Comoglio, P.M., . . . Biamonti, G., 2005. Cell motility is controlled by SF2/ASF through alternative splicing of the Ron protooncogene. Molecular cell 20, 881-890.

Ghilardi, N., Skoda, R.C., 1997. The leptin receptor activates janus kinase 2 and signals for proliferation in a factor-dependent cell line. Mol Endocrinol 11, 393-399.

Giacca, M., Zacchigna, S., 2012. Virus-mediated gene delivery for human gene therapy. Journal of controlled release : official journal of the Controlled Release Society 161, 377-388.

Gilbert, W., 1987. The exon theory of genes. Cold Spring Harb Symp Quant Biol 52, 901-905.

Goel, R., Beard, W.A., Kumar, A., Casas-Finet, J.R., Strub, M.P., Stahl, S.J., . . . et al., 1993. Structure/function studies of HIV-1(1) reverse transcriptase: dimerization-defective mutant L289K. Biochemistry 32, 13012-13018.

Göttlinger, H.G., Dorfman, T., Sodroski, J.G., Haseltine, W.A., 1991. Effect of mutations affecting the p6 gag protein on human immunodeficiency virus particle release. Proc Natl Acad Sci U S A 88, 3195-3199.

Goujon, C., Moncorge, O., Bauby, H., Doyle, T., Ward, C.C., Schaller, T., . . . Malim, M.H., 2013. Human MX2 is an interferon-induced post-entry inhibitor of HIV-1 infection. Nature 502, 559-562.

Grez, M., Reichenbach, J., Schwable, J., Seger, R., Dinauer, M.C., Thrasher, A.J., 2011. Gene therapy of chronic granulomatous disease: the engraftment dilemma. Mol Ther 19, 28-35.

Grosveld, F., van Assendelft, G.B., Greaves, D.R., Kollias, G., 1987. Position-independent, high-level expression of the human β-globin gene in transgenic mice. Cell 51, 975-985.

Guo, F., Cen, S., Niu, M., Saadatmand, J., Kleiman, L., 2006. Inhibition of tRNA(3)(Lys)-primed reverse transcription by human APOBEC3G during human immunodeficiency virus type 1 replication. Journal of virology 80, 11710-11722.

Gupta, S., Purcell, N.H., Lin, A., Sen, S., 2002. Activation of nuclear factor-kappaB is necessary for myotrophin-induced cardiac hypertrophy. The Journal of cell biology 159, 1019-1028.

Hacein-Bey Abina, S., Gaspar, H.B., Blondeau, J., Caccavelli, L., Charrier, S., Buckland, K., . . . Cavazzana, M., 2015. Outcomes following gene therapy in patients with severe Wiskott-Aldrich syndrome. Jama 313, 1550-1563.

Hacein-Bey-Abina, S., Garrigue, A., Wang, G.P., Soulier, J., Lim, A., Morillon, E., . . . Cavazzana-Calvo, M., 2008. Insertional oncogenesis in 4 patients after retrovirus-mediated gene therapy of SCID-X1. The Journal of clinical investigation 118, 3132-3142.

Hacein-Bey-Abina, S., Hauer, J., Lim, A., Picard, C., Wang, G.P., Berry, C.C., . . . Cavazzana-Calvo, M., 2010. Efficacy of gene therapy for X-linked severe combined immunodeficiency. The New England journal of medicine 363, 355-364.

Haemmerle, R., Phaltane, R., Rothe, M., Schroder, S., Schambach, A., Moritz, T., Modlich, U., 2014. Clonal Dominance With Retroviral Vector Insertions Near

the ANGPT1 and ANGPT2 Genes in a Human Xenotransplant Mouse Model. Molecular therapy. Nucleic acids 3, e200.

Halene, S., Wang, L., Cooper, R.M., Bockstoce, D.C., Robbins, P.B., Kohn, D.B., 1999. Improved expression in hematopoietic and lymphoid cells in mice after transplantation of bone marrow transduced with a modified retroviral vector. Blood 94, 3349-3357.

Hall, S.L., Padgett, R.A., 1994. Conserved sequences in a class of rare eukaryotic nuclear introns with non-consensus splice sites. Journal of molecular biology 239, 357-365.

Hanson, P.I., Roth, R., Lin, Y., Heuser, J.E., 2008. Plasma membrane deformation by circular arrays of ESCRT-III protein filaments. The Journal of cell biology 180, 389-402.

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

Harkey, M.A., Kaul, R., Jacobs, M.A., Kurre, P., Bovee, D., Levy, R., Blau, C.A., 2007. Multiarm high-throughput integration site detection: limitations of LAM-PCR technology and optimization for clonal analysis. Stem Cells Dev 16, 381-392.

Harris, H., 1959. Turnover of nuclear and cytoplasmic ribonucleic acid in two types of animal cell, with some further observations on the nucleolus. Biochemical Journal 73, 362-369.

Hayashi, T., Shioda, T., Iwakura, Y., Shibuta, H., 1992. RNA packaging signal of human immunodeficiency virus type 1. Virology 188, 590-599.

Hebsgaard, S.M., Korning, P.G., Tolstrup, N., Engelbrecht, J., Rouze, P., Brunak, S., 1996. Splice site prediction in Arabidopsis thaliana pre-mRNA by combining local and global sequence information. Nucleic acids research 24, 3439-3452.

Heckl, D., Schwarzer, A., Haemmerle, R., Steinemann, D., Rudolph, C., Skawran, B., . . . Modlich, U., 2012. Lentiviral vector induced insertional haploinsufficiency of Ebf1 causes murine leukemia. Mol Ther 20, 1187-1195.

Heinzinger, N.K., Bukrinsky, M.I., Haggerty, S.A., Ragland, A.M., Kewalramani, V., Lee, M.A., . . . Emerman, M., 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, 7311-7315.

Hernandez-Lopez, H.R., Graham, S.V., 2012. Alternative splicing in human tumour viruses: a therapeutic target? The Biochemical journal 445, 145-156.

Herrera-Carrillo, E., Berkhout, B., 2015. Bone Marrow Gene Therapy for HIV/AIDS. Viruses 7, 3910-3936.

Hock, R.A., Miller, A.D., 1986. Retrovirus-mediated transfer and expression of drug resistance genes in human haematopoietic progenitor cells. Nature 320, 275-277.

Hoffmann, M., Wu, Y.J., Gerber, M., Berger-Rentsch, M., Heimrich, B., Schwemmle, M., Zimmer, G., 2010. Fusion-active glycoprotein G mediates the cytotoxicity of vesicular stomatitis virus M mutants lacking host shut-off activity. The Journal of general virology 91, 2782-2793.

Hossini, A.M., Eberle, J., Fecker, L.F., Orfanos, C.E., Geilen, C.C., 2003. Conditional expression of exogenous Bcl-X(S) triggers apoptosis in human melanoma cells in vitro and delays growth of melanoma xenografts. FEBS letters 553, 250-256.

Howe, S.J., Mansour, M.R., Schwarzwaelder, K., Bartholomae, C., Hubank, M., Kempski, H., . . . Thrasher, A.J., 2008. Insertional mutagenesis combined with acquired somatic mutations causes leukemogenesis following gene therapy of SCID-X1 patients. The Journal of clinical investigation 118, 3143-3150.

Hu, W.S., Hughes, S.H., 2012. HIV-1 reverse transcription. Cold Spring Harbor perspectives in medicine 2.

Huang, H., Bhat, A., Woodnutt, G., Lappe, R., 2010. Targeting the ANGPT-TIE2 pathway in malignancy. Nature reviews. Cancer 10, 575-585.

Hulme, A.E., Perez, O., Hope, T.J., 2011. Complementary assays reveal a relationship between HIV-1 uncoating and reverse transcription. Proc Natl Acad Sci U S A 108, 9975-9980.

Hunter, E., Swanstrom, R., 1990. Retrovirus envelope glycoproteins. Current topics in microbiology and immunology 157, 187-253.

Hutter, G., Nowak, D., Mossner, M., Ganepola, S., Mussig, A., Allers, K., . . . Thiel, E., 2009. Long-term control of HIV by CCR5 Delta32/Delta32 stem-cell transplantation. The New England journal of medicine 360, 692-698.

Hutter, G., Thiel, E., 2011. Allogeneic transplantation of CCR5-deficient progenitor cells in a patient with HIV infection: an update after 3 years and the search for patient no. 2. AIDS 25, 273-274.

Ibrahim, E.C., Schaal, T.D., Hertel, K.J., Reed, R., Maniatis, T., 2005. Serine/arginine-rich protein-dependent suppression of exon skipping by exonic splicing enhancers. Proc Natl Acad Sci U S A 102, 5002-5007.

Iordanskiy, S., Berro, R., Altieri, M., Kashanchi, F., Bukrinsky, M., 2006. Intracytoplasmic maturation of the human immunodeficiency virus type 1 reverse transcription complexes determines their capacity to integrate into chromatin. Retrovirology 3, 4.

Jackson, I.J., 1991. A reappraisal of non-consensus mRNA splice sites. Nucleic acids research 19, 3795-3798.

Jacquenet, S., Mereau, A., Bilodeau, P.S., Damier, L., Stoltzfus, C.M., Branlant, C., 2001. A second exon splicing silencer within human immunodeficiency virus type 1 tat exon 2 represses splicing of Tat mRNA and binds protein hnRNP H. The Journal of biological chemistry 276, 40464-40475.

Jacques, D.A., McEwan, W.A., Hilditch, L., Price, A.J., Towers, G.J., James, L.C., 2016. HIV-1 uses dynamic capsid pores to import nucleotides and fuel encapsidated DNA synthesis. Nature 536, 349-353.

Janvier, K., Pelchen-Matthews, A., Renaud, J.B., Caillet, M., Marsh, M., Berlioz-Torrent, C., 2011. The ESCRT-0 component HRS is required for HIV-1 Vpu-mediated BST-2/tetherin down-regulation. PLoS pathogens 7, 1001265.

Jeanmougin, F., Thompson, J.D., Gouy, M., Higgins, D.G., Gibson, T.J., 1998. Multiple sequence alignment with Clustal X. Trends in biochemical sciences 23, 403-405.

Johnson, J.M., Castle, J., Garrett-Engele, P., Kan, Z., Loerch, P.M., Armour, C.D., . . . Shoemaker, D.D., 2003. Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. Science 302, 2141-2144.

Jostock, T., Mullberg, J., Ozbek, S., Atreya, R., Blinn, G., Voltz, N., . . . Rose-John, S., 2001. Soluble gp130 is the natural inhibitor of soluble interleukin-6 receptor transsignaling responses. European journal of biochemistry 268, 160-167.

Kafri, T., van Praag, H., Ouyang, L., Gage, F.H., Verma, I.M., 1999. A packaging cell line for lentivirus vectors. Journal of virology 73, 576-584.

Kai, N., Mishina, M., Yagi, T., 1997. Molecular cloning of Fyn-associated molecules in the mouse central nervous system. J Neurosci Res 48, 407-424.

Kajaste-Rudnitski, A., Naldini, L., 2015. Cellular innate immunity and restriction of viral infection: implications for lentiviral gene therapy in human hematopoietic cells. Human gene therapy 26, 201-209.

Kane, M., Yadav, S.S., Bitzegeio, J., Kutluay, S.B., Zang, T., Wilson, S.J., . . . Bieniasz, P.D., 2013. MX2 is an interferon-induced inhibitor of HIV-1 infection. Nature 502, 563-566.

Kang, E., Gennery, A., Hematopoietic Stem Cell Transplantation for Primary Immunodeficiencies By Elizabeth Kang and Andrew Gennery. Hematol Oncol Clin North Am. 2014 Dec;28(6):1157-70. Epub 2014 Sep 16 doi:10.1016/j.hoc.2014.08.006.

Kang, E.M., Choi, U., Theobald, N., Linton, G., Long Priel, D.A., Kuhns, D., Malech, H.L., 2010. Retrovirus gene therapy for X-linked chronic granulomatous disease can achieve stable long-term correction of oxidase activity in peripheral blood neutrophils. Blood 115, 783-791.

Kaplan, A.H., Zack, J.A., Knigge, M., Paul, D.A., Kempf, D.J., Norbeck, D.W., Swanstrom, R., 1993. Partial inhibition of the human immunodeficiency virus type 1 protease results in aberrant virus assembly and the formation of noninfectious particles. Journal of virology 67, 4050-4055.

Karn, J., Stoltzfus, C.M., 2012. Transcriptional and posttranscriptional regulation of HIV-1 gene expression. Cold Spring Harbor perspectives in medicine 2, a006916.

Kasprzycka, M., Marzec, M., Liu, X., Zhang, Q., Wasik, M.A., 2006. Nucleophosmin/anaplastic lymphoma kinase (NPM/ALK) oncoprotein induces the T regulatory cell phenotype by activating STAT3. Proc Natl Acad Sci U S A 103, 9964-9969.

Kataoka, N., Bachorik, J.L., Dreyfuss, G., 1999. Transportin-SR, a nuclear import receptor for SR proteins. The Journal of cell biology 145, 1145-1152.

Kato, K., 1990. Sequence of a novel carbonic anhydrase-related polypeptide and its exclusive presence in Purkinje cells. FEBS letters 271, 137-140.

Kaufmann, K.B., Buning, H., Galy, A., Schambach, A., Grez, M., 2013. Gene therapy on the move. EMBO molecular medicine 5, 1642-1661.

Kearney, M., Palmer, S., Maldarelli, F., Shao, W., Polis, M.A., Mican, J., . . . Mellors, J.W., 2008. Frequent polymorphism at drug resistance sites in HIV-1 protease and reverse transcriptase. AIDS 22, 497-501.

Keele, B.F., Giorgi, E.E., Salazar-Gonzalez, J.F., Decker, J.M., Pham, K.T., Salazar, M.G., . . . Shaw, G.M., 2008. Identification and characterization of transmitted and early founder virus envelopes in primary HIV-1 infection. Proc Natl Acad Sci U S A 105, 7552-7557.

Keren, H., Lev-Maor, G., Ast, G., 2010. Alternative splicing and evolution: diversification, exon definition and function. Nature reviews. Genetics 11, 345-355.

Kim, E., Goren, A., Ast, G., 2008. Alternative splicing: current perspectives.

Bioessays 30, 38-47.

Kim, K.T., Baird, K., Ahn, J.Y., Meltzer, P., Lilly, M., Levis, M., Small, D., 2005. Pim-1 is up-regulated by constitutively activated FLT3 and plays a role in FLT3-mediated cell survival. Blood 105, 1759-1767.

Kim, Y.K., Bourgeois, C.F., Isel, C., Churcher, M.J., Karn, J., 2002. Phosphorylation of the RNA polymerase II carboxyl-terminal domain by CDK9 is directly responsible for human immunodeficiency virus type 1 Tat-activated transcriptional elongation. Molecular and cellular biology 22, 4622-4637.

Kinoshita, S., Chen, B.K., Kaneshima, H., Nolan, G.P., 1998. Host control of HIV-1 parasitism in T cells by the nuclear factor of activated T cells. Cell 95, 595-604.

Klein, C., Bueler, H., Mulligan, R.C., 2000. Comparative analysis of genetically modified dendritic cells and tumor cells as therapeutic cancer vaccines. The Journal of experimental medicine 191, 1699-1708.

Klein, C., Nguyen, D., Liu, C.H., Mizoguchi, A., Bhan, A.K., Miki, H., . . . Snapper, S.B., 2003. Gene therapy for Wiskott-Aldrich syndrome: rescue of T-cell signaling and amelioration of colitis upon transplantation of retrovirally transduced hematopoietic stem cells in mice. Blood 101, 2159-2166.

Knight, S., Bokhoven, M., Collins, M., Takeuchi, Y., 2010. Effect of the internal promoter on insertional gene activation by lentiviral vectors with an intact HIV long terminal repeat. Journal of virology 84, 4856-4859.

Knight, S., Collins, M., Takeuchi, Y., 2013. Insertional mutagenesis by retroviral vectors: current concepts and methods of analysis. Current gene therapy 13, 211-227.

Knight, S., Zhang, F., Mueller-Kuller, U., Bokhoven, M., Gupta, A., Broughton, T., . . . Takeuchi, Y., 2012. Safer, silencing-resistant lentiviral vectors: optimization of the ubiquitous chromatin-opening element through elimination of aberrant splicing. Journal of virology 86, 9088-9095.

Knudson, A.G., Jr., 1971. Mutation and cancer: statistical study of retinoblastoma. Proc Natl Acad Sci U S A 68, 820-823.

Koh, E.Y., Chen, T., Daley, G.Q., 2004. Genetic complementation of cytokine signaling identifies central role of kinases in hematopoietic cell proliferation. Oncogene 23, 1214-1220.

Koh, Y., Wu, X., Ferris, A.L., Matreyek, K.A., Smith, S.J., Lee, K., . . . Engelman, A., 2013. Differential effects of human immunodeficiency virus type 1 capsid and cellular factors nucleoporin 153 and LEDGF/p75 on the efficiency and specificity

of viral DNA integration. Journal of virology 87, 648-658.

Krishnan, L., Engelman, A., 2012. Retroviral Integrase Proteins and HIV-1 DNA Integration. Journal of Biological Chemistry 287, 40858-40866.

Kruttgen, A., Grotzinger, J., Kurapkat, G., Weis, J., Simon, R., Thier, M., . . . et al., 1995. Human ciliary neurotrophic factor: a structure-function analysis. The Biochemical journal 309 (Pt 1), 215-220.

Kuno, Y., Abe, A., Emi, N., Iida, M., Yokozawa, T., Towatari, M., . . . Saito, H., 2001. Constitutive kinase activation of the TEL-Syk fusion gene in myelodysplastic syndrome with t(9;12)(q22;p12). Blood 97, 1050-1055.

Kurahashi, H., Takami, K., Oue, T., Kusafuka, T., Okada, A., Tawa, A., . . . Nishisho, I., 1995. Biallelic inactivation of the APC gene in hepatoblastoma. Cancer research 55, 5007-5011.

Kutluay, S.B., Zang, T., Blanco-Melo, D., Powell, C., Jannain, D., Errando, M., Bieniasz, P.D., 2014. Global changes in the RNA binding specificity of HIV-1 gag regulate virion genesis. Cell 159, 1096-1109.

Kwong, P.D., Wyatt, R., Robinson, J., Sweet, R.W., Sodroski, J., Hendrickson, W.A., 1998. Structure of an HIV gp120 envelope glycoprotein in complex with the CD4 receptor and a neutralizing human antibody. Nature 393, 648-659.

Laguette, N., Sobhian, B., Casartelli, N., Ringeard, M., Chable-Bessia, C., Segeral, E., . . . Benkirane, M., 2011. SAMHD1 is the dendritic- and myeloid-cell-specific HIV-1 restriction factor counteracted by Vpx. Nature 474, 654-657.

Lahaye, X., Satoh, T., Gentili, M., Cerboni, S., Conrad, C., Hurbain, I., . . . Manel, N., 2013. The capsids of HIV-1 and HIV-2 determine immune detection of the viral cDNA by the innate sensor cGAS in dendritic cells. Immunity 39, 1132-1142.

Lalwani, A.K., Goldstein, J.A., Kelley, M.J., Luxford, W., Castelein, C.M., Mhatre, A.N., 2000. Human nonsyndromic hereditary deafness DFNA17 is due to a mutation in nonmuscle myosin MYH9. American journal of human genetics 67, 1121-1128.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., . . . Szustakowki, J., 2001. Initial sequencing and analysis of the human genome. Nature 409, 860-921.

Lane, R.F., St George-Hyslop, P., Hempstead, B.L., Small, S.A., Strittmatter, S.M., Gandy, S., 2012. Vps10 family proteins and the retromer complex in aging-related neurodegeneration and diabetes. The Journal of neuroscience :

the official journal of the Society for Neuroscience 32, 14080-14086.

Lannutti, B.J., Drachman, J.G., 2004. Lyn tyrosine kinase regulates thrombopoietin-induced proliferation of hematopoietic cell lines and primary megakaryocytic progenitors. Blood 103, 3736-3743.

Leboulch, P., 2013. Gene therapy: primed for take-off. Nature 500, 280-282.

Levine, B.L., Humeau, L.M., Boyer, J., MacGregor, R.R., Rebello, T., Lu, X., . . . June, C.H., 2006. Gene transfer in humans using a conditionally replicating lentiviral vector. Proc Natl Acad Sci U S A 103, 17372-17377.

Lewinski, M.K., Yamashita, M., Emerman, M., Ciuffi, A., Marshall, H., Crawford, G., . . . Bushman, F.D., 2006. Retroviral DNA integration: viral and cellular determinants of target-site selection. PLoS pathogens 2, e60.

Lewis, B.P., Green, R.E., Brenner, S.E., 2003. Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. Proc Natl Acad Sci U S A 100, 189-192.

LeWitt, P.A., Rezai, A.R., Leehey, M.A., Ojemann, S.G., Flaherty, A.W., Eskandar, E.N., . . . Feigin, A., 2011. AAV2-GAD gene therapy for advanced Parkinson's disease: a double-blind, sham-surgery controlled, randomised trial. The Lancet. Neurology 10, 309-319.

Lin, Y.C., Boone, M., Meuris, L., Lemmens, I., Van Roy, N., Soete, A., . . . Callewaert, N., 2014. Genome dynamics of the human embryonic kidney 293 lineage in response to cell biology manipulations. Nature communications 5, 4767.

Lindenboim, L., Yuan, J., Stein, R., 2000. Bcl-xS and Bax induce different apoptotic pathways in PC12 cells. Oncogene 19, 1783-1793.

Liu, J., Perkins, N.D., Schmid, R.M., Nabel, G.J., 1992. Specific NF-kappa B subunits act in concert with Tat to stimulate human immunodeficiency virus type 1 transcription. Journal of virology 66, 3883-3887.

Liu, J., Sorensen, A.B., Wang, B., Wabl, M., Nielsen, A.L., Pedersen, F.S., 2009. Identification of novel Bach2 transcripts and protein isoforms through tagging analysis of retroviral integrations in B-cell lymphomas. BMC molecular biology 10, 2.

Liu, Z., Pan, Q., Ding, S., Qian, J., Xu, F., Zhou, J., . . . Liang, C., 2013. The interferon-inducible MxB protein inhibits HIV-1 infection. Cell host & microbe 14, 398-410.

Llano, M., Saenz, D.T., Meehan, A., Wongthida, P., Peretz, M., Walker, W.H., . . .

Poeschla, E.M., 2006. An essential role for LEDGF/p75 in HIV integration. Science 314, 461-464.

Llano, M., Vanegas, M., Fregoso, O., Saenz, D., Chung, S., Peretz, M., Poeschla, E.M., 2004. LEDGF/p75 determines cellular trafficking of diverse lentiviral but not murine oncoretroviral integrase proteins and is a component of functional lentiviral preintegration complexes. Journal of virology 78, 9524-9537.

Logsdon, J.M., Jr., 1998. The recent origins of spliceosomal introns revisited. Curr Opin Genet Dev 8, 637-648.

Lorioli, L., Cesani, M., Regis, S., Morena, F., Grossi, S., Fumagalli, F., . . . Biffi, A., 2014. Critical issues for the proper diagnosis of Metachromatic Leukodystrophy. Gene 537, 348-351.

Louahed, J., Grasso, L., De Smet, C., Van Roost, E., Wildmann, C., Nicolaides, N.C., . . . Renauld, J.C., 1999. Interleukin-9-induced expression of M-Ras/R-Ras3 oncogene in T-helper clones. Blood 94, 1701-1710.

Ma, Z., Cools, J., Marynen, P., Cui, X., Siebert, R., Gesk, S., . . . Morris, S.W., 2000. Inv(2)(p23q35) in anaplastic large-cell lymphoma induces constitutive anaplastic lymphoma kinase (ALK) tyrosine kinase activation by fusion to ATIC, an enzyme involved in purine nucleotide biosynthesis. Blood 95, 2144-2149.

Ma, Z., Morris, S.W., Valentine, V., Li, M., Herbrick, J.A., Cui, X., . . . Hitzler, J.K., 2001. Fusion of two novel genes, RBM15 and MKL1, in the t(1;22)(p13;q13) of acute megakaryoblastic leukemia. Nature genetics 28, 220-221.

Madsen, J.M., Stoltzfus, C.M., 2006. A suboptimal 5' splice site downstream of HIV-1 splice site A1 is required for unspliced viral mRNA accumulation and efficient virus replication. Retrovirology 3, 10.

Maldarelli, F., Wu, X., Su, L., Simonetti, F.R., Shao, W., Hill, S., . . . Hughes, S.H., 2014. HIV latency. Specific HIV integration sites are linked to clonal expansion and persistence of infected cells. Science 345, 179-183.

Malech, H.L., Maples, P.B., Whiting-Theobald, N., Linton, G.F., Sekhsaria, S., Vowells, S.J., . . . Gallin, J.I., 1997. Prolonged production of NADPH oxidase-corrected granulocytes after gene therapy of chronic granulomatous disease. Proc Natl Acad Sci U S A 94, 12133-12138.

Marino, G., Uria, J.A., Puente, X.S., Quesada, V., Bordallo, J., Lopez-Otin, C., 2003. Human autophagins, a family of cysteine proteinases potentially implicated in cell degradation by autophagy. The Journal of biological chemistry 278, 3671-3678.

Markowitz, S.D., Bertagnolli, M.M., 2009. Molecular Origins of Cancer:

Molecular Basis of Colorectal Cancer. The New England journal of medicine 361, 2449-2460.

Marquet, R., Isel, C., Ehresmann, C., Ehresmann, B., 1995. tRNAs as primer of reverse transcriptases. Biochimie 77, 113-124.

Marshall, H.M., Ronen, K., Berry, C., Llano, M., Sutherland, H., Saenz, D., . . . Bushman, F.D., 2007. Role of PSIP1/LEDGF/p75 in lentiviral infectivity and integration targeting. PloS one 2.

Martin Stoltzfus, C., 2009. Chapter 1 Regulation of HIV-1 Alternative RNA Splicing and Its Role in Virus Replication, Advances in Virus Research. Academic Press, pp. 1-40.

Martinez-Glez, V., Lapunzina, P., 2007. Sotos syndrome is associated with leukemia/lymphoma. Am J Med Genet A 1, 1244-1245.

Marvel, J., Perkins, G.R., Lopez Rivas, A., Collins, M.K., 1994. Growth factor starvation of bcl-2 overexpressing murine bone marrow cells induced refractoriness to IL-3 stimulation of proliferation. Oncogene 9, 1117-1122.

Matreyek, K.A., Engelman, A., 2013. Viral and cellular requirements for the nuclear entry of retroviral preintegration nucleoprotein complexes. Viruses 5, 2483-2511.

Mattaj, I.W., Englmeier, L., 1998. Nucleocytoplasmic transport: the soluble phase. Annual review of biochemistry 67, 265-306.

Matute, J.D., Arias, A.A., Wright, N.A., Wrobel, I., Waterhouse, C.C., Li, X.J., . . . Dinauer, M.C., 2009. A new genetic subgroup of chronic granulomatous disease with autosomal recessive mutations in p40 phox and selective defects in neutrophil NADPH oxidase activity. Blood 114, 3309-3315.

Maus, M.V., Grupp, S.A., Porter, D.L., June, C.H., 2014. Antibody-modified T cells: CARs take the front seat for hematologic malignancies. Blood 123, 2625-2635.

May, C., Rivella, S., Callegari, J., Heller, G., Gaensler, K.M.L., Luzzatto, L., Sadelain, M., 2000. Therapeutic haemoglobin synthesis in [beta]-thalassaemic mice expressing lentivirus-encoded human [beta]-globin. Nature 406, 82-86.

Mazeyrat, S., Saut, N., Sargent, C.A., Grimmond, S., Longepied, G., Ehrmann, I.E., . . . Mitchell, M.J., 1998. The mouse Y chromosome interval necessary for spermatogonial proliferation is gene dense with syntenic homology to the human AZFa region. Human molecular genetics 7, 1713-1724.

Mazzarella, R.A., Green, M., 1987. ERp99, an abundant, conserved

glycoprotein of the endoplasmic reticulum, is homologous to the 90-kDa heat shock protein (hsp90) and the 94-kDa glucose regulated protein (GRP94). The Journal of biological chemistry 262, 8875-8883.

Mbisa, J.L., Barr, R., Thomas, J.A., Vandegraaff, N., Dorweiler, I.J., Svarovskaia, E.S., . . . Pathak, V.K., 2007. Human immunodeficiency virus type 1 cDNAs produced in the presence of APOBEC3G exhibit defects in plus-strand DNA transfer and integration. Journal of virology 81, 7099-7110.

McClure, M.O., Sommerfelt, M.A., Marsh, M., Weiss, R.A., 1990. The pH independence of mammalian retrovirus infection. The Journal of general virology 71 (Pt 4), 767-773.

McGuire, M.J., Lipsky, P.E., Thiele, D.L., 1997. Cloning and characterization of the cDNA encoding mouse dipeptidyl peptidase I (cathepsin C). Biochimica et biophysica acta 1351, 267-273.

Mettus, R.V., Litvin, J., Wali, A., Toscani, A., Latham, K., Hatton, K., Reddy, E.P., 1994. Murine A-myb: evidence for differential splicing and tissue-specific expression. Oncogene 9, 3077-3086.

Miller, A.D., Garcia, J.V., von Suhr, N., Lynch, C.M., Wilson, C., Eiden, M.V., 1991. Construction and properties of retrovirus packaging cells based on gibbon ape leukemia virus. Journal of virology 65, 2220-2224.

Mirza, H., Schmidt, V.A., Derian, C.K., Jesty, J., Bahou, W.F., 1997. Mitogenic responses mediated through the proteinase-activated receptor-2 are induced by expressed forms of mast cell alpha- or beta-tryptases. Blood 90, 3914-3922.

Mitchell, R.S., Beitzel, B.F., Schroder, A.R., Shinn, P., Chen, H., Berry, C.C., . . . Bushman, F.D., 2004. Retroviral DNA integration: ASLV, HIV, and MLV show distinct target site preferences. PLoS biology 2, E234.

Mitrophanous, K., Yoon, S., Rohll, J., Patil, D., Wilkes, F., Kim, V., . . . Mazarakis, N., 1999. Stable gene transfer to the nervous system using a non-primate lentiviral vector. Gene therapy 6, 1808-1818.

Miyauchi, K., Kim, Y., Latinovic, O., Morozov, V., Melikyan, G.B., 2009. HIV enters cells via endocytosis and dynamin-dependent fusion with endosomes. Cell 137, 433-444.

Modell, B., Darlison, M., 2008. Global epidemiology of haemoglobin disorders and derived service indicators. Bulletin of the World Health Organization 86, 480-487.

Modlich, U., Bohne, J., Schmidt, M., von Kalle, C., Knoss, S., Schambach, A., Baum, C., 2006. Cell-culture assays reveal the importance of retroviral vector

design for insertional genotoxicity. Blood 108, 2545-2553.

Modlich, U., Navarro, S., Zychlinski, D., Maetzig, T., Knoess, S., Brugman, M.H., . . . Baum, C., 2009. Insertional transformation of hematopoietic cells by self-inactivating lentiviral and gammaretroviral vectors. Mol Ther 17, 1919-1928.

Moiani, A., Paleari, Y., Sartori, D., Mezzadra, R., Miccio, A., Cattoglio, C., . . . Mavilio, F., 2012. Lentiviral vector integration in the human genome induces alternative splicing and generates aberrant transcripts. The Journal of clinical investigation 122, 1653-1666.

Molling, K., Bolognesi, D.P., Bauer, H., Busen, W., Plassmann, H.W., Hausen, P., 1971. Association of viral reverse transcriptase with an enzyme degrading the RNA moiety of RNA-DNA hybrids. Nat New Biol 234, 240-243.

Montiel-Equihua, C.A., Zhang, L., Knight, S., Saadeh, H., Scholz, S., Carmo, M., . . . Gaspar, H.B., 2012. The beta-globin locus control region in combination with the EF1alpha short promoter allows enhanced lentiviral vector-mediated erythroid gene expression with conserved multilineage activity. Mol Ther 20, 1400-1409.

Montini, E., Cesana, D., Schmidt, M., Sanvito, F., Bartholomae, C.C., Ranzani, M., . . . Naldini, L., 2009. The genotoxic potential of retroviral vectors is strongly modulated by vector design and integration site selection in a mouse model of HSC gene therapy. The Journal of clinical investigation 119, 964-975.

Montini, E., Cesana, D., Schmidt, M., Sanvito, F., Ponzoni, M., Bartholomae, C., . . . Naldini, L., 2006. Hematopoietic stem cell gene transfer in a tumor-prone mouse model uncovers low genotoxicity of lentiviral vector integration. Nature biotechnology 24, 687-696.

Moreno-Carranza, B., Gentsch, M., Stein, S., Schambach, A., Santilli, G., Rudolf, E., . . . Grez, M., 2009. Transgene optimization significantly improves SIN vector titers, gp91phox expression and reconstitution of superoxide production in X-CGD cells. Gene therapy 16, 111-118.

Moser, H.W., Mahmood, A., Raymond, G.V., 2007. X-linked adrenoleukodystrophy. Nature clinical practice. Neurology 3, 140-151.

Mosser, J., Douar, A.M., Sarde, C.O., Kioschis, P., Feil, R., Moser, H., . . . Aubourg, P., 1993. Putative X-linked adrenoleukodystrophy gene shares unexpected homology with ABC transporters. Nature 361, 726-730.

Mukherjee, S., Thrasher, A.J., 2013. Gene therapy for PIDs: progress, pitfalls and prospects. Gene 525, 174-181.

Murakami, T., Freed, E.O., 2000. Genetic evidence for an interaction between

human immunodeficiency virus type 1 matrix and alpha-helix 2 of the gp41 cytoplasmic tail. Journal of virology 74, 3548-3554.

Muromoto, R., Kuroda, M., Togi, S., Sekine, Y., Nanbo, A., Shimoda, K., . . . Matsuda, T., 2010. Functional involvement of Daxx in gp130-mediated cell growth and survival in BaF3 cells. European journal of immunology 40, 3570-3580.

Naldini, L., 2015. Gene therapy returns to centre stage. Nature 526, 351-360.

Naldini, L., Blomer, U., Gallay, P., Ory, D., Mulligan, R., Gage, F.H., . . . Trono, D., 1996. In vivo gene delivery and stable transduction of nondividing cells by a lentiviral vector. Science 272, 263-267.

Nandurkar, H.H., Hilton, D.J., Nathan, P., Willson, T., Nicola, N., Begley, C.G., 1996. The human IL-11 receptor requires gp130 for signalling: demonstration by molecular cloning of the receptor. Oncogene 12, 585-593.

Neckers, L., 2007. Heat shock protein 90: the cancer chaperone. Journal of biosciences 32, 517-530.

Negre, O., Eggimann, A.V., Beuzard, Y., Ribeil, J.A., Bourget, P., Borwornpinyo, S., . . . Payen, E., 2016. Gene Therapy of the beta-Hemoglobinopathies by Lentiviral Transfer of the beta(A(T87Q))-Globin Gene. Human gene therapy 27, 148-165.

Neil, S.J.D., Zang, T., Bieniasz, P.D., 2008. Tetherin inhibits retrovirus release and is antagonized by HIV-1 Vpu. Nature 451, 425-430.

Neklason, D.W., Solomon, C.H., Dalton, A.L., Kuwada, S.K., Burt, R.W., 2004. Intron 4 mutation in APC gene results in splice defect and attenuated FAP phenotype. Fam Cancer 3, 35-40.

Nevins, J.R., Darnell, J.E., Jr., 1978. Steps in the processing of Ad2 mRNA: poly(A)+ nuclear sequences are conserved and poly(A) addition precedes splicing. Cell 15, 1477-1493.

Nielsen, P.J., Trachsel, H., 1988. The mouse protein synthesis initiation factor 4A gene family includes two related functional genes which are differentially expressed. The EMBO journal 7, 2097-2105.

Nishida, K., Kaziro, Y., Satoh, T., 1999. Anti-apoptotic function of Rac in hematopoietic cells. Oncogene 18, 407-415.

Novembre, J., Galvani, A.P., Slatkin, M., 2005. The geographic spread of the CCR5 Delta32 HIV-resistance allele. PLoS biology 3, 18.

Nyamweya, S., Hegedus, A., Jaye, A., Rowland-Jones, S., Flanagan, K.L., Macallan, D.C., 2013. Comparing HIV-1 and HIV-2 infection: Lessons for viral immunopathogenesis. Reviews in medical virology 23, 221-240.

O'Neal, K.D., Yu-Lee, L.Y., 1994. Differential signal transduction of the short, Nb2, and long prolactin receptors. Activation of interferon regulatory factor-1 and cell proliferation. The Journal of biological chemistry 269, 26076-26082.

Oblinger, J.L., Burns, S.S., Akhmametyeva, E.M., Huang, J., Pan, L., Ren, Y., . . . Chang, L.S., 2016. Components of the eIF4F complex are potential therapeutic targets for malignant peripheral nerve sheath tumors and vestibular schwannomas. Neuro-oncology 18, 1265-1277.

Ocwieja, K.E., Brady, T.L., Ronen, K., Huegel, A., Roth, S.L., Schaller, T., . . . Bushman, F.D., 2011. HIV integration targeting: a pathway involving Transportin-3 and the nuclear pore protein RanBP2. PLoS pathogens 7, e1001313.

Okazaki, Y., Furuno, M., Kasukawa, T., Adachi, J., Bono, H., Kondo, S., . . . Hayashizaki, Y., 2002. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. Nature 420, 563-573.

Onafuwa-Nuga, A., Telesnitsky, A., 2009. The Remarkable Frequency of Human Immunodeficiency Virus Type 1 Genetic Recombination. Microbiology and Molecular Biology Reviews 73, 451-480.

Ono, A., Ablan, S.D., Lockett, S.J., Nagashima, K., Freed, E.O., 2004. Phosphatidylinositol (4,5) bisphosphate regulates HIV-1 Gag targeting to the plasma membrane. Proc Natl Acad Sci U S A 101, 14889-14894.

Ono, A., Freed, E.O., 2004. Cell-type-dependent targeting of human immunodeficiency virus type 1 assembly to the plasma membrane and the multivesicular body. Journal of virology 78, 1552-1563.

Osorio, F.G., Navarro, C.L., Cadinanos, J., Lopez-Mejia, I.C., Quiros, P.M., Bartoli, C., . . . Lopez-Otin, C., 2011. Splicing-directed therapy in a new mouse model of human accelerated aging. Science translational medicine 3, 106ra107.

Ott, M.G., Schmidt, M., Schwarzwaelder, K., Stein, S., Siler, U., Koehl, U., . . . Grez, M., 2006. Correction of X-linked chronic granulomatous disease by gene therapy, augmented by insertional activation of MDS1-EVI1, PRDM16 or SETBP1. Nature medicine 12, 401-409.

Overbaugh, J., Miller, A.D., Eiden, M.V., 2001. Receptors and entry cofactors for retroviruses include single and multiple transmembrane-spanning proteins as well as newly described glycophosphatidylinositol-anchored and secreted proteins. Microbiology and molecular biology reviews : MMBR 65, 371-389, table

of contents.

Palacios, R., Steinmetz, M., 1985. Il-3-dependent mouse clones that express B-220 surface antigen, contain Ig genes in germ-line configuration, and generate B lymphocytes in vivo. Cell 41, 727-734.

Palfi, S., Gurruchaga, J.M., Ralph, G.S., Lepetit, H., Lavisse, S., Buttery, P.C., . . . Mitrophanous, K.A., 2014. Long-term safety and tolerability of ProSavin, a lentiviral vector-based gene therapy for Parkinson's disease: a dose escalation, open-label, phase 1/2 trial. Lancet 383, 1138-1146.

Paling, N.R., Welham, M.J., 2002. Role of the protein tyrosine phosphatase SHP-1 (Src homology phosphatase-1) in the regulation of interleukin-3-induced survival, proliferation and signalling. The Biochemical journal 368, 885-894.

Park, L.S., Friend, D., Gillis, S., Urdal, D.L., 1986. Characterization of the cell surface receptor for human granulocyte/macrophage colony-stimulating factor. The Journal of experimental medicine 164, 251-262.

Parrish-Novak, J., Dillon, S.R., Nelson, A., Hammond, A., Sprecher, C., Gross, J.A., . . . Foster, D., 2000. Interleukin 21 and its receptor are involved in NK cell expansion and regulation of lymphocyte function. Nature 408, 57-63.

Paruzynski, A., Arens, A., Gabriel, R., Bartholomae, C.C., Scholz, S., Wang, W., . . . von Kalle, C., 2010. Genome-wide high-throughput integrome analyses by nrLAM-PCR and next-generation sequencing. Nat Protoc 5, 1379-1395.

Pawliuk, R., Westerman, K.A., Fabry, M.E., Payen, E., Tighe, R., Bouhassira, E.E., . . . Leboulch, P., 2001. Correction of sickle cell disease in transgenic mouse models by gene therapy. Science 294, 2368-2371.

Pennica, D., Shaw, K.J., Luoh, S.M., Wood, W.I., 1995. Isolation of cDNA clones encoding the mouse protein V-1. Gene 158, 305-306.

Perkins, G.R., Marshall, C.J., Collins, M.K., 1996. The role of MAP kinase kinase in interleukin-3 stimulation of proliferation. Blood 87, 3669-3675.

Poiesz, B.J., Ruscetti, F.W., Gazdar, A.F., Bunn, P.A., Minna, J.D., Gallo, R.C., 1980. Detection and isolation of type C retrovirus particles from fresh and cultured lymphocytes of a patient with cutaneous T-cell lymphoma. Proc Natl Acad Sci U S A 77, 7415-7419.

Pollard, V.W., Malim, M.H., 1998. The HIV-1 Rev protein. Annual review of microbiology 52, 491-532.

Ponting, C.P., 2008. The functional repertoires of metazoan genomes. Nature reviews. Genetics 9, 689-698.

Pradhan, A., Lambert, Q.T., Reuther, G.W., 2007. Transformation of hematopoietic cells and activation of JAK2-V617F by IL-27R, a component of a heterodimeric type I cytokine receptor. Proc Natl Acad Sci U S A 104, 18502-18507.

Praparattanapan, J., Kotarathititthum, W., Chaiwarith, R., Nuntachit, N., Sirisanthana, T., Supparatpinyo, K., 2012. Resistance-associated mutations after initial antiretroviral treatment failure in a large cohort of patients infected with HIV-1 subtype CRF01_AE. Curr HIV Res 10, 647-652.

Preston, B.D., Poiesz, B.J., Loeb, L.A., 1988. Fidelity of HIV-1 reverse transcriptase. Science 242, 1168-1171.

Pujol, A., Hindelang, C., Callizot, N., Bartsch, U., Schachner, M., Mandel, J.L., 2002. Late onset neurological phenotype of the X-ALD gene inactivation in mice: a mouse model for adrenomyeloneuropathy. Human molecular genetics 11, 499-505.

Purcell, D.F., Martin, M.A., 1993. Alternative splicing of human immunodeficiency virus type 1 mRNA modulates viral protein expression, replication, and infectivity. Journal of virology 67, 6365-6378.

Putnam, D., 2006. Polymers for gene delivery across length scales. Nature materials 5, 439-451.

Rangarajan, A., Weinberg, R.A., 2003. Opinion: Comparative biology of mouse versus human cells: modelling human cancer in mice. Nature reviews. Cancer 3, 952-959.

Ranzani, M., Annunziato, S., Calabria, A., Brasca, S., Benedicenti, F., Gallina, P., . . . Montini, E., 2014. Lentiviral vector-based insertional mutagenesis identifies genes involved in the resistance to targeted anticancer therapies. Mol Ther 22, 2056-2068.

Ranzani, M., Cesana, D., Bartholomae, C.C., Sanvito, F., Pala, M., Benedicenti, F., . . . Montini, E., 2013. Lentiviral vector-based insertional mutagenesis identifies genes associated with liver cancer. Nature methods 10, 155-161.

Rasaiyaah, J., Tan, C.P., Fletcher, A.J., Price, A.J., Blondeau, C., Hilditch, L., . . . Towers, G.J., 2013. HIV-1 evades innate immune recognition through specific cofactor recruitment. Nature 503, 402-405.

Ratner, L., Haseltine, W., Patarca, R., Livak, K.J., Starcich, B., Josephs, S.F., . . . Wong-Staal, F., 1985. Complete nucleotide sequence of the AIDS virus, HTLV-III. Nature 313, 277-284.

Raza, F., Waldron, J.A., Quesne, J.L., 2015. Translational dysregulation in

cancer: eIF4A isoforms and sequence determinants of eIF4A dependence. Biochemical Society transactions 43, 1227-1233.

Relander, T., Johansson, M., Olsson, K., Ikeda, Y., Takeuchi, Y., Collins, M., Richter, J., 2005. Gene transfer to repopulating human CD34+ cells using amphotropic-, GALV-, or RD114-pseudotyped HIV-1-based vectors from stable producer cells. Mol Ther 11, 452-459.

Rezgaoui, M., Hermey, G., Riedel, I.B., Hampe, W., Schaller, H.C., Hermans-Borgmeyer, I., 2001. Identification of SorCS2, a novel member of the VPS10 domain containing receptor family, prominently expressed in the developing mouse brain. Mech Dev 100, 335-338.

Richardson, J.H., Kaye, J.F., Child, L.A., Lever, A.M., 1995. Helper virus-free transfer of human immunodeficiency virus type 1 vectors. The Journal of general virology 76 (Pt 3), 691-696.

Rittner, K., Churcher, M.J., Gait, M.J., Karn, J., 1995. The human immunodeficiency virus long terminal repeat includes a specialised initiator element which is required for Tat-responsive transcription. Journal of molecular biology 248, 562-580.

Robberson, B.L., Cote, G.J., Berget, S.M., 1990. Exon definition may facilitate splice site selection in RNAs with multiple exons. Molecular and cellular biology 10, 84-94.

Robbins, P.F., Kassim, S.H., Tran, T.L., Crystal, J.S., Morgan, R.A., Feldman, S.A., . . . Rosenberg, S.A., 2015. A pilot trial using lymphocytes genetically engineered with an NY-ESO-1-reactive T-cell receptor: long-term follow-up and correlates with response. Clinical cancer research : an official journal of the American Association for Cancer Research 21, 1019-1027.

Roe, T., Reynolds, T.C., Yu, G., Brown, P.O., 1993. Integration of murine leukemia virus DNA depends on mitosis. The EMBO journal 12, 2099-2108.

Rogozin, I.B., Carmel, L., Csuros, M., Koonin, E.V., 2012. Origin and evolution of spliceosomal introns. Biology Direct 7, 11-11.

Romfo, C.M., Alvarez, C.J., van Heeckeren, W.J., Webb, C.J., Wise, J.A., 2000. Evidence for splice site pairing via intron definition in Schizosaccharomyces pombe. Molecular and cellular biology 20, 7955-7970.

Rosa, A., Chande, A., Ziglio, S., De Sanctis, V., Bertorelli, R., Goh, S.L., . . . Pizzato, M., 2015. HIV-1 Nef promotes infection by excluding SERINC5 from virion incorporation. Nature 526, 212-217.

Rous, P., 1910. A Transmissible Avian Neoplasm. (Sarcoma of the Common

Fowl.). The Journal of experimental medicine 12, 696-705.

Rutherford, M.N., Bayly, G.R., Matthews, B.P., Okuda, T., Dinjens, W.M., Kondoh, H., LeBrun, D.P., 2001. The leukemogenic transcription factor E2a-Pbx1 induces expression of the putative N-myc and p53 target gene NDRG1 in Ba/F3 cells. Leukemia 15, 362-370.

Sakamaki, K., Miyajima, I., Kitamura, T., Miyajima, A., 1992. Critical cytoplasmic domains of the common beta subunit of the human GM-CSF, IL-3 and IL-5 receptors for growth signal transduction and tyrosine phosphorylation. The EMBO journal 11, 3541-3549.

Sakuma, T., Barry, M.A., Ikeda, Y., 2012. Lentiviral vectors: basic to translational. The Biochemical journal 443, 603-618.

Sanber, K.S., Knight, S.B., Stephen, S.L., Bailey, R., Escors, D., Minshull, J., . . . Takeuchi, Y., 2015. Construction of stable packaging cell lines for clinical lentiviral vector production. Scientific reports 5, 9021.

Satoh, T., Fantl, W.J., Escobedo, J.A., Williams, L.T., Kaziro, Y., 1993. Platelet-derived growth factor receptor mediates activation of ras through different signaling pathways in different cell types. Molecular and cellular biology 13, 3706-3713.

Sauter, D., 2014. Counteraction of the multifunctional restriction factor tetherin. Front Microbiol 5.

Scaramuzza, S., Biasco, L., Ripamonti, A., Castiello, M.C., Loperfido, M., Draghici, E., . . . Aiuti, A., 2013. Preclinical safety and efficacy of human CD34(+) cells transduced with lentiviral vector for the treatment of Wiskott-Aldrich syndrome. Mol Ther 21, 175-184.

Schaller, T., Ocwieja, K.E., Rasaiyaah, J., Price, A.J., Brady, T.L., Roth, S.L., . . . Towers, G.J., 2011. HIV-1 capsid-cyclophilin interactions determine nuclear import pathway, integration targeting and replication efficiency. PLoS pathogens 7, 8.

Schambach, A., Zychlinski, D., Ehrnstroem, B., Baum, C., 2013. Biosafety features of lentiviral vectors. Human gene therapy 24, 132-142.

Schmidt, M., Schwarzwaelder, K., Bartholomae, C., Zaoui, K., Ball, C., Pilz, I., . . . von Kalle, C., 2007. High-resolution insertion-site analysis by linear amplification-mediated PCR (LAM-PCR). Nature methods 4, 1051-1057.

Scholler, J., Brady, T.L., Binder-Scholl, G., Hwang, W.T., Plesa, G., Hege, K.M., . . . June, C.H., 2012. Decade-long safety and function of retroviral-modified chimeric antigen receptor T cells. Science translational

medicine 4, 132ra153.

Schur, F.K., Hagen, W.J., Rumlova, M., Ruml, T., Muller, B., Krausslich, H.G., Briggs, J.A., 2015. Structure of the immature HIV-1 capsid in intact virus particles at 8.8 A resolution. Nature 517, 505-508.

Schwall, R.H., Chang, L.Y., Godowski, P.J., Kahn, D.W., Hillan, K.J., Bauer, K.D., Zioncheck, T.F., 1996. Heparin induces dimerization and confers proliferative activity onto the hepatocyte growth factor antagonists NK1 and NK2. The Journal of cell biology 133, 709-718.

Scotti, M.M., Swanson, M.S., 2016. RNA mis-splicing in disease. Nature reviews. Genetics 17, 19-32.

Segal, B.H., Leto, T.L., Gallin, J.I., Malech, H.L., Holland, S.M., 2000. Genetic, biochemical, and clinical features of chronic granulomatous disease. Medicine 79, 170-200.

Serra-Moreno, R., 2014. The end of Nef's tether. Trends in microbiology 22, 662-664.

Shakoor, K.A., Saleh, A., Khanzada, M.S., 2002. Usefulness of K-1 (CD-30) marker in Hodgkin's disease. J Pak Med Assoc 52, 442-447.

Shapiro, E., Krivit, W., Lockman, L., Jambaqué, I., Peters, C., Cowan, M., . . . Aubourg, P., 2000. Long-term effect of bone-marrow transplantation for childhood-onset cerebral X-linked adrenoleukodystrophy. The Lancet 356, 713-718.

Shatkin, A.J., 1974. Methylated Messenger RNA Synthesis In Vitro by Purified Reovirus. Proceedings of the National Academy of Sciences 71, 3204-3207.

Sheehy, A.M., Gaddis, N.C., Choi, J.D., Malim, M.H., 2002. Isolation of a human gene that inhibits HIV-1 infection and is suppressed by the viral Vif protein. Nature 418, 646-650.

Shehu-Xhilaga, M., Crowe, S.M., Mak, J., 2001. Maintenance of the Gag/Gag-Pol Ratio Is Important for Human Immunodeficiency Virus Type 1 RNA Dimerization and Viral Infectivity. Journal of virology 75, 1834-1841.

Sherer, N.M., Jin, J., Mothes, W., 2010. Directional spread of surface-associated retroviruses regulated by differential virus-cell interactions. Journal of virology 84, 3248-3258.

Shi, M., Cooper, J.C., Yu, C.L., 2006. A constitutively active Lck kinase promotes cell proliferation and resistance to apoptosis through signal transducer and activator of transcription 5b activation. Mol Cancer Res 4, 39-45.

Shi, Y., Wang, R., Sharma, A., Gao, C., Collins, M., Penn, L., Mills, G.B., 1997. Dissociation of cytokine signals for proliferation and apoptosis. J Immunol 159, 5318-5328.

Shibayama, H., Anzai, N., Ritchie, A., Zhang, S., Mantel, C., Broxmeyer, H.E., 1998. Interleukin-3 and Flt3-ligand induce adhesion of Baf3/Flt3 precursor B-lymphoid cells to fibronectin via activation of VLA-4 and VLA-5. Cellular immunology 187, 27-33.

Shiota, S., von Wronski, M.A., Tano, K., Bigner, D.D., Brent, T.P., Mitra, S., 1992. Characterization of cDNA encoding mouse DNA repair protein O6-methylguanine-DNA methyltransferase and high-level expression of the wild-type and mutant proteins in Escherichia coli. Biochemistry 31, 1897-1903.

Shope, R.E., Hurst, E.W., 1933. Infectious Papillomatosis of Rabbits : With a Note on the Histopathology. The Journal of experimental medicine 58, 607-624.

Sickmier, E.A., Frato, K.E., Shen, H., Paranawithana, S.R., Green, M.R., Kielkopf, C.L., 2006. Structural Basis for Polypyrimidine-Tract Recognition by the Essential pre-mRNA Splicing Factor U2AF(65). Molecular cell 23, 49-59.

Siliciano, P.G., Guthrie, C., 1988. 5' splice site selection in yeast: genetic alterations in base-pairing with U1 reveal additional requirements. Genes & development 2, 1258-1267.

Singh, R., Gupta, S.C., Peng, W.X., Zhou, N., Pochampally, R., Atfi, A., . . . Mo, Y.Y., 2016. Regulation of alternative splicing of Bcl-x by BC200 contributes to breast cancer pathogenesis. Cell death & disease 7, e2262.

Sjoblom, T., Boureux, A., Ronnstrand, L., Heldin, C.H., Ghysdael, J., Ostman, A., 1999. Characterization of the chronic myelomonocytic leukemia associated TEL-PDGF beta R fusion protein. Oncogene 18, 7055-7062.

Smith, A., Ramos-Morales, F., Ashworth, A., Collins, M., 1997. A role for JNK/SAPK in proliferation, but not apoptosis, of IL-3-dependent cells. Curr Biol 7, 893-896.

Spitali, P., Aartsma-Rus, A., 2012. Splice modulating therapies for human disease. Cell 148, 1085-1088.

Staley, J.P., Guthrie, C., 1999. An RNA switch at the 5' splice site requires ATP and the DEAD box protein Prp28p. Molecular cell 3, 55-64.

Stearne, P.A., Pietersz, G.A., Goding, J.W., 1985. cDNA cloning of the murine transferrin receptor: sequence of trans-membrane and adjacent regions. J Immunol 134, 3474-3479.

Stein, S., Ott, M.G., Schultze-Strasser, S., Jauch, A., Burwinkel, B., Kinner, A., . . . Grez, M., 2010. Genomic instability and myelodysplasia with monosomy 7 consequent to EVI1 activation after gene therapy for chronic granulomatous disease. Nature medicine 16, 198-204.

Strang, B.L., Ikeda, Y., Cosset, F.L., Collins, M.K., Takeuchi, Y., 2004. Characterization of HIV-1 vectors with gammaretrovirus envelope glycoproteins produced from stable packaging cells. Gene therapy 11, 591-598.

Stremlau, M., Owens, C.M., Perron, M.J., Kiessling, M., Autissier, P., Sodroski, J., 2004. The cytoplasmic body component TRIM5alpha restricts HIV-1 infection in Old World monkeys. Nature 427, 848-853.

Stutz, F., Izaurralde, E., 2003. The interplay of nuclear mRNP assembly, mRNA surveillance and export. Trends in cell biology 13, 319-327.

Sugnet, C.W., Kent, W.J., Ares, M., Jr., Haussler, D., 2004. Transcriptome and genome conservation of alternative splicing events in humans and mice. Pac Symp Biocomput, 66-77.

Sumantran, V.N., Ealovega, M.W., Nunez, G., Clarke, M.F., Wicha, M.S., 1995. Overexpression of Bcl-XS sensitizes MCF-7 cells to chemotherapy-induced apoptosis. Cancer research 55, 2507-2510.

Suzuki, J., Kaziro, Y., Koide, H., 1997. An activated mutant of R-Ras inhibits cell death caused by cytokine deprivation in BaF3 cells in the presence of IGF-I. Oncogene 15, 1689-1697.

Suzuki, T., Minehata, K., Akagi, K., Jenkins, N.A., Copeland, N.G., 2006. Tumor suppressor gene identification using retroviral insertional mutagenesis in Blm-deficient mice. The EMBO journal 25, 3422-3431.

Suzuki, T., Shen, H., Akagi, K., Morse, H.C., Malley, J.D., Naiman, D.Q., . . . Copeland, N.G., 2002. New genes involved in cancer identified by retroviral tagging. Nature genetics 32, 166-174.

Svasti, S., Suwanmanee, T., Fucharoen, S., Moulton, H.M., Nelson, M.H., Maeda, N., . . . Kole, R., 2009. RNA repair restores hemoglobin expression in IVS2-654 thalassemic mice. Proc Natl Acad Sci U S A 106, 1205-1210.

Sweet, B.H., Hilleman, M.R., 1960. The vacuolating virus, S.V. 40. Proc Soc Exp Biol Med 105, 420-427.

Tabin, C.J., Hoffmann, J.W., Goff, S.P., Weinberg, R.A., 1982. Adaptation of a retrovirus as a eucaryotic vector transmitting the herpes simplex virus thymidine kinase gene. Molecular and cellular biology 2, 426-436.

Tago, K., Kaziro, Y., Satoh, T., 1998. Functional involvement of mSos in interleukin-3 and thrombin stimulation of the Ras, mitogen-activated protein kinase pathway in BaF3 murine hematopoietic cells. J Biochem 123, 659-667.

Takeuchi, K., Kawashima, A., Nagafuchi, A., Tsukita, S., 1994. Structural diversity of band 4.1 superfamily members. Journal of cell science 107, 1921-1928.

Takeuchi, Y., Nagumo, T., Hoshino, H., 1988. Low fidelity of cell-free DNA synthesis by reverse transcriptase of human immunodeficiency virus. Journal of virology 62, 3900-3902.

Tazi, J., Bakkour, N., Marchand, V., Ayadi, L., Aboufirassi, A., Branlant, C., 2010. Alternative splicing: regulation of HIV-1 multiplication as a target for therapeutic action. The FEBS journal 277, 867-876.

Tazi, J., Bakkour, N., Stamm, S., 2009. Alternative splicing and disease. Biochimica et biophysica acta 1792, 14-26.

Tebas, P., Stein, D., Binder-Scholl, G., Mukherjee, R., Brady, T., Rebello, T., . . . June, C.H., 2013. Antiviral effects of autologous CD4 T cells genetically modified with a conditionally replicating lentiviral vector expressing long antisense to HIV. Blood 121, 1524-1533.

Tebas, P., Stein, D., Tang, W.W., Frank, I., Wang, S.Q., Lee, G., . . . June, C.H., 2014. Gene editing of CCR5 in autologous CD4 T cells of persons infected with HIV. The New England journal of medicine 370, 901-910.

Temin, H.M., Mizutani, S., 1970. RNA-dependent DNA polymerase in virions of Rous sarcoma virus. Nature 226, 1211-1213.

Terwilliger, E.F., Godin, B., Sodroski, J.G., Haseltine, W.A., 1989. Construction and use of a replication-competent human immunodeficiency virus (HIV-1) that expresses the chloramphenicol acetyltransferase enzyme. Proc Natl Acad Sci U S A 86, 3857-3861.

Thomas, D.C., Voronin, Y.A., Nikolenko, G.N., Chen, J., Hu, W.S., Pathak, V.K., 2007. Determination of the ex vivo rates of human immunodeficiency virus type 1 reverse transcription by using novel strand-specific amplification analysis. Journal of virology 81, 4798-4807.

Thomas, J., Leverrier, Y., Marvel, J., 1998. Bcl-X is the major pleiotropic anti-apoptotic gene activated by retroviral insertion mutagenesis in an IL-3 dependent bone marrow derived cell line. Oncogene 16, 1399-1408.

Thomas, P., Smart, T.G., 2005. HEK293 cell line: a vehicle for the expression of recombinant proteins. J Pharmacol Toxicol Methods 51, 187-200.

Thrasher, A.J., Burns, S.O., 2010. WASP: a key immunological multitasker. Nature reviews. Immunology 10, 182-192.

Tubo, N.J., Jenkins, M.K., 2014. TCR signal quantity and quality in CD4+ T cell differentiation. Trends in immunology 35, 591-596.

Turunen, J.J., Niemela, E.H., Verma, B., Frilander, M.J., 2013. The significant other: splicing by the minor spliceosome. Wiley interdisciplinary reviews. RNA 4, 61-76.

Ueda, S., Mizuki, M., Ikeda, H., Tsujimura, T., Matsumura, I., Nakano, K., . . . Kanakura, Y., 2002. Critical roles of c-Kit tyrosine residues 567 and 719 in stem cell factor-induced chemotaxis: contribution of src family kinase and PI3-kinase on calcium mobilization and cell migration. Blood 99, 3342-3349.

Uren, A.G., Kool, J., Berns, A., van Lohuizen, M., 2005. Retroviral insertional mutagenesis: past, present and future. Oncogene 24, 7656-7672.

Usami, Y., Wu, Y., Gottlinger, H.G., 2015. SERINC3 and SERINC5 restrict HIV-1 infectivity and are counteracted by Nef. Nature 526, 218-223.

van der Plas, D.C., Smiers, F., Pouwels, K., Hoefsloot, L.H., Lowenberg, B., Touw, I.P., 1996. Interleukin-7 signaling in human B cell precursor acute lymphoblastic leukemia cells and murine BAF3 cells involves activation of STAT1 and STAT5 mediated via the interleukin-7 receptor alpha chain. Leukemia 10, 1317-1325.

van Lohuizen, M., Verbeek, S., Krimpenfort, P., Domen, J., Saris, C., Radaszkiewicz, T., Berns, A., 1989. Predisposition to lymphomagenesis in pim-1 transgenic mice: cooperation with c-myc and N-myc in murine leukemia virus-induced tumors. Cell 56, 673-682.

Vanderver, A., Prust, M., Tonduti, D., Mochel, F., Hussey, H.M., Helman, G., . . . van der Knaap, M.S., 2015. Case definition and classification of leukodystrophies and leukoencephalopathies. Molecular genetics and metabolism 114, 494-500.

Venkatachalam, S., Shi, Y.P., Jones, S.N., Vogel, H., Bradley, A., Pinkel, D., Donehower, L.A., 1998. Retention of wild-type p53 in tumors from p53 heterozygous mice: reduction of p53 dosage can promote cancer formation. The EMBO journal 17, 4657-4667.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., . . . Zhu, X., 2001. The sequence of the human genome. Science 291, 1304-1351.

Votteler, J., Sundquist, W.I., 2013. Virus budding and the ESCRT pathway. Cell host & microbe 14, 232-241.

Wagner, T.A., McLaughlin, S., Garg, K., Cheung, C.Y., Larsen, B.B., Styrchak, S., . . . Frenkel, L.M., 2014. HIV latency. Proliferation of cells with HIV integrated into cancer genes contributes to persistent infection. Science 345, 570-573.

Wain-Hobson, S., Sonigo, P., Danos, O., Cole, S., Alizon, M., 1985. Nucleotide sequence of the AIDS virus, LAV. Cell 40, 9-17.

Walz, C., Crowley, B.J., Hudon, H.E., Gramlich, J.L., Neuberg, D.S., Podar, K., . . . Sattler, M., 2006. Activated Jak2 with the V617F point mutation promotes G1/S phase transition. The Journal of biological chemistry 281, 18177-18183.

Wan, W., Albom, M.S., Lu, L., Quail, M.R., Becknell, N.C., Weinberg, L.R., . . . Cheng, M., 2006. Anaplastic lymphoma kinase activity is essential for the proliferation and survival of anaplastic large-cell lymphoma cells. Blood 107, 1617-1623.

Wang, E.T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., . . . Burge, C.B., 2008a. Alternative isoform regulation in human tissue transcriptomes. Nature 456, 470-476.

Wang, G.P., Berry, C.C., Malani, N., Leboulch, P., Fischer, A., Hacein-Bey-Abina, S., . . . Bushman, F.D., 2010. Dynamics of gene-modified progenitor cells analyzed by tracking retroviral integration sites in a human SCID-X1 gene therapy trial. Blood 115, 4356-4366.

Wang, G.P., Garrigue, A., Ciuffi, A., Ronen, K., Leipzig, J., Berry, C., . . . Bushman, F.D., 2008b. DNA bar coding and pyrosequencing to analyze adverse events in therapeutic gene transfer. Nucleic acids research 36, 14.

Wang, J.K., Gao, G., Goldfarb, M., 1994. Fibroblast growth factor receptors have different signaling and mitogenic potentials. Molecular and cellular biology 14, 181-188.

Wang, M.Y., Cutler, M., Karimpour, I., Kleene, K.C., 1992. Nucleotide sequence of a mouse testis poly(A) binding protein cDNA. Nucleic acids research 20, 3519.

Weinberger, A.D., Weinberger, L.S., 2013. Stochastic fate selection in HIV-infected patients. Cell 155, 497-499.

Weiss, R.A., Vogt, P.K., 2011. 100 years of Rous sarcoma virus. The Journal of experimental medicine 208, 2351-2355.

Weissenhorn, W., Dessen, A., Harrison, S.C., Skehel, J.J., Wiley, D.C., 1997. Atomic structure of the ectodomain from HIV-1 gp41. Nature 387, 426-430.

Wente, S.R., Rout, M.P., 2010. The nuclear pore complex and nuclear transport.

Cold Spring Harbor perspectives in biology 2, a000562.

Wilen, C.B., Tilton, J.C., Doms, R.W., 2012. HIV: cell binding and entry. Cold Spring Harbor perspectives in medicine 2.

Williams, D.A., Lemischka, I.R., Nathan, D.G., Mulligan, R.C., 1984. Introduction of new genetic material into pluripotent haematopoietic stem cells of the mouse. Nature 310, 476-480.

Wilsker, D., Probst, L., Wain, H.M., Maltais, L., Tucker, P.W., Moran, E., 2005. Nomenclature of the ARID family of DNA-binding proteins. Genomics 86, 242-251.

Wirth, B., Brichta, L., Hahnen, E., 2006. Spinal muscular atrophy: from gene to therapy. Seminars in pediatric neurology 13, 121-131.

Wolstein, O., Boyd, M., Millington, M., Impey, H., Boyer, J., Howe, A., . . . Symonds, G.P., 2014. Preclinical safety and efficacy of an anti-HIV-1 lentiviral vector containing a short hairpin RNA to CCR5 and the C46 fusion inhibitor. Molecular therapy. Methods & clinical development 1, 11.

Wong, K.K., Stillwell, L.C., Dockery, C.A., Saffer, J.D., 1996. Use of tagged random hexamer amplification (TRHA) to clone and sequence minute quantities of DNA--application to a 180 kb plasmid isolated from Sphingomonas F199. Nucleic acids research 24, 3778-3783.

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

Yahata, T., Takanashi, T., Muguruma, Y., Ibrahim, A.A., Matsuzawa, H., Uno, T., . . . Ando, K., 2011. Accumulation of oxidative DNA damage restricts the self-renewal capacity of human hematopoietic stem cells. Blood 118, 2941-2950.

Yan, N., Chen, Z.J., 2012. Intrinsic antiviral immunity. Nature immunology 13, 214-222.

Yancopoulos, G.D., Davis, S., Gale, N.W., Rudge, J.S., Wiegand, S.J., Holash, J., 2000. Vascular-specific growth factors and blood vessel formation. Nature 407, 242-248.

Yang, H., Slupska, M.M., Wei, Y.F., Tai, J.H., Luther, W.M., Xia, Y.R., . . . Miller, J.H., 2000. Cloning and characterization of a new member of the Nudix hydrolases from human and mouse. The Journal of biological chemistry 275, 8844-8853.

Yin, H., Kanasty, R.L., Eltoukhy, A.A., Vegas, A.J., Dorkin, J.R., Anderson, D.G., 2014. Non-viral vectors for gene-based therapy. Nature reviews. Genetics 15, 541-555.

Ymer, S., Tucker, W.Q., Sanderson, C.J., Hapel, A.J., Campbell, H.D., Young, I.G., 1985. Constitutive synthesis of interleukin-3 by leukaemia cell line WEHI-3B is due to retroviral insertion near the gene. Nature 317, 255-258.

Yoshinaga, T., Fujiwara, T., 1995. Different roles of bases within the integration signal sequence of human immunodeficiency virus type 1 in vitro. Journal of virology 69, 3233-3236.

Yu, S.F., von Rüden, T., Kantoff, P.W., Garber, C., Seiberg, M., Rüther, U., . . . Gilboa, E., 1986. Self-inactivating retroviral vectors designed for transfer of whole genes into mammalian cells. Proceedings of the National Academy of Sciences 83, 3194-3198.

Yu, X., Yu, Y., Liu, B., Luo, K., Kong, W., Mao, P., Yu, X.F., 2003. Induction of APOBEC3G ubiquitination and degradation by an HIV-1 Vif-Cul5-SCF complex. Science 302, 1056-1060.

Zammarchi, F., de Stanchina, E., Bournazou, E., Supakorndej, T., Martires, K., Riedel, E., . . . Cartegni, L., 2011. Antitumorigenic potential of STAT3 alternative splicing modulation. Proc Natl Acad Sci U S A 108, 17779-17784.

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

Zhang, F., Wilson, S.J., Landford, W.C., Virgen, B., Gregory, D., Johnson, M.C., . . . Hatziioannou, T., 2009. Nef proteins from simian immunodeficiency viruses are tetherin antagonists. Cell host & microbe 6, 54-67.

Zhang, H., Yang, B., Pomerantz, R.J., Zhang, C., Arunachalam, S.C., Gao, L., 2003. The cytidine deaminase CEM15 induces hypermutation in newly synthesized HIV-1 DNA. Nature 424, 94-98.

Zhang, J., Tamilarasu, N., Hwang, S., Garber, M.E., Huq, I., Jones, K.A., Rana, T.M., 2000a. HIV-1 TAR RNA enhances the interaction between Tat and cyclin T1. The Journal of biological chemistry 275, 34314-34319.

Zhang, S., Fukuda, S., Lee, Y., Hangoc, G., Cooper, S., Spolski, R., . . . Broxmeyer, H.E., 2000b. Essential role of signal transducer and activator of transcription (Stat)5a but not Stat5b for Flt3-dependent signaling. The Journal of experimental medicine 192, 719-728.

Zhou, Y.Q., He, C., Chen, Y.Q., Wang, D., Wang, M.H., 2003. Altered

expression of the RON receptor tyrosine kinase in primary human colorectal adenocarcinomas: generation of different splicing RON variants and their oncogenic potential. Oncogene 22, 186-197.

Zhu, L., Zhang, Y., Zhang, W., Yang, S., Chen, J.-Q., Tian, D., 2009. Patterns of exon-intron architecture variation of genes in eukaryotic genomes. BMC Genomics 10, 47.

Zou, J., Presky, D.H., Wu, C.Y., Gubler, U., 1997. Differential associations between the cytoplasmic regions of the interleukin-12 receptor subunits beta1 and beta2 and JAK kinases. The Journal of biological chemistry 272, 6073-6077.

Zufferey, R., Donello, J.E., Trono, D., Hope, T.J., 1999. Woodchuck hepatitis virus posttranscriptional regulatory element enhances expression of transgenes delivered by retroviral vectors. Journal of virology 73, 2886-2892.

Zufferey, R., Nagy, D., Mandel, R.J., Naldini, L., Trono, D., 1997. Multiply attenuated lentiviral vector achieves efficient gene delivery in vivo. Nature biotechnology 15, 871-875.