Lentiviral vector packaging cell line development using genome editing to target optimal loci discovered by high-throughput DNA barcoding

BY ALBERTO MOLINA GIL

INSTITUTE OF CHILD HEALTH / SCHOOL OF PHRAMACY

GLAXOSMITHKLINE

UNIVERSITY COLLEGE LONDON

A thesis submitted for the degree of Doctor of Philosophy

2017

Declaration

I, Alberto Molina Gil, 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

Lentiviral vectors are increasingly used as delivery methods in gene therapy clinical trials due to their high efficiency transducing cells and stability of transgene expression. The development of packaging and producer cell lines for the production of lentiviral vectors has always been a labour-intensive and lengthy process. Sequential introduction of vector components, adaptability to suspension cultures, autotransduction and genetic, transcriptional or cell line growth instability are some of the limitations that cause significant drops in productivity. Improved transcription of self-inactivating vectors leading to high titers has been attempted in different ways with the intent to find a high stable producer clone.

In this project, we studied the use of lentiviral vectors as a tool to target and identify high-transcribing loci in the genome of our host cells for lentiviral packaging cell line development. Third generation lentiviral vectors carrying eGFP under the control of an endogenous clinically-tested promoter (short EF1α) were produced, containing a variable DNA sequence tag (barcode) in their long terminal repeat (LTR). The aim of the barcode is to uniquely tag, identify and track a particular clone within the heterologous expressing population. Human embryonic kidney cell lines (HEK-293) were transduced with a barcoded lentiviral library at a low multiplicity of infection. We demonstrated that integration site analysis and next-generation sequencing of lentiviral barcoded vector junctions by ligation-mediated PCR (LM-PCR) coupled with RNA-Seq allows for quantification of the relative abundance of each barcode variant in each specific genomic position. Expression cassettes containing lentiviral vector components were then site-specifically integrated into these genomes sites using the CRISPR-Cas9 technology.

The barcoding lentiviral system allows for rapid and high-resolution high-throughput screening of gene expression in a large number of genomic positions naturally targeted for optimal vector expression but also of lower expressing sites in order to meet lentiviral cytotoxicity and stoichiometric constraints.

Acknowledgements

I am deeply indebted to my GSK supervisor Conrad Vink, not only for his brilliant science and all I learned from him but also because he has unwittingly taught me how to think in scientific research. Thanks for keeping me motivated and making me enjoy our meetings.

Thank you as well to Ahad Rahim for his supervision and support in every decision, conference, form, etc. Your contributions and feedback in any scientific discussion has been very enriching and your guidance very fruitful.

My sincere appreciation to Steven Howe, my (first) ICH supervisor, for the opportunity to take this PhD and all I have learned, for his help and support when circumstances have been adverse and for putting me in contact with so many enriching people whereof my project has benefited.

I would also like to express my gratitude to Christine Kinnon as well as Nina Kotsopoulou for giving me the opportunity to work in the Molecular Immunology Unit and the Advanced Therapies Delivery teams, respectively. Thank you both for your generosity when I needed it the most.

I am grateful to our collaborator Martijn H Brugman Leids Universitair Medisch Centrum for his help with LAM-PCR and analysis of barcoded libraries. Thanks to Tony Brooks and Mike Hubank for their help with next-generation sequencing. Thanks to Yilong Lee for his computing expertise and Steve Deharo for letting me run integration site analysis pipelines on his server, after my laptop crashed.

At GSK, I would like to thank Eirini for our conversations trying to fix the world. Thanks to Celeste for being my older sister in the lab and Michela for her understanding and patience and also Alessia, Chris, Yu Hua and Sabine because our 'lenti' team has been awesome place thanks to you. Thanks to Ian for his necessary distractions and crazy plans with the van and Simon, my captain for your vision and friendship, the VOQ will never be forgotten. To Sara for the most therapeutic runs ever. Thanks to Mario, Beatriz, Katie, Kenny and everyone in the

Cell and Gene Therapy team at GSK because it has been an amazing group to work in.

Special thanks to my ICH lab mates Mustafa, Kathy, Anna, Jo H and Jonathan who made me feel very welcome in Ste's group and also to everyone at MCI/MIU who contributed to its wonderful atmosphere; in particular to Neelam, Kathryn, Diego, Pili, Ben, Miguel, Sujal, Mike, MariaA, Dave, Jude, Julien, Dessi, Karolin, Marlene, Emma and Christos to keep me connected with ICH while I was away.

Finally, I would also like to thank GlaxoSmithKline and the Biotechnology and Biological Sciences Research Council for funding the project I present.

Y ahora que no nos oye nadie, también quería dar las gracias a una serie de personas que, a otro nivel, han estado detrás apoyándome durante todo este tiempo. En primer lugar, sin duda alguna, a Estrella, porque día tras día me has acompañado también en esta etapa y me parece ya imposible entenderse a uno mismo sin mirarse en el otro. Al Sergi i a la Marta per compartir tantes estones plenes de riures que es troben molt a faltar. Special thanks pels meus amics Marc and Adrià, per riure'ns tant, per unes nits i converses tan maques com necessàries i per mantenir-me conectat amb la vida i faràndula de la capital. 'Nylons' una abraçada molt gran i sincera, us he d'agraïr molt.

Mil gracias a mi familia: a mis abuelos, a mi tía y a mis tíos por vuestro apoyo todo este tiempo. Pero sobretodo quiero agradecerles a mis padres y a mi hermano la fuerza incondicional que me han dado cuando las cosas no iban bien; la mayoría de cosas de las que puedo presumir hoy en día son vuestras, como también es vuestro este logro. A ti mamá porque cuando he estado lejos te he sentido más cerca que nunca y me has demostrado que estaba equivocado contigo; ahora sí, se bienvenida a tu catedral. A tu, pare, moltes gràcies també perquè amb 'comportació' puc dir que ho he sabut disfrutar i ho he tirat 'endavant'. No m'hauria perdonat mai haver-te fallat. Y a ti Alejandro porque tu potencial me fascina tanto como me inspira y por ser una de las personas que me entienden mejor.

A l’ocell blau

Table of contents

List of figures

List of tables

List of Abbreviations

A	Adenine
AAV	Adeno-associated virus
ACE	Artificial Chromosome Expression
Ad/ADV	Adenovirus
ADA	Adenosin deaminase
Ag	Antigen
AIDS	Acquired immune deficiency syndrome
ALD	Adrenoleukodystrophy
ANOVA	Analysis of variance
APRT	Adenine phosphoribosyltransferase
ARE	Adenylate-uridylate-rich elements
ASLV	Avian sarcoma-leukosis virus
ASV	Avian sarcoma virus
ATP	Adenosine triphosphate
BALB	Bagg Albino (inbred research mouse strain)
BAF	Barrier to autointegration factor
BLAT	Basic Local Alignment Tool
Bleo	Bleomycin
BHK	Baby hamster kidney
BET	Bromodomain and Extra-Terminal
bp	Base pair
BSA	Bovine serum albumin
BWA	Burrows-Wheeler Aligner
C	Cytosine
CA	Capsid protein or State of California
CAGE	Cap Analysis of Gene Expression
CAR	Chimeric antigen receptor
^{o}C	Degrees Celsius
$CaCl_2$	Calcium chloride
CCR5	Chemokine receptor type 5
cDNA	Complementary deoxyribonucleic acid
CD	Cluster of differentiation
CDA	Cytosine deaminase
CDS	Coding DNA sequence
CHMP	Committee for Medicinal Products for Human Use
CHO	Chinese hamster ovary
Chr	Chromosome
CIS	Common integration/insertion sites
CMC	Chemistry Manufacturing and Controls
CMV	Cytomegalovirus
CO_2	Carbon dioxide
CNS	Central nervous system
cPPT	Central polypurine tract
Cre	Causes recombination
crRNA	CRISPR RNA
CRISPR	Clustered regularly interspaced short palindromic repeats
cSH4	Chicken β-globin insulator

cSIN	Conditional self-inactivating
C-ter	C-terminal
CT	Cycle threshold
CTE	Constitutive Transport Element
CTS	Central termination sequence
C3-Sp	C3-Spacer
D	Adenine, guanine or thymine
DC	Dendritic cell
DHFR	Dihydrofolate reductase
DMEM	Dulbecco's Modified Eagle's Medium
DMSO	Dimethyl sulfoxide
DNA	Deoxyribonucleic acid
dNTPs	Deoxyribonucleotides triphosphate
ds	Double stranded
DSB	Double strand break
DT	Diphteria toxin
DTT	Dithiothreitol
EBNA	Epstein Barr nuclear antigen
EBV	Epstein Barr virus
E.coli	Escherichia Coli
EDTA	Ethylenediaminetetraacetic acid
EEW	EFS-eGFP-WPRE
EF1α	Elongation factor 1 α subunit
EFS	EF1α short promoter
eGFP	Enhanced green fluorescent protein
ELISA	Enzyme-linked immunosorbent assay
EMA	European Medicines Agency
Env	Envelope
ESC	Embryonic stem cell
EST	Expressed Sequence Tag
F	Farad
FACS	Fluorescence-activated cell sorting
FCS/FBS	Fetal calf/bovine serum
FDA	Food and drug Administration
FISH	Fluorescence in situ hybridisation
FITC	Fluorescein isothiocyanate
FLASH	Fast ligation-based automatable solid-phase high-throughput
Flp	Flippase
FRT	Flp recongnition targets
FU	Fluorescent units
FV	Foamy viruses
Fwd	Forward
G	Guanine
GALV	Gibbon ape leukaemia virus
Gag	Group-specific antigen
g	Gram or gravitational force units
gDNA	Genomic deoxyribonucleic acid
GDP	Guanosine diphosphate

GMP	Good manufacturing practices
GHT	Glycine, hypoxanthine, thymidine
GIRI	Genetic Information Research Institute
GM-CSF	Granulocyte-macrophage colony-stimulating factor
GOI	Gene of interest
gp	Glycoprotein
GRAS	Generally regarded as safe
GS	Glutamine synthetase
GRCh37	Genome Reference Consortium Human genome build 37
GSK	GlaxoSmithKline
GTP	Guanosine triphosphate
h	Hours
H	Adenine, cytosine or thimine
HA	Homology arms
HAC	Human artificial chromosome
HBV	Hepatitits B virus
HBS	Hepes buffer saline
HDAC	Histone deacetyilase
HDR	Homology directed repair
HeLa	Cell line derived from cervical cancer
HEK	Human Embryonic Kidney
hERVS	Human endogenous retroviral sequences
hisD	Histidinol dehydrogenase
HIV	Human Immunodeficiency Virus
HMG	High-mobility-group
HPH	Hygromycin-B-phosphotransferase
HPRT	Hypoxanthine-guanine phosphoribosyltransferase
HR	Homologous recombination
HSC	Hematopoietc stem cell
HSR-TIGET	Hospitale San Raffaele- Telethon Institute for Gene Therapy
HSV	Herpes simplex virus
HTLV	Human T-lymphotropic virus
HTP	High-throughput
IBD	Integrase binding domain
ICH	Institute of Child Health
ID	Identity
IDLV	Integration deficient lentiviral vector
IL2RG	Interleukin-2 receptor common gamma chain
IN	Integrase
Indel	Insertion and/or deletion
(int)	Integration
iPSC	Induced pluripotent stem cell
IRES	Internal ribosome entry site
IS	Insertion/integration site
ITR	Internal Terminal Repeats
IUPAC	International Union of Pure and Applied Chemistry
k	Kilo (10^3)
KCl	Potassium chloride
KOH	Potassium hydroxyde

L	Litre
LCR	Locus control regions
LDC	Limiting dilution cloning
LEDGF	Lens epithelium-derived growth factor
LIC	Ligation independent cloning
LINE	Long interspersed elements
LM-PCR	Ligation-mediated polymerase chain reaction
LN_2	Liquid nitrogen
loxP	locus of X-over of P1
LPS	Lipopolysaccharide
LTR	Long terminal repeat
LV	Lentivirus or lentiviral
LVV	Lentiviral vectors
m	Milli (10^{-3})
M	Molar or million
MA	Matrix
MCS	Multicloning site
$MgSO_4$	Magnesium sulphate
$MgCl_2$	Magnesium chloride
min	Minute
mAb	Monoclonal antibody
MLD	Metachromatic leukodystrophy
MLV	Murine leukaemia virus
MMCT	Microcell-mediated chromosome transfer
MN	Meganuclease
$MnCl_2$	Manganese chloride
MRN	Mre11-Rad50-Nbs1 complex
mRNA	Messenger ribonucleic acid
MOI	Multiplicity of infection
mRNA	Messenger ribonucleic acid
MSX	Methionine sulphoximine
MTX	Methotrexate
μ	Micro (10^{-6})
n	Nano(10^{-9})
NaBu	Sodium butirate
NC	Nucleocapsid
NCBI	National Center for Biotechnology Information
Nef	Negative regulatory factor
Neo	Neomycin
NHEJ	Non-homologous end joining
NIH	National Institutes of Health
NLS	Nuclear leading sequences
nm	Nanometer (10^{-9} meters)
NaN_3	Sodium azide
NILV	Non-integrating lentiviral vector
NGS	Next-generation sequencing
NRC-BRI	National Research Council of Canada-Biotechnology Research institute
NSG	NOD scid gamma (NOD.Cg-Prkdc^{scid} $\mathit{Il2rg}^{tm1Wjl}$/SzJ

NS0	Cell line derived from murine myeloma cells
Nt (or N)	Nucleotide
NTC	Non-template control
N-ter	N-terminal
NUC	Nuclease subunit of the CRISPR-Cas9 complex
ON	Overnight
ORF	Open reading frame
P	Poisson
PAC	Puromycin-N-acetyl transferase
PAM	Protospacer adjacent motif
pmol	picomole (10^{-12} mol)
PBS	Primer binding sites
PCL	Packaging/producer cell line
PCR	Polymerase chain reaction
PD	Parkinson's disease
PE	Paired end
PEG	Polyethylenglycol
PEI	Polyethylenimine
PEST	Polypeptide regions rich in proline (P), glutamic acid (E), serine (S), and threonine (T)
PGK	Phosphoglygerate kinase
PH	Pleckstrin homology
Phleo	Phleomycin
PI	Propidium iodide
PIC	Pre-integration complex
Pol	Polymerase
PNK	Polynucleotide kinase
PR	Protease
PRE	Posttranscriptional regulatory element
Puro	Puromycin
PSIP1	PC4 and SFRS1 interacting protein 1
PTM	Post-translational modification
PWWP	Pro-Trp-Trp-Pro
q	Specific productivity
qPCR	Quantitative polymerase chain reaction
R	Repeat region
RCA	Rolling circle amplification
rCHO	Recombinant Chinese Hamster ovary cell
RCL	Replication competent lentivirus
RCR	Replication competent retrovirus
REAL	Restriction Enzyme And Ligation
REC	Recognition subunit of CRISPR-Cas9 complex
Rev	Reverse or regulator of expression of virion proteins
RFP	Red fluorescent protein
RGEN	RNA-guided endonucleases
RMCE	Recombinase mediated cassette exchange
RNA	Ribonucleic acid
RNA-Seq	NGS of RNA
RNP	Ribonucleoprotein

Rpm	Revolutions per minute
RRE	Rev response element
RRV	Ross River virus
RSV	Rous sarcoma virus
RT	Reverse transcription/reverse transcriptase or room temperature (25^{o}C)
RTC	Reverse transcription complex
RV	Retrovirus or retroviral
RVV	Retroviral vectors
R1 or R2	NGS read1 or read2 dataset
S	Guanine or cytosine (G or C in IUPAC nomenclature)
SA	Suspension adapted
SBE	Society for Biological Engineering
SD or s.d.	Standard deviation
SE	Single end
seBFP	strongly enhanced BFP
SFDA	State Food and Drug Administration
SFFV	Spleen focus-forming virus
SFV	Semliki Forest virus
sgRNA	Single guide RNA
SIN	Self Inactivating
SINE	Short interspersed elements
S/MAR	scaffold/matrix attachment region
SMRT	Single-molecule real-time
SOC	Super optimal broth with catabolite repression
SF	Serum free
SFFV	Spleen focus-forming virus
SP2/0	Cell line from murine myeloma
ss	Single stranded
SU	Surface protein
SV40	Simian virus 40
SYNT	Synthetic
T	Thymine
TAE	Tris-acetate-EDTA
TALENs	Transcription activator-like effector nucleases
TAR	Trans-activation response element
Tat	Transactivator protein
Tas	Transactivator of spumavirus
TCA	Tricarboxylic acids
TCL	Temperature cycling ligation
TCR	T-cell receptor
TE	Transfection efficiency of Tris-EDTA
Tet	Tetracycline
Tet	tetracycline repressor
TK	Thymidine kinase
TIL	Tumour infiltrating lymphocytes
TM	Transmembrane protein
TPA	Tissue plasminogen activator
tracrRNA	Trans-activating crRNA

TRE	Tet regulatory element
TSS	Transcription start sites
tTA	Tetracycline-controlled transactivator
TU	Transducing units
TV	Transfer vector
U	Unit
UCL	University College London
UCOE	Ubiquitous chromatin opening element
UCSC	University of California Santa Cruz
UCLA	University of California Los Angeles
U3	Unique at 3' region
U5	Unique at 5' region
UT	Untransduced
UTR	Untranslated terminal repeat
UV	Ultraviolet
V	Volt
VacV, VV	Vaccinia virus
VISA	Vector Integration Site Analysis
vLB	Vegitone lysogeny broth
Vpr	Viral protein R
Vpu	Viral protein U
Vpx	Viral protein X
v/v	Volume/volume
VSV-G	Vesicular stomatitis virus glycoprotein
W	Adenine or thymine (A or T in IUPAC nomenclature)
WAS	Wiskott Aldrich syndrome
WHV	Woodchuck hepatitis virus
WPRE	Woodchuck hepatitis virus posttranscriptional pegulatory element
w/v	Weight/volume
w/w	Weight/ weight
X-CGD	X-linked chronic granutomalous disease
XGPRT	Xanthine-guanine phosphoribosyltransferase
X-SCID	X-linked severe combined immunodeficiency
Zeo	Zeocin
ZFN	Zinc finger nuclease
ZFP	Zinc finger protein
+ve	Positive
-ve	Negative
-d	Distance (Starcode parameter)
-r	Ratio (Starcode parameter)
ψ	Packaging signal
Ω	Ohms

Chapter 1

INTRODUCTION

1.1 Lentiviral gene therapy and packaging cell lines

1.1.1 A brief overview of the history of gene therapy

Long before the concept of the gene was discovered, the human race has intentionally bred animals and plants with the intent of achieving more productive specimens and consequently meet the increasing nutritional demands. This early form of genetic engineering (or selection) has persisted over generations as discussed in the Sixth International Congress of Genetics in 1932 in Ithaca[1]. With the discovery of gene transmission via nucleic acids by Avery *et al.,* in 1944[2] and the posterior discovery of the structure of DNA by Watson, Crick (and Franklin) in 1953[3,4], the use of genetic concepts changed the scope towards a potential therapeutic application. As Avery stated in his article: "*Biologists have long attempted by chemical means to induce in higher organisms predictable and specific changes which thereafter could be transmitted in series as hereditary characters".* Such *chemical means* turned out to imply viruses when some authors demonstrated their ability to transfer genes into bacterial host genomes (i.e., *Salmonella* and bacteriophages)[5]. These findings were extended to animal cells when Rous sarcoma virus (RSV) and Simian virus (SV40) viral genes were found to be responsible for cell transformation. In 1966, Edward Tatum speculated about the potential of this tool affirming that *"viruses will be effectively used for man's benefit, in theoretical studies in somatic-cell genetics and possibly in genetic therapy…"*

Although the terms 'gene surgery' and 'gene therapy' were originally coined in the early 1960s, the genetic code, recombinant DNA and prokaryotes relegated the general interest to the end of the next decade. Nonetheless, a first (and premature) early approach of gene therapy was put into practice in the late 1960s when Rogers *et al.,* treated hyperargininaemia patients with whole Shope papilloma virus arguing that the viral genome contained a copy of the arginase gene[6,7]. However, the experiment did not show any influence on the metabolic profile of the patients[8]. From the 1970s, the recombinant DNA era revolutionised the field and provided the tools to feasibly develop conceptual gene engineering. The improvement in the knowledge of molecular genetics and gene delivery methods such as with calcium phosphate[9] and the subsequent proof-of-concept *in vitro* correction of hypoxanthine-guanine phosphoribosyl transferase (HPRT) deficient cells[10] (although its efficiency was not sufficient for treating patients) as well as the general acceptance that viruses could be used for therapeutic use led to a new reconsideration of the potential of gene therapy. A clear reflection of that perception was the title of Anderson's article *"Gene therapy in human beings: When is it ethical to begin?"*[11]

In a study that was probably considered ahead of its time, the controversial 'Cline experiment' demonstrated the first non-viral gene therapy clinical trial published in Nature in 1980. His team transfected bone marrow cells with the beta-globin gene using the calcium phosphate technique and transplanted them back into patients[12]. Although the theoretical key principles behind this approach were reasonable, the experiment provided no meaningful data and such practices were methodologically and scientifically questioned[13,14] (costing him his chairmanship at UCLA and his NIH funding) and highlighted the important role of regulatory agencies on the authorization of such practices in the clinic. Proof of this was the creation of the DNA Advisory Committee of the NIH, in 1974, to specifically regulate any activities derived from the use of DNA as a therapeutic tool.

During the 1980s, and thanks to the advances in the understanding of molecular biology of retroviruses (in particular the discovery of RNA polymerase by Mizutani Temin and Baltimore in 1970[15,16]), the first retroviral vectors were developed and demonstrated efficacy in complementing patient cells defective in HPRT[17,18] and two years later in ADA-SCID models[19]. The first attempt to utilise vector-mediated gene-modified cells in humans took place in 1998; Rosenberg *et al.,* demonstrated safety and feasibility in patients with advanced melanoma that were successfully treated using retrovirally marked tumor-infiltrating lymphocytes (TIL) with a gene conferring resistance to G418 (a neomycin analogue)[20]. In the 1990s, a better understanding of viral vectors combined with the expertise acquired in DNA-manipulating techniques made researchers raise expectations that gene therapy would eventually provide a safe and feasible alternative to allogeneic bone marrow transplant for the treatment of hereditary monogenic diseases (when there is no suitable donor match). Blaese and Anderson (NIH, 1990) and Bordignon and Mavilio (HSR-TIGET, 1992) performed the first approved clinical trials using retroviral-mediated gene transfer of the adenosine deaminase (ADA) gene in T cells and peripheral blood lymphocytes (and hematopoietic stem cells), respectively. Both attempts led to short-term immune system reconstitution and temporary response although the treatment was not completely curative since the patients continued requiring enzyme replacement therapy[21,22].

A few years later, the death of 18 year-old patient Jesse Gelsinger (suffering from a mild ornithine transcarbamylase deficiency) as a result of adenoviral vector-associated toxicity and subsequent multi-organ failure struck the gene therapy scientific community and caused a great stir in the media. The investigation revealed serious violations on the reporting of previous adverse events incompatible with the inclusion criteria of the clinical trial and resulted in serious fines and other consequences for the leading researchers and institutions[23].

In the next decade, gene therapy treatment for X-linked severe combined immunodeficiency (X-SCID) resulted in the sustained expression of the functional IL-2 common γ-chain receptor (IL2RG) followed by consequent successful repopulation of the patient's bone marrow and restoration of the immune function[24]. However, early successes were also accompanied by serious side effects derived from the integration properties of gamma-retroviral vectors used for the gene delivery. Cavazzana-Calvo's team reported the onset of leukaemia in five of the twenty patients (of whom one died) as a consequence of insertional mutagenesis of murine leukaemia gamma-retroviral vector in the LMO2 proto-oncogene leading to its activation and T-cell proliferation. Similar results were obtained in the British clinical trial led by Thrasher and Gaspar[25]. The Jesse Gelsinger case and the lethal case of leukaemia highlighted the need for greater vigilance, safety studies and investigation into any potential adverse effects associated to viral vectors.

The monogenic nature of primary immunodeficiencies makes them attractive targets for gene therapy. Although allogeneic haematopoietic stem cell transplantation provides a curative option and new protocols are reducing the effects of conditioning chemotherapy and graft-versus host diseases, gene therapy has demonstrated to be an efficacious alternative, especially for patients for whom HLA–matching donors are not available. Autologous transplantation of stably genetically modified hematopoietic stem/progenitor CD34+ cells has been able to treat a sizable number of hereditary rare diseases within the last 25 years.

In the Italian trial led by Roncarolo and Bordignon, patients with adenosine deaminase deficiency (ADA-SCID) with no enzyme replacement therapy available responded satisfactorily to engraftment of engineered HSC with non-myeloablative conditioning and did not present any adverse effects after up to 15 years of follow-up. Moreover, polyethylene glycol–modified bovine ADA (PEG-ADA) discontinuation (3 months after reinfusion) favoured the selective outgrowth of transduced T lymphocytes, which led to sustained immune function restoration[26,27]. Similar results were observed in the parallel British clinical trial[28]. Diseases affecting the myeloid compartment such as X-linked chronic granulomatous disease (X-CGD) have also been successfully treated; two adults received retroviral transfer of gp91^{phox} that restored the oxidative antimicrobial activity of phagocytes[29]. In the case of X-linked adrenoleukodystrophy, a fatal brain demyelinating disease, Cartier *et al.*, demonstrated safety and efficacy using 3rd generation lentiviral vectors to complement the ABCD1 faulty gene and maintain ALD protein expression up to 16 month post-infusion[30]. More prevalent inherited diseases like beta-haemoglobinopathies –affecting adult age patients- have also benefited from gene therapy. Trials for beta-thalassemia have showed promising results after 33 months following HMGA2 gene transfer using lentiviral vectors. Recently, two more Phase I/II trials using lentiviral vectors have approached the treatment of Wiskott–Aldrich syndrome (WAS)[31] and metachromatic leukodystrophy (MLD) *ex vivo*[32]. *ProSavin*® vector has been used for the *in vivo* treatment of Parkinson disease (PD) in a Phase I/II trial[33]. The successful application of retroviral vetors has subsequently led to significant milestone of GSK releasing *Strimvelis*®*,* a commercial retroviral gene therapy product for ADA-SCID primary immunodeficiency[34] in 2016. Previously, only UniQure's *Glybera*®, a variant (Ser447X) of the human LPL gene under the control of the CMV promoter transferred using recombinant adeno-associated virus type 1 (rAAV1) vector, had received market authorisation in 2012 for a subcohort group of patients with severe pancreatitis attacks. However, the lack of demand forced UniQure to withdraw *Glybera*® from the market. Other gene therapy products approved by Chinese regulatory agencies comprise *Gendicine*® (SiBiono GenTech) a recombinant adenoviral p53 (rAd-p53) gene-replacement for head and neck squamous cell carcinoma or *Oncorine*® (a recombinant oncolytic human adenoviral vector type 5, rAd5) for nasopharingeal carcinoma in 2005 by Shanghai Sunway Biotech. Besides the aforementioned products, only the Russian *Neovasculgen*® (intramuscular injection of a plasmid encoding the pCMV-vegf165 cassette, carrying vascular endothelial growth factor) to treat peripheral arterial disease (PAD, including atherosclerotic lower limb ischemia) received national marketing authorization in 2012[35]. More recently, BioVex/Amgen released *OncoVex/T-Vec*® (talimogene laherparepvec, an oncolytic HSV) in 2015 for the treatment of advanced melanoma.

1.1.2 Vectors as tools for gene delivery

Considering the wide range of potential applications with an increasing demand that gene therapy is required to address, this technology faces several conceptual and technical limitations. Although the efficiency of gene delivery of a therapeutic gene to a population of dividing and non-dividing cells remains as the main concern, current drawbacks also comprise the sustainability of expression in the target tissue, potential adverse events derived from insertional mutagenesis of integration vectors (also referred as genotoxicity), the potential host immune response and the high costs associated to solve the aforementioned complications. In order to provide efficient and safe gene delivery of the therapeutic gene in different medical contexts, different types of viral and non-viral vectors have been engineered and developed to fulfil the conditions of each treatment.

In general, non-viral methods (grouped in naked DNA, cationic lipids and molecular conjugates) are easier to produce and scale up[36], allow larger genetics payloads and present lower immunogenicity and carcinogenesis as no integration of the transgene into the host genome occurs. Nonetheless, one of the main challenges to overcome in order to augment the current (21%) representation in on-going clinical trials is the low gene delivery efficacy together with poor nanoparticle stability[37]. However, as most of the current approaches use viral vectors, we will focus the scope of this research study on them.

Viruses have naturally evolved to target cells and transfer their gene content to be replicated. Transduction of viral vectors enables highly efficient gene transfer with a lower impact on the cell physiology and viability of the target cells. In addition, some viral (and modifiable) surface molecules confer specific tropism. Non-essential *cis* viral components have been replaced with therapeutic cassettes for viral delivery. A number of virus families have been explored for gene delivery

Adenoviruses are linear dsDNA viruses that can infect both dividing and non-dividing cells and cause common upper respiratory infections in humans. First generation replication-deficient adenoviral vectors (ADV, Ad) were first engineered by replacing the E1 protein with the gene of interest and can be prepared at concentrations higher than other vectors (10^{11} - 10^{12} particles per mL)[38]. Three generations of adenoviral vectors have been generated by deleting viral genes (also termed *gut-less* or *gutted)* showing higher capacity (up to 36kb), extended expression periods and reduced immune response. ADV are characterised by being capable of inducing a potent immune response and thus, short-term expression, which make them useful for cancer applications. The first trial approved using Ad (in which Jesse Gelsinger took part) was for the delivery of the OTC (ornithine transcarbamoylase) gene[39].

Vaccinia virus (VV or VacV) is a dsDNA virus, member of the *orthopoxviridae* family that has shown promise in *in vivo* gene delivery applications. VV remains episomal and its arrest of the host protein machinery function allows it to form mature virions 4-6 hours post-infection. Its broad tropism, high capacity and levels of transgene expression compensate the large size of viral particles and make it a good candidate for gene therapy treatment for cancer[40]. Like herpesviruses, poxviruses have been used as oncolytic vectors for the treatment of cancer. Oncolytic viruses selectively replicate in cancer cells directly inducing their lysis or triggering an immune modulatory response towards cancer cells. Moreover, additional selective targeting can be achieved by genetic engineering of the vector. Despite the entry pathway driving hepatic tropism of VV has not been characterised in detail to date, heparan sulfate proteoglycans have been suggested as potential receptors. Oncolytic viral products like JX-594 (by Jennerex) armed with granulocyte macrophage colony stimulating factor (GM-CSF) successfully completed Phase II trials for hepatocellular carcinoma[41].

Herpes Simplex Viruses (HSV) have also been extensively used as gene transfer vector. Their broad tropism, very high packaging capacity (the viral genome is 153kb) and its ability to transduce dividing and quiescent cells (especially neurons) have attracted the interest of researchers for the treatment of neurological disorders. In addition to that, the vector remains episomally recircularised in the nucleus and is thought to replicate as a concatemer via a rolling circle-like mechanism[42]. Different types of HSV-1 vectors exist: (i) replication defective rHSV-1 vectors lack essential replication genes, which are provided in trans by the cell line, (ii) attenuated rHSV-1 carry deletions in non-essential genes that hamper their replication *in vivo* but not *in vitro*. (iii) HSV-1-derived amplicon vectors are a safer approach since the amplicon plasmid only contains an origin of viral replication, a packaging signal as well as the cassette of interest. Therefore, replication, structural and packaging proteins must be supplied in *trans* by a helper/packaging cell line (lacking packaging signal). The historical concern in HSV has been the presence of contaminating helper viruses in vector stocks, which can induce immune responses.

Adeno-associated viruses (AAV) are 25nm, non-enveloped ssDNA virus, members of the *Parvoviridae* family and not currently known to be pathogenic for humans. Their genome consists of two ORFs (*cap* and *rep*) that contain the genes responsible for viral replication (normally replaced with transgene for gene therapy applications) flanked by 145-bp inverted terminal repeats (ITRs). *Cap/rep*-deficient AAVs require helper functions to be provided in *trans* (adenovirus or herpes simplex virus) to enable the expression of the *cap* gene and thus synthesis of the capsid proteins[43]. Recombinant adeno-associated viral (rAAV) vector production is commonly achieved by cell lysate harvest 72 hours prior to triple transfection into human embryonic kidney (HEK) 293 cells. Following promising clinical trials results, type 1 rAAV encoding for the LPL^{S447X} gene variant became the first commercially available gene therapy product (alipogene tiparvovec or *Glybera®* by UniQure*)* in the Western world after the approval from the European Union in October, 2012 for the treatment of familial lipoprotein lipase deficiency suffering from pancreatitis attacks. rAAVs have also demonstrated their clinical efficacy and safety in the treatment of cystic fibrosis[44], Leber's congenital amaurosis type 2[45], choroideremia[46], and hemophilia B[47] and other neurological disorder due to its ability to travers the blood brain barrier. However, among their downsides limited capacity (5kb) and the pre-existing immunity to some AAV serotypes[48]. Recently, in a controversial article, insertional mutagenesis caused by AAV2 vectors has been shown to result in hepatocellular carcinoma[49], contrary to the safe profile associated to AAV vectors.

Among retroviruses, gamma-retroviruses are historically the most widely used for gene therapy applications, mainly for the availability of cell lines for their production. Gamma-retroviruses differ from lentiviruses in several ways. Splicing is regulated by the formation of a secondary structure in the RNA upstream of the major splice donor, which results in 2 species of mRNA[50], rev protein is not required to export unspliced mRNA[51]. In terms of viral integration, gamma-retroviruses also integrate in gene-rich regions but, unlike lentiviruses, they do not interact with LEDGF/p75, which shifts their integration pattern towards the body of actively transcribed genes[52]. Gamma-retroviruses cannot pass through the nucleopore and thus require nuclear envelope breakdown to access the cell genome. As a consequence, they cannot infect non-dividing cells and integrate more frequently in cell cycle-related genes. Gamma-retroviruses display a more genotoxic integration profile as a result of the positional effect of the vector components. Upon integration of a retroviral vector, the activity of the internal promoter or 5'LTR U3 promoter/enhancer vector regions (if the vector is not self-inactivated) can induce the expression of proto-oncogenes located upstream or downstream of the vector [53]. Alternatively, the integration of the vector within a tumor suppressor gene can disrupt its function might also trigger tumor activity.

Foamy viruses (FV) are members of the *Spumaretrovirinae* subfamily that owe their name to the appearance of cells under cytopathic effect[54]. Their life cycle is particularly different from the other retroviruses because its reverse transcription takes place during viral assembly prior to the budding of the viral particle[55]. This allows for longer persistence of infective dsDNA, especially in quiescent target cells. However, FV can also transduce non-dividing cells as nuclear membrane breakdown is not required for the FV pre-integration complex to enter the nucleus[56]. FVs mediate nuclear export of mRNAs through interaction of constitutive transport element with NXF1 and NXT1 transporter proteins and the interaction of a viral mRNA element with kariopherin CRM1 via a virus encoded protein[57]. Besides *gag*, *pol* and *env*, FV genome also contains an internal promoter located in the *env* gene driving basal expression of *tas* (TransActivator of Spumavirus, or *bel-1*) and *bet* accessory genes. While *bet* participates in the inhibition of APOBEC3 cell restriction factors[58], *tas* acts as a transactivator of transcription[59]. Vectorisation of FVs includes the removal of *tas* and *bet* accessory genes and separation of the packaging genes into several helper plasmids[60]. Interestingly, *gag* and *pol* can be provided from separate plamids, as they are translated from different viral RNAs and its relative expression is not regulated by a ribosomal frameshift as in other retroviruses[61].

Alpha-retroviruses mainly differ from lenti- and gamma-retroviruses in their integration preferences. The integration profile of alpha-retroviruses is more neutral and does not favour gene-rich regions, transcription units or TSS, which makes them attractive candidates for gene therapy in the future[62]. However, a model for alpha-retroviral integration or the existence of a tethering protein that modulate their integration profile are yet to be discovered.

Non-HIV lentiviruses such as bovine immunodeficiency virus (BIV), caprine arthritis-encephalitis virus (CEAV), feline immunodeficiency virus (FIV), simian immunodeficiency virus (SIV) and equine infectious anaemia virus (EIAV) have also given rise to lentiviral vector systems. However, only the latter has been used by Oxford Biomedica for the treatment of Parkinson's disease, due to its no-association to any disease in humans[63]. The vector contains three genes encoding for the critical enzymes involved in the production of dopamine (tyrosine hydroxylase, aromatic L-amino acid decarboxylase and guanosine 5'-triphosphate cyclohydrolase 1).

From the molecular biology prespective, non-HIV lentiviruses share most of their features with HIV, with only a few differences. EIAV possesses three accessory proteins (Tat, rev and S2) compared to the seven proteins present in HIV-1. S2 protein contributes to viral replication and its removal reduces the virulence of the infection although its function is poorly understood[64]. Differences between HIV and FIV comprise its cell entry through CD134 receptor instead of CD4[65] and *rev* gene overlaps and shares reading frame with *env* (like BIV and CEAV)[66]. Also, unlike primate lentiviruses, FIV infection is not limited by the presence of tetherin[67]; tetherin binds viral particles to the cell membrane to prevent further infection. Another difference found in FIV compared to primate lentiviruses is the sequence homology between both polypurine tracts[68]. SIV follows a different pathway to enter the nucleus[69] and shares low homology with HIV, reducing the risks of recombination.

Finally, HIV-1 derived lentiviral vectors (derived from lentiviruses and thus retroviruses) were used in this study and thus will be explained in more detail in the Section 1.1.3.

Table 1.1. Characteristics of the main vectors used in gene therapy.

Vector	Gamma-retroviral	Lentiviral	Vaccinia	Adenoviral	AAV	Herpes viral	Naked/plasmid DNA
Nucleic acid form	ssRNA	ssRNA	dsDNA	dsDNA	ssDNA	dsDNA	dsDNA
Size of the particle	100nm	100nm	360 × 270 × 250nm	90-100nm	25nm	120-300nm	10-100nm
Maximum insert size	8-10kb	8-10kb	25kb	35kb*	5kb	152-155kb	Unlimited
Enveloped	Yes	Yes	Yes	No	No	Yes	NA
Titer (TU/mL)	>10^{9**}	>10^{9**}	>10^{9}	>10^{11}	>10^{12}	>10^{12}	No limitation
Transduce non-dividing cells	No	Yes	No	Yes	Yes	Yes	Yes
***In/Ex vivo* applications**	*Ex vivo*	*Ex vivo*	*Ex/in vivo*	*Ex/In vivo*	*Ex/In vivo*	*Ex/In vivo*	*Ex/In vivo*
Integration	Yes	Yes	No	No	No/(Yes***)	No	No
Duration of expression	Long	Long	Short	>1year	<1year	>6 months	Short
Scale up adaptability	Pilot scale up	Not tested	Easy	Easy	Difficult	Difficult	Easy
Immunological problems	Few	Few	Extensive	Extensive	Not known	Few	None
Pre existing host immunity	Unlikely	Unlikely	Yes	Yes	Yes	Yes	No
Main limitation	Insertional mutagenesis	Insertional mutagenesis[1]	Inflammatory response, toxicity	Inflammatory response, toxicity of the capsid	Capacity. Immune response in 2nd dose	Inflammatory response, toxicity	Low efficacy
Main application	PI	PI	Cancer	Cancer	Retina, CNS, liver, muscle	CNS	Various (mainly cancer)
Use in clinical trials	18.4%	5%	7.2%	22.2%	6%	2.9%	17.4%

*** if rep protein present, into AAVS1 site, chr19 q13.4; ** concentrated; * gutless ADV 1 although no genotoxic events seen; 2 integration pattern more random than lentivirus; CNS, central nervous system; PI, primary immunodeficiencies.

1.1.3 Lentiviral vectors

Lentiviral genome and structure

Lentiviral vectors are derived from lentiviruses that belong to the *Retroviridae* family. Retroviruses are enveloped 100nm diameter viruses that possess 2 copies of +ssRNA in their virion form. Out of the 7 genera of retroviruses comprising alpha-, beta- delta-, epsilon-, spuma- gamma- and lenti- (retro)viruses, the latter two (detailed in Table 1.1) have been used for gene therapy. The different genera differ in morphology, integration profile, genomic complexity and structure. Given their use in this study, lentiviral vectors will be covered in detail.

Lentiviruses, like all other retroviruses, are characterised by their replicative strategy, which requires reverse transcription of its viral RNA in order to integrate into the host cell genome as a proviral dsDNA fragment. The virus indefinitely persists in the host and uses the host's expression machinery to express viral genes. Viral particles are budded using the host cell membrane as envelope. The genome of lentiviral virions is composed of 2 molecules of positive strand RNA 7-10kb in length. Unlike gamma-retroviral vectors, lentiviral pre-integration complexes possess the singular feature of being able to actively traverse the nuclear envelope and therefore integrate into the genome of dividing and non-dividing (quiescent) cells[70,71]. Another distinctive characteristic is the high efficiency displayed and the ability to transduce a wider range of cell types.

The genome of a complex retrovirus can be divided into coding and non-coding sequences. The coding sequences contain the group-specific antigen (gag) proteins that form the structural components of the virion (capsid CA, matrix MA, nucleocapsid NC and p6), the protease (PR) that catalyses gag and pol polyprotein cleavage during viral maturation, pol proteins responsible for the viral enzymes (integrase IN and reverse transcriptase RT) and env proteins (surface SU and transmembrane TM) that confer the virus the ability to enter cells. The functions of regulatory proteins encoded by the genes *rev* (regulator of virion) and *tat* (trans-activator of transcription) and accessory proteins (vpr, vpu, vif, nef) are described in more detail the Section 1.1.3 ('Molecular biology of HIV' subsection).

Non-coding sequences include the long terminal repeats, which in turn are subdivided into U3 (unique in 3'), R (repeat) and U5 (unique in 5') regions (named after their exclusivity on one of the ends of the viral RNA). The U3 region (455bp) has enhancer/promoter activity for the expression of viral transcripts[72]. The R region (95bp) acts as a polyadenylation signal (AAUAAA) in the 3'LTR, and coordinates reverse transcription together with the 5'LTR R region. The U5 region (81bp) helps processing the preceding polyadenylation signal[73]. Following 5' to 3' direction, the primer binding site (PBS) is a *cis*-regulatory sequence that binds 18bp in the $\text{tRNA}^{\text{Lys3}}$, which primes extension of the DNA minus strand during reverse transcription[74,75]. The packaging signal (Ψ) allows encapsidation of full-length RNA transcripts into virions[76]. The central polypurine tract (cPPT, located in the IN region of the *pol* gene)[77] and the polypurine tract (located upstream of the 3'LTR) are 16bp-priming regions that enable the extension of the DNA plus strand during reverse transcription[78].

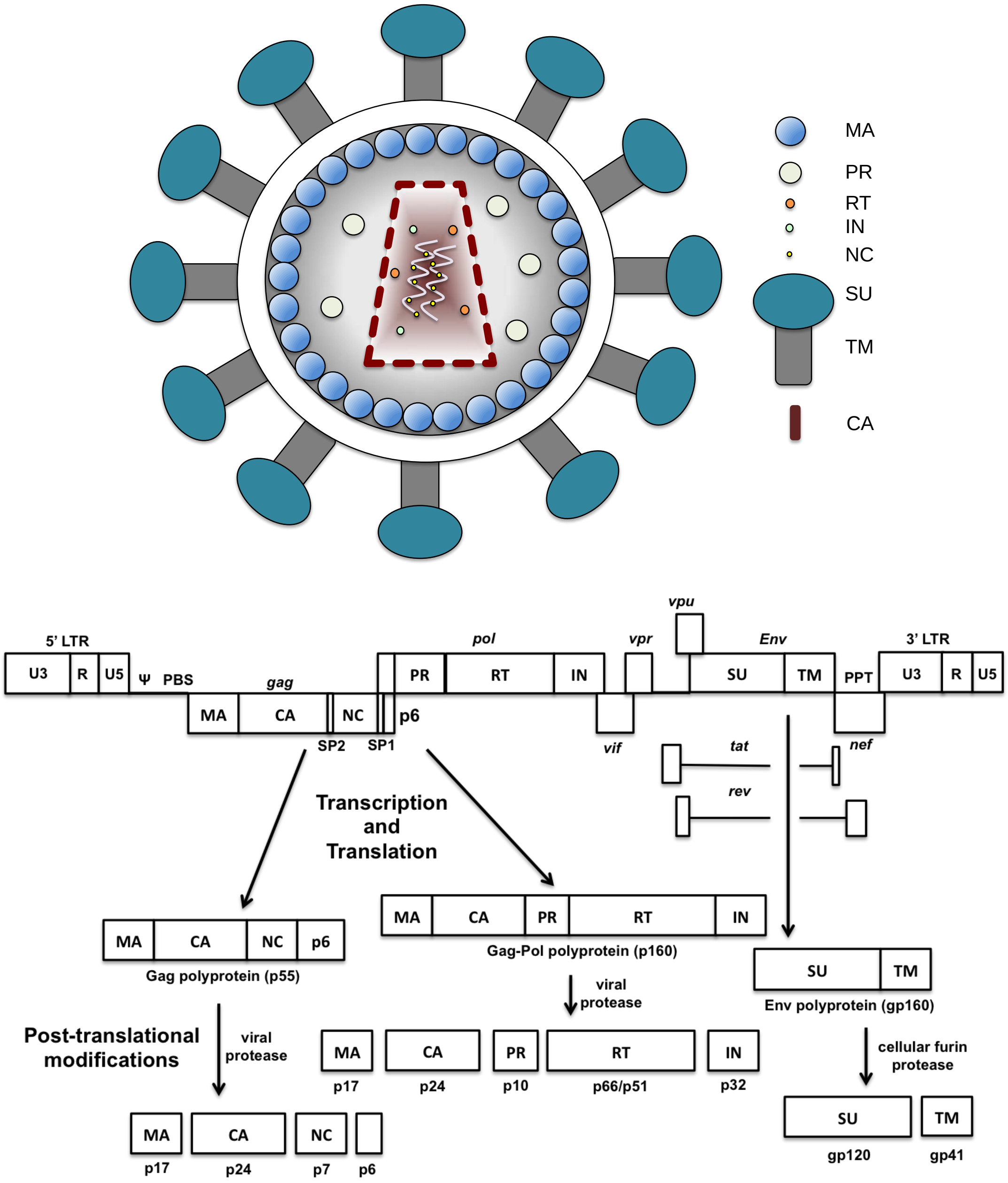


Figure 1.1. Schematic of the structure of a mature HIV-1 virion and expression of lentiviral genes.

Transmembrane glycoprotein (TM); surface glycoprotein (SU); matrix (MA); protease (PR); capsid (CA); nucleocapsid (NC); integrase (IN); reverse transcriptase (RT); long terminal repeat (LTR); unique in 3' (U3); repeat (R); unique in 5' (U5); trans-activator of transcription (tat); regulator of virion (rev); negative regulatory factor (nef); viral protein R (vpr); viral protein U (vpu); group-specific antigen (gag); polymerase (pol); envelope (env); Spacer peptide 1 (SP1); Spacer peptide 2 (SP2).

Molecular biology of HIV-1

Transcription of integrated proviral DNA

The 5'LTR U3 region contains enhancer/promoter sequences that enable RNA polymerase II to drive the expression of viral transcript from the integrated proviral dsDNA. In the absence of Tat, hypophosphorylation of RNApol II[79] results in transcription of short and unpolyadenylated transcripts. However, phosphorylation of the C-terminal of RNA pol II prevents premature transcript termination, enabling the expression of regulatory proteins Tat, Rev, Nef[80]. Trans-activating factor Tat feeds back the phosphorylation of RNA pol II through its binding to transcription factor II H[81]. In turn, Tat binds to the transactivation-responsive (TAR) stem-loop in the RNA and recruits cyclin T1 and Cdk9, which also hyperphosphorylate RNApolII to mediate elongation of HIV mRNA[79]. Over 40 alternative spliced forms of RNA derived from four splice donors and eight splice acceptors are translated giving rise to multiple viral proteins. However, they can be grouped in unspliced 9kb genomic RNA, the 4kb incompletely/partially spliced mRNAs (comprising *vif*, *vpr*, *env/vpu* and *tat* exon 1) and the 1.8kb completely spliced mRNAs (comprising *tat* exon 1-2, *rev* and *nef*). Control over the splicing is finely regulated by the strength of splice donors and acceptors and additionally by intronic and exonic splicing silencers and enhancers. At this point, lentiviral RNA represents about a 1% of the total cellular RNA[82]. However, expression of viral RNA is not detectable if the virus persists latently in a cell reservoir[83].

Latency is a reversibly non-productive state of infection that the virus can undergo after infection. Pre-integration latency allows persistence of unintegrated forms in the cytoplasm of CD4+ T cells for one day although cannot form long-term reservoirs[84]. Post-integration latency consists of a state of reversible blockage of expression of viral genes at a transcriptional level due to several potential mechanisms: chromatin structure at the site of integration, nuclear architecture/chromosomal disposition, transcriptional interference, transcription factors or repressors, concentration of tat, epigenetics[85].

Several regulatory elements mediate the addition of the polyA tail upon recognition of a polyadenylation signal (AAUAAA) located in the R region of the LTR. The U3 and 5' side of *nef* gene contain a upstream enhancer elements that promote recognition by the cleavage/polyadenylation specificity factor[86]. Binding of factors to the secondary structure favour the access to the 3'LTR polyA[87]. U1 snRNP splicing factor binds 5'polyA site precluding its use[88].

Nuclear export of viral RNA

Unspliced mRNA and partially spliced RNA species (4kb transcripts encoding accessory proteins), are exported to the cytoplasm in a process mediated by Rev[89]. Cooperative oligomerisation of Rev protein on the Rev responsive element (RRE) stem–loop present in the env region of spliced mRNA and mediate nuclear export through the nuclear pore complex with the participation of Ran-GTP and Cellular exportin-1 (Crm-1)[90,91] owing to a 10 amino acid nuclear export signal rich in leucines[92]. Once in the cytoplasm, the GTP of the former is hydrolysed to GDP by Ran GAP and Ran BP1, which dissociates the complex and enables return of Rev to the nucleus[93]. Fully-spliced HIV-1 RNAs are not retained in the nucleus and follow the cellular RNA export pathway.

Translation of viral proteins

Unlike cellular mRNAs, translation of HIV-mRNAs is not initiated at the first AUG from the 5' end due to the presence of secondary structures (TAR, PBS, 5' polyA and packaging signal) precluding it. Initiation of translation can take place either via cap- or IRES-dependent mechanism. The cap mechanism requires 12 eIF and the interaction of the eIF2 GTP Met-tRNA$^{Met}{}_{i}$ ternary complex with the complex formed by the 40S ribosome and 3 eukaryotic initiation factors (eIFs)[94]. The IRES mechanism relies on the secondary structure of these sequences located in the 5'UTR and in the *gag* coding region but also on their interaction with the 40S subunit and IRES *trans*-acting factors[95]. However, the mechanism of translation initiation through IRES is not completely understood[96].

The *env* gene is translated from spliced mRNA in the rough endoplasmic reticulum and glycosylated to gp160 glycoprotein in order not to be recognised by neutralising antibodies[97]. Unspliced *gag* mRNA is translated by polyribosomes into Gag p55 polyprotein. *Gag-pol* is also translated from unspliced mRNA and also contains reverse transcriptase (RT or p51/p66). Gag-pol polyprotein is translated from the same unspliced mRNA as gag at a lower frequency. A slippery hexanucleotide followed by a stem-loop pseudoknot at the end of NC protein ORF[98] increases the chances of frameshift to 5-10%[99]. The consequent 1:20 *gag:gag-pol* ratio is critical for infectivity[100]. As a result the ribosome reads through the gag stop codon and translation of gag-pol polyprotein containing the viral enzymes protease (PR or p10), integrase (IN or p32), and reverse transcriptase occurs.

Accessory proteins are also synthesized from spliced mRNA and, among other functions, participate in the neutralisation of host restriction factors. The dimerization domain of viral infectivity factor (vif) is involved in neutralisation of the APOBEC3G host factor[101]. The function of this host factor is to prevent retroviral replication by deaminating cytidine residues during reverse transcription, which causes detrimental mutations in the proviral genome. Similarly, in HIV-2 and SIV, vpx counteracts the action of SAMHD1 restriction factor[102]. The dNTPase activity of SAMHD1 protects the cells from viral infection by reducing the levels of nucleotides in order to inhibit reverse transcription. Viral protein R (vpr) has been associated with multiple functionalities involving facilitation of reverse transcription, nuclear import of HIV-1 PIC, transcription and has also been reported as toxic inducing cellular apoptosis[103]. Viral protein U (vpu) and negative regulatory factor (nef) downregulate the expression of CD4 cell receptor preventing the interaction between premature envelope protein and its receptor[104,105]. Vpu also interacts with tetherin host restriction factor mediating its degradation[106].

Assembly, maturation and budding

Assembly of immature viral particles lasts approximately 10 minutes[107] and is mainly mediated by polyprotein Gag, which in turn is cleaved to give rise to Matrix

(MA or p17), Capsid (CA or p24), Nucleocapsid (NC or p7, separated by p2 and p1 spacer peptides) and p6 proteins[108]. MA contributes to the assembly via myristoylation of Gag N-terminus, which in addition to its membrane binding domain triggers its targeting to the cell membrane[109]. In addition, MA also recruits Env protein[110]. MA and CA mediate protein interaction and virion stabilisation[111]. Highly basic NC recognises the 4–hairpin secondary structure of the packaging signal (Ψ) and allows packaging of two molecules of polyadenylated and capped full size ssRNA virion genome together with the LysRS/tRNALys packaging complex. LysRS interacts with C-terminal domain of Gag protein and IN-RT-PR polyprotein, which also interacts with tRNALys3. C-terminal domains of different Gag and Gag-pol proteins also interact. The stoichiometry within the tRNALys packaging complex is 12:4:1:8 Gag:gag-pol: LysRS:tRNA^{Lys3} [112,113]. Protein p6 also participates in virion budding by interacting with IN-RT-PR polyprotein and mediating incorporation of viral accessory proteins like vpr[114]. Spacer peptides SP1 and SP2 help accommodating the conformational changes occurring during these processes[115]. Mature transmembrane Env protein is glycosylated in the Golgi and transported to the membrane through the cellular secretory pathway. During this process cleavage of the gp160 polyprotein in the Golgi by furin-like proteases gives rise to the surface (SU, gp120) and transmembrane (TM, gp41) mature glycoproteins[116]. SU and TM are then delivered to the cell membrane to become part of the viral particles. Maturation does not conclude prior to viral budding. When the viral particle is budded, dimerisation of the protease causes its activation through a mechanism that is not well understood. As a consequence, the nine peptide bonds contained in the gag and gag-pol polyproteins are cleaved by the viral protease (PR) to give rise to CA, MA, NC, p6 and PR, RT and IN mature enzymes. Polyprotein proteolysis is necessary to yield mature infective virions.

Virion budding is mediated by the host endosomal-sorting complex required for transport (ESCRT) machinery. Briefly, p6 binds to ALIX and TSG101 proteins, which trigger recruitment of VPS4 ATPases and ESCTR-III complexes creating a ‘dome’ that induces a fission of the membrane neck[110].

Viral entry

Following non-specific interactions with negatively charged heparan sulfate proteoglycans and more specific interactions with α4β7 integrin[117] or DC-SIGN[118], viral entry is mediated through the specific interaction between the viral envelope protein and the human receptor CD4 in T lymphocytes and macrophages. The viral envelope protein is composed of three transmembrane (TM) gp41 non-covalently bound to trimeric forms of gp120 surface (SU) glycoproteins. Interaction between CD4 receptor and gp120 SU glycoprotein induces conformational changes in gp41 and gp120, which allow binding of the gp120 variable loop 3 with CCR5 (predominantly found in macrophages) or CXCR4 (mainly in hematopoietic progenitor cells) chemokine co-receptors binding site[119,120]. Such interaction triggers rearrangement within gp41 resulting in insertion of a fusion peptide into the target cell membrane and the formation of a six-helix bundle and ultimately the fusion of the virion and host cell membranes[121].

Uncoating and cytosolic transport

HIV-1 core uncoating occurs in the cell cytoplasm between viral entry and the nuclear import of reverse transcribed dsDNA and other proteins of the PIC. The stages and mechanism of the uncoating process is not well understood. Three models have been proposed to explain the CA disassembly during the uncoating process. The early immediate uncoating model supports complete CA disassembly right after membrane fusion and virion entry and the migration through the nucleopore of a reverse transcription complex (RTC) devoid of CA[122]. The RTC is a transitory structure consisting of MA, CA, NC, IN, Vpr and RT whose function is to enable reverse transcription of the viral RNA in the cytoplasm prior to transition into pre-integration complex (PIC) for nuclear import and proviral integration. The cytoplasmic uncoating model suggests a progressive process where CA is removed in the cytoplasm and highlights the role of remaining CA associated with the RTC mediating interaction with host factors and nucleopore complexes[123]. The nucleopore uncoating model suggests complete CA removal in the nucleopore complex, protecting the RTC from cytosolic DNA sensing pathway proteins[124,125]. Experimental evidence of CA associated with PIC in the nucleus reinforces the two latter models[123,126].

Viral proteins also interact host proteins during the uncoating process; for example, peptidyl-prolyl isomerase Cyclophilin A (CypA) binds to CA and protects the viral DNA genome enhancing its replication[127]. The transport of RTC/PIC is mediated by the interaction of viral proteins with actin microfilaments and microtubules. Diffusion of DNA macromolecules is a slow and random process due to the steric hindrance occurring between the molecules present in the cytoplasm[128]. Transport of RTC/PIC structures is actively mediated through actin filaments (using myosin VI complexes[129]) and the microtubules tubuline[130] (using dynein complexes, mechanism also observed in other viruses)[131,132] from the cell membrane and to the nuclear envelope

Reverse transcription

Reverse transcription constitutes one of the defining features of retroviruses and is an essential step in viral replication. This process is carried out by the two subunits of the reverse transcriptase heterodimer: p66 subunit, which contains two domains that play the catalytic roles and the p51 subunit whose four subdomains have structural function. Within the p66 subunit, RNAse H domain, that allows degradation of RNA in DNA-RNA hybrids and DNA polymerase domain allows DNA synthesis and elongation from RNA and DNA templates[133]. Within the 50 molecules of RT present in a virion, RT with defective p66 or p51 activity can complement each other and restore the infective phenotype, indicating that they interact in a cooperative manner[134].

During uncoating MA, NC, vpr, RT and IN remain associated with the tRNA^{Lys3} and the viral genome RNA in the RTC while reverse transcription takes place[135]. Reverse transcription of lentiviral vectors is initiated when the host-derived tRNA^{Lys3} serves as a primer and hybridises the primer binding site (PBS). The choice of host tRNA can vary among retroviruses. DNA synthesis creates a short (-) single strand of DNA from the PBS in the 5'LTR upstream to the 5' R[136]. This DNA fragment (tRNA^{Lys3}-U5-R DNA) undergoes strand transfer and primes the R regions in the 3'LTR for the synthesis of the remaining a full-length (-) genome single strand of DNA up to the 5' end of the viral RNA genome (until the PBS, included). RNase H activity of the RT cleaves the RNA from the DNA–RNA hybrid except a purine-rich PPT[137], which primes initiation of synthesis of the DNA (+) strand. Synthesis of the DNA (+) strand commences towards the 3'LTR using the (-) strand as a template. Formation of cPPT DNA-RNA and tRNA-DNA duplexes allow the degradation by RNAse H activity[138]. Then, (+) strand transfer and hybridisation of primer binding sites from both strands allow bidirectional elongation of both DNA strands ending with U3-R-U5 regions on both LTRs[139]. Renda *et al.*, demonstrated that methylation of the adenosine58 residue in tRNA^{Lys3} (1-methyladenosine 58, m(1)A58) is required to stop elongation of (+) strand during reverse transcription[140]. The reverse transcription takes place in the cytoplasm and is completed when all RNA has been reverse transcribed to dsDNA and the RTC gives rise to pre-integration complexes (PIC). PICs are integration competent complexes in vitro [140,141].

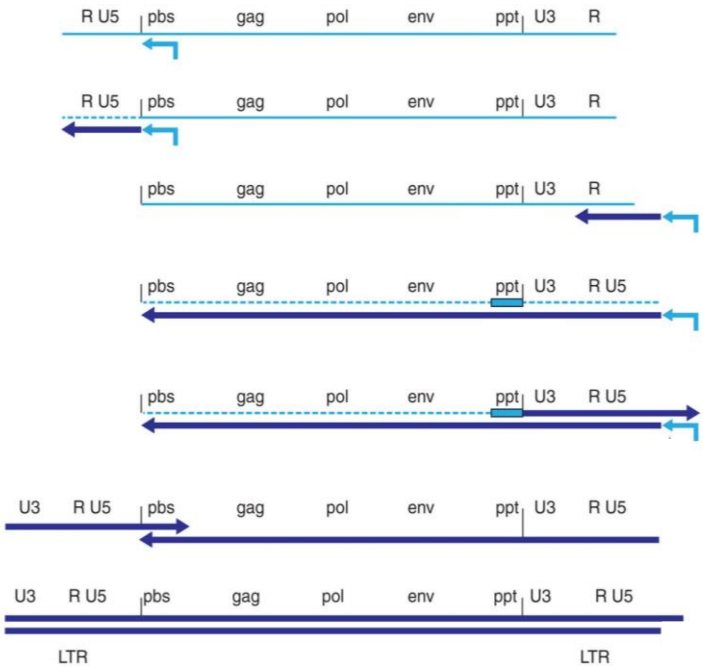


Figure 1.2. Schematics of the retroviral reverse transcription process.
Dark blue line represents DNA and light blue represent RNA. Dashed light blue line represents degradation of the RNA strand by RNAse H. primer binding site (PBS); long terminal repeat (LTR); unique in 3' (U3); repeat (R); unique in 5' (U5); group-specific antigen (gag); polypurine tract (ppt); polymerase (pol); envelope (env). Figure extracted from Hu and Hugher 2012[135].

Nuclear import

Unlike γ-retroviruses, which require nuclear envelope dismantlement to access the nucleus, lentiviral vector PIC can actively enter the nuclear envelope through nucleopore complexes and thus can transduce dividing and non-dividing cells. The ability of the PIC to traverse the nuclear envelope barrier was thought to be due to the IN (through interaction with Nup153 and their nuclear localisation sequences, NLS)[141,142], Vpr and MA (through two weak NLS)[143,144] and the cPPT. However, depletion of cPPT[145], Vpr (in the third generation lentiviral vectors[146]) and IN and MA nuclear localisation signals proved their role is not essential in nuclear entry[147]. Dispensability of karyophilic components is thus controversial since the virus is still able to replicate but nuclear import kinetics are considerably affected by the removal of some of these components[148]. Instead, CA protein has been shown to play a crucial role in nuclear import[149]; MLV CA/HIV-1 chimeras showed significantly impaired nuclear import. However, although it is known that TNPO3[150] is involved, the mechanism by which this occurs is not well understood[151].

Several cellular proteins have also been shown to be implicated in HIV-1 nuclear import such as α2 Rch1[152], importin 7[153] and transportin SR-2 (also called TNPO3)[154] and several nucleoporines. Nucleoporine 358 (Nup358 358kDa, also called RanBP2), Nup153 and Nup98 mediate uncoating and active transport of the PIC into the nucleus[155]. Nup358 has also been implicated with target site preference[156,157].

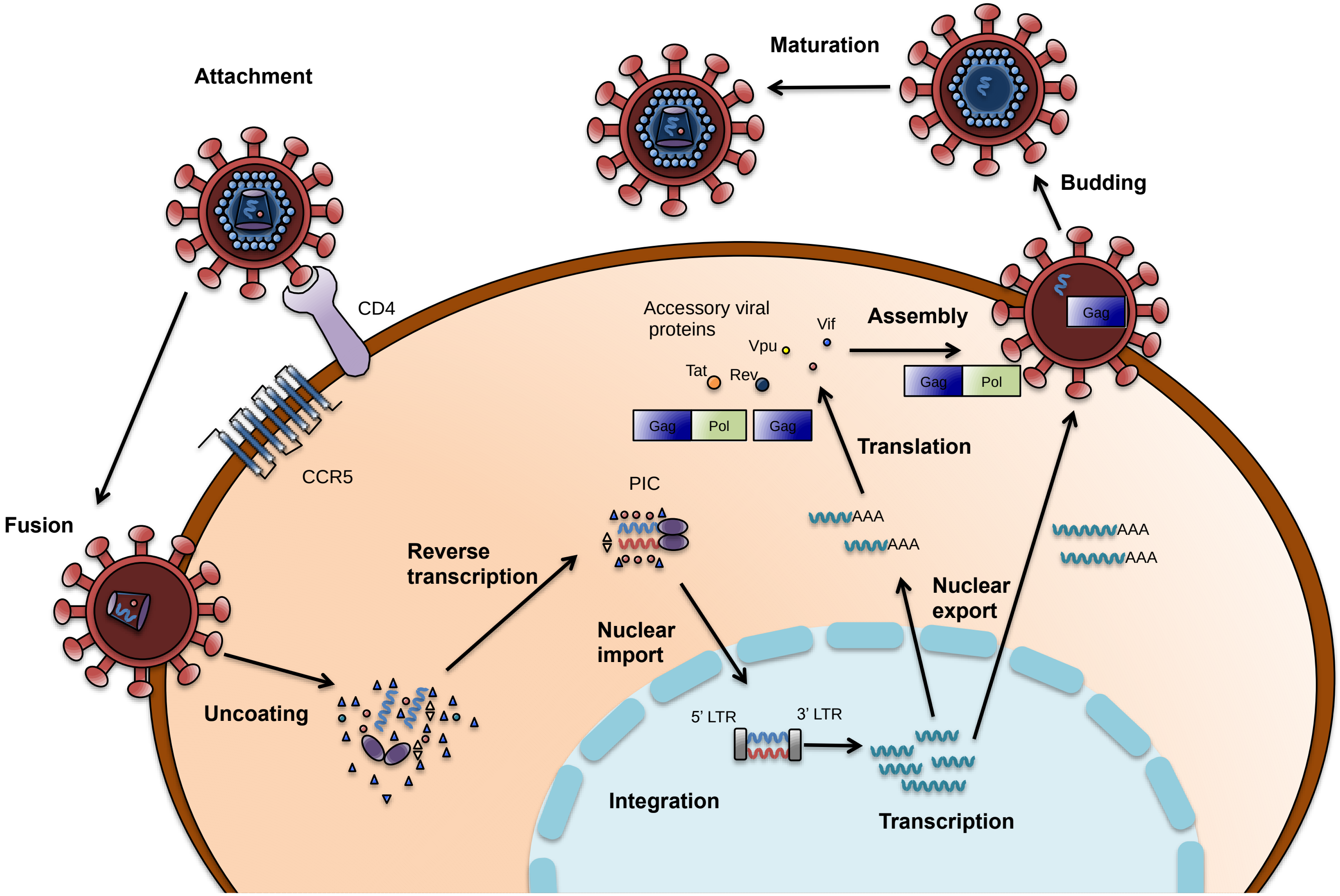


Figure 1.3. Stages of the HIV-1 life cycle.

CD4, cluster of differentiation 4; CCR5, chemokine receptor type 5; PIC, Pre-integration complex; LTR, long terminal repeat; rev, regulator of virion; nef, negative regulatory factor; vpr, viral protein R; vpu, viral protein U; gag, group-specific antigen; pol, polymerase. Modified from Engelman and Cherepanov 2012[158].

1.1.4 Lentiviral integration

Proviral integration constitutes another essential step in retroviral replication. In this step, linear proviral double stranded DNA is covalently inserted into the host chromosomal DNA by the viral integrase. Viral DNA serves as a template to generate either mRNA and subsequently viral proteins or unspliced genomic RNA for future virions.

Integrase protein

Viral integration is catalysed by the integrase in a complex process that involves viral and host proteins. Integrase (IN) was first described as a non-specific endonuclease in avian myeloblastosis virus[159]. Its function was confirmed when knock-out of certain residues blocked viral replication[160]. HIV integrase (IN) is a 32kDa non-specific endonuclease protein whose sequence shares homology with the avian sarcoma virus (ASV) reverse transcriptase[161] but also RNAse H and RuvC resolvase and Mu transposase[162]. This protein is encoded and translated from the *gag-pol* open reading frame and processed by the viral protease (PR) to its mature form. The 50 amino acid N-terminal is composed of 3 alpha-helixes with a HHCC motif that when coordinated with zinc acts as a DNA binding domain[163,164]. The 160-amino acid central core domain is composed of a mixed alpha-helix and β-sheet[165] and contains the catalytic function. Extensive mutational and substitution analysis performed on the catalytic core revealed that the D64, D116 and E152 triad (in HIV) is functionally critical[166,167]. The 80-amino acid C-terminal domain possesses a nuclear localization signals and a 5-stranded β-barrel conforming a SH3-like domain whose function is DNA binding[168,169]. Experiments combining truncated C- or N- terminal domains imply that all three domains are likely to form dimers independently and that IN acts as a dimer or a higher order multimer[170,171]. Crystallization of the FV IN revealed the interaction between the N-terminal and the catalytic core of monomers forming a dimer and these in turn homotetramer structures[172].

Together with the retroviral DNA, viral proteins including matrix (MA)[173], reverse transcriptase (RT)[174] nucleocapsid (NC)[175] and cellular proteins like barrier to autointegration factor (BAF)[176] and high-mobility-group (HMG-I(Y))[177], the integrase protein (IN) form the pre-integration complex (PIC). HMGs are DNA-binding proteins that contribute to lentiviral integration in many ways: recognising DNA secondary structure, binding to the minor grove, interacting with supercoiled, bent and non-B-DNA structures and modulating chromatin structure. Pre-integration complexes (which can be isolated from infected cells[174,178,179]) are able to catalyse integration *in vitro* in the presence of Mg^{+2} and target DNA[180–185]. PIC is guided to the nucleus due to the nuclear leading sequences (NLS), present in MA protein, which together with Vpr and CA, are involved in lentiviral nuclear import via the nucleopore and integrate into non-dividing cells[186–188]. Other proteins such as the transcriptional co-activator lens epithelium derived growth factor (LEDGF or p75) also play a role as co-factors of the viral integrase and participate in the target site selection of the HIV pre-integration complex.

The integration mechanism

The integration reaction takes place in two catalytic steps. In the first step, referred to as end processing, IN specifically recognises sequences on both ends of the long terminal repeats (LTR)[179,189–191]. Crosslinking and nucleotide substitution experiments have confirmed specific interactions between the IN and the LTR termini sequence[182,192]. This way, HIV-1 IN recognises 20bp[182] on the U5 and U3 ends, murine leukaemia virus (MLV) integrase recognizes 11-12 bp and ASV recognises 15bp[181]. Modification of these recognition sequence was also found to affect integration efficiency[193]. Next, the integrase cleaves a highly-conserved 5'-GT-3' dinucleotide resulting in a 5'-CA-3' overhang (5' protruding) on both ends of the viral DNA[194,195] in an ATP-independent process[181]. In the strand transfer step, the exposed oxygens of the hydroxyl group in the resected 3' viral ends attack the phosphodiester bonds on the host cell target DNA[196,197]. The integration reaction concludes when the host DNA repair machinery fills the gaps between the recessed LTR termini and the host DNA[198,199]. As a result, five nucleotides at the integration site are duplicated on both ends of the integration. Circular intermediates containing 1-LTR or 2-LTR, formed as dead-end product of reverse transcription reaction, are not integration competent[195]. 2-LTR forms are generated as result of recircularisation events[174] in a process that involves the non-homologous end joining (NHEJ) machinery[200]. 1-LTR forms are derived from incomplete reverse transcription[201] or recombination involving the MRN complex[202] . Additionally, if not prevented by the host cell protein BAF[176,203], HIV can target itself (autointegration) yielding a smaller truncated autointegrant (if targeting the same DNA strand) or a 2(inverted)-LTR internally arranged circle (if targeting a different one)[195].

Integration preferences of HIV and other retroviruses

Lentiviral vector integration preferences are not random but a complex puzzle for which an integral and accurate model has not been described. Integration site selection is partially determined and modulated by a combination of interrelated factors including accessibility to chromatin[204–206], nuclear disposition of chromosomes[207,208], tethering proteins[209–211], topological features and the primary sequence via the integrase viral protein[212].

Initial thoughts on retroviral integration suggested it may be driven by a particular nucleotide sequence. Withers-Ward *et al.,* revealed no strong correlation between integration pattern and host's DNA primary sequences[213]. However, re-examination by Shih *et al.,* described the existence of integration hotspots with specific base pairs[214]. Integration downstream the local 5'-TpN-3' pattern[215] and 5' G/C residues[159] were found more predominant than random for HIV and Avian sarcoma leukosis virus (ASLV), respectively. More recently, integration base-patterns [-3]TDG(int)GTWACCHA[7] for HIV, [–4]DNST(int)VVTRBSAV[7] for MLV and [–4]STNN(int)SNNNNSNAAS[9] for ASLV were observed in a 2-fold higher-than-expected frequency. However, the absence of a strong consensus sequence indicates that sequence recognition is not a strict requirement and only modulates structural features in the integration site selection process[216]. Katz *et al.,* demonstrated that sequences containing inverted repeats were also shown to enhance integration through the formation of a hairpin loop[184]. Loop-forming oligonucleotides were also shown to be preferred as integration targets *in vitro*[217].

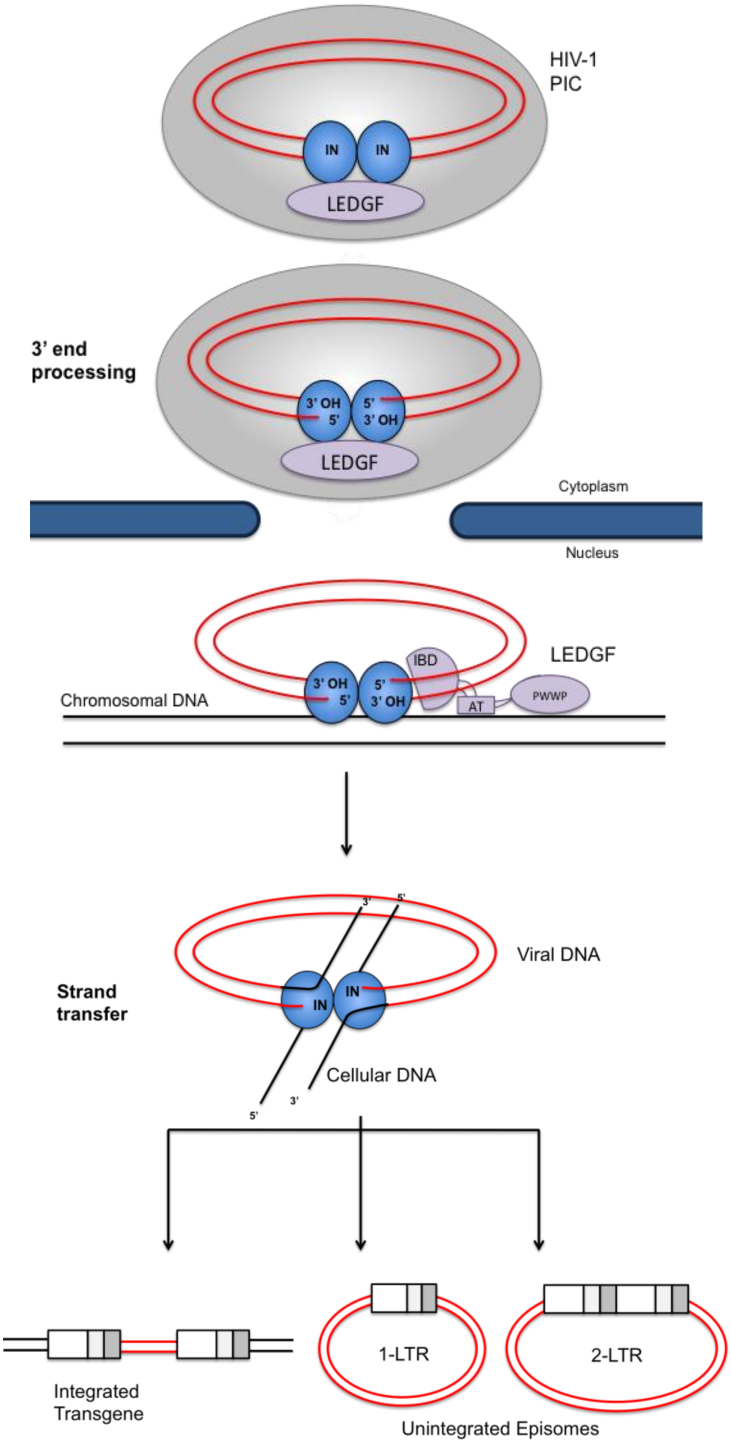


Figure 1.4. Schematics of lentiviral integration.

LTR, Long Terminal Repeat; LEDGF, lens epithelium-derived growth factor; IN, integrase; PWWP, Pro-Trp-Trp-Pro domain; AT, A/T (adenine/thymidine) DNA hook; IDB, Integrase binding domain.

The idea that retroviral vectors preferentially integrate into open euchromatin suggested nucleosomes may exert steric hindrance preventing retroviral integration[218–221]. However, early experiments on retroviral integration performed with isolated PICs *in vitro* revealed integration on DNA wrapped around the histone octameric core is preferred over naked DNA[215,222–224], possibly because deformation is needed to complete the strand transfer step[172]. Lentiviral integration was found to be favoured in positions presenting high levels of distortion in their chromatin structure[215]. Integration frequency seem to follow 10.4bp periodic windows in A/T-rich regions[221] corresponding to the number of bases per turn whose phosphates are outwardly exposed in the major groove in nucleosomal DNA *in vitro*[215,223] (confirmed also in lentiviruses in recent genome-wide analysis[225,226]). This link between DNA sequence and chromatin structure may also explain diminished GC content in lentiviral target sequences over short intervals (<2kb)[227]. Topological preferences correlate with local chromatin structure on lentiviral integration targeting. HIV-1 integration was found to be disfavoured in centromeric and telomeric regions due to the high degree of condensation and poor accessibility of constitutive heterochromatin[228] (same occurs with facultative heterochromatin).

With the publication of the human genome[229] and the arrival of next-generation sequencing technologies, integration site selection was analysed at a higher throughput level leading to more representative conclusions. Murine leukaemia virus (MLV, as a representative of a *gamma-retroviridae* family) was found to integrate within gene enriched regions, highly/actively transcribed genes (transcribed by RNA pol-II or protein coding genes) often collocated with CpG island-dense regions and is 5 times more likely to integrate near (<2kb) RefSeq transcription start sites[230–234] showing only a moderate preference for transcription units (with respect to HIV). Xenotropic murine leukaemia virus showed the same integration features[235,236]. These results correlate with their over representation around annotated CpG islands, conserved transcription factors binding sites and non-coding sequences[52,237].

A genome-wide analysis analyzing 8,250 unique integration sites of avian sarcoma-leukosis virus (ASLV) against the UCSC hg19 genome confirmed the results of previous studies performed at a lower scale[231,238]. ASLV (as a representative of the alpha-retroviridae family) presents a relatively neutral pattern with only a very modest predisposition towards genes and promoter regions when compared to SIN-MLV and SIN-HIV-derived vectors[239]. ASLV integration near repetitive or satellite elements was close to random[239].

HIV-1 (and lentiviral vectors in general) instead tend to integrate within active transcription units in areas with high gene density (60-70% according to Kursun *et al.,*[240]), collocated with CpG islands, regions, DNAseI cleavage sites and G/C-rich regions[241]. Unlike gamma-retroviruses, more likely to be integrated near TSS, lentiviral integration events are evenly distributed over the length the gene[231,234,242–250]. Within genes, more than 90% of the integrations take place within introns due to their relative longer proportion.

Early studies on retroviral integration did not report a preferential integration near transcription units or repetitive elements[251,252]. Leclercq *et al.,* did not observe this integration pattern in human T-lymphotropic virus (HTLV)[253]. In the same line, Weidhaas *et al.,* suggested high levels of transcription might interfere with avian leukosis virus (ALV) integration[254]. However, the representation of some of these studies was questioned due to the limited number of integration events and the fact the human sequence was not available yet and conclusions were drawn from a few model genes. In recent studies, MLV integrations are under represented around repetitive regions (e.g. LINEs) with the exception of SINEs frequently located in transcribed regions and contain PolII promoters[255,256]. Satellite elements were clearly underrepresented.

As shown before, a number of studies have presented divergences between integration profiles depending on genera. Experiments examining integration preferences of chimeric HIV-1 containing the MLV integrase showed a gamma-retroviral-like integration pattern[245], narrowing down the determinant factor to the viral integrase and specific tethering factors.

Tethering proteins and other factors influencing the retroviral integration profile

Contrary to early studies showing that DNA binding proteins blocked the access of pre-integration complex to the target DNA, host cellular proteins have been demonstrated to play an important role in target site selection. As previously mentioned, the lens epithelium-derived growth factor (LEDGF, also known as p75 protein encoded by the PC4 and SFRS1 interacting protein 1 gene, *PSIP 1*) binds the lentiviral IN and strongly influences the targeting of the proviral DNA. LEDGF/p75 was reported to bind lentiviral IN through its integrase binding domain (IBD), whose structure was published by Cherepanov *et* al.,[257]. LEDGF/p75 prevents IN degradation by the proteasome[258] and also participates in its nuclear import and localisation via the interaction of its integrase binding domain (IBD)[259] with the N-terminal and core domains of the integrase; its knock-down prevented their colocalisation in the nucleus[260]. Integrases of other viruses such as MLV, RSV, HTLV did not show this interaction indicating it is specific to lentiviruses[261]. LEDGF/p75 knock-down showed a decrease in infectivity[262], which identified the LEDGF/p75 as a candidate target for integrase inhibitors[263]. LEDGF/p75 knock-down did not present a biased affinity towards transcription units and instead, integrations were more distributed among CpG islands and promoter regions identifying 'LEDGF/p75 islands'[243,264]. The tethering model supports that LEDGF/p75 modulates the integration preferences via the interaction of its C-terminal domain with the viral IN and secondly via the interaction between two AT DNA hooks (+NLS) with chromatin. In addition, the N-terminal of the LEDGF PWWP domain has been shown to interact with chromatin. Although this interaction is key for viral replication, fusion proteins consisting of LEDGF and LANA or H1 proteins instead of the PWWP domain have shown to enable replication[265]. If theAT hooks and PWWP domains of LEDGF are replaced with chromatin binding domains (CBDs), viral replication can still be supported and the integration patterns are redirected to those of the CBDs[248].

However, the mechanism of interaction between the PWWP domain/AT-hooks and the chromatin is poorly understood[266]. In 2013, Eidahl *et al.*, published a solution structure of LEDGF PWWP with a peptide of H3K36me3 and examined its binding with DNA by nuclear magnetic resonance[267].

Therefore, retroviral vector integration distribution pattern is also consistent with their epigenetic marks in different cell types[52]. MLV-derived vectors are more frequently integrated in regions with methylation marks associated with active RNA PolII promoters and enhancers (H3K4me1 and me3)[241]. HIV-derived integrations are often found close to epigenetic markers associated with transcribed gene bodies (H3K36me3)[237,268]. SIN-ASLV shows poor correlation with epigenetic markers found in transcribed regions and slightly higher than random close to regulatory regions. When histone modification mediates gene silencing through chromatin conformation (for example via trimethylation of H3K9me2 and me3 and HP1 protein[269] or H3K27me2 and me3 marks) the integration frequency drops in all retroviruses. Histone acetylation (commonly linked to active gene expression) located near to TSS is less frequent along the HIV targeting sites and vice versa[241]. Acetylation markers associated with TSS and proto-oncogene rich regions are preferentially associated with bromodomain and extraterminal (BET) proteins in MLV[241]. The BET proteins (Brd2, -3, -4) were recently identified to play an equivalent role as cellular binding partners targeting MLV integrase at the TSS[270,271].

To a lesser extent, the Ini1 protein (integrase-interacting protein 1), part of the SWI/SNF ATP-dependent chromatin remodelling complex, was found to enhance HIV-1 integration by tethering integration machinery to specific DNA sites and increasing the catalytic activity of IN[272]. Miller and Bushman hypothesized the recruitment of this chromatin remodelling complex could facilitate chromatin accessibility and thus subsequent retroviral integration[273].

The viral protein CA may also play a role in integration site targeting of PICs through its interaction the cellular splicing factor CPSF6. Variations in proteins that interact with CA such as cyclophillin or N74D CA mutant affect its binding to

CPSF6 and alters integration preferences although the underlying mechanisms of this and the involvement of other cell factors needs to be clarified[155].

Cell type might also be involved in lentiviral integration site selection, probably through one of the previously described factors. In Jurkat cells, Lewinski *et al.,* showed favoured integration in alphoid repeats typically found in gene deserts of centromeric heterochromatin (although its expression is repressed)[274]. A potential explanation could be satellite DNA contains the conserved motif (TGGAA)n which attracts nuclear proteins. In general, differential exposure of chromatin and the presence of remodelling complexes affect acceptor site selection. In post-mitotic/non-dividing/quiescent/resting retinal and neuronal cell types, Brady *et al.,* and Bartholomae *et al.,* reported 30% reduced frequency of lentiviral targeting into transcription units[227,275]. Concordant histone modifications and decreased levels of associated features such as CpG islands, GC content and DNAseI hypersensitive sites corroborate that pattern[227]. Marshal *et al*., showed that there is a correlation between the levels of integration close to transcription units and the endogenous level of tethering protein LEDGF/p75[243] of different cell lines. However, while LEDGF/p75 is the driving force in dividing cells and its knock-down leads to a shift towards regions rich in TSS and CpG islands, in quiescent cells its effect is less pronounced. In LEDGF/p75 knocked-down quiescent cells, a decrease in the targeting of transcription units was described but integration near TSS and CpG islands was not enhanced[276]. This suggested that small traces of LEDGF/p75 are sufficient to the tethering effect or that other factors can also influence the integration preferences. Moreover, in their study, *in vivo* delivery of LVV into brain and eye cells resulted in integration in non-expressed genes indicating that target tissue (and possibly the delivery method) tune the integration profile.

1.1.5 Development of lentiviral vectors

Lentiviral vectors are derived from HIV-1. However, a range of gene delivery vectors have also been derived from other lentiviruses such as the HIV-2[277], the simian[278], feline[279], bovine[280] immunodeficiency or the caprine arthritis-encephalitis virus[281] and equine infectious anaemia virus (EIAV)[282] the latter of which has reached clinical trial for the treatment of Parkinson's disease (NCT00627588 and NCT01856439).

The generation of lentiviral vectors is based on the separation of the viral genome into the transfer vector (provided in *cis*) and genes encoding structural and enzymatic functions (provided in *trans*) in order to reduce the risk of generating replication competent lentivirus (RCL) in target/infected cells. While *cis*-acting sequences comprising the LTR, RRE, splice donor and acceptor, Ψ, PBS, PPT and cPPT have been maintained in the transfer vector, packaging genes (*gag-pol*, *rev*, *env*) have been iteratively separated from the viral genome into several helper plasmids or replaced with the objective of producing safer retroviral vectors. The vectors are replication deficient (only one round of transduction is observed) and the likelihood of generating a replication competent retrovirus (or lentivirus, RCR/RCL) from the transduction of target cells is significantly reduced[146].

In the first generation of lentiviral vectors[70], the viral components were split in three plasmids containing (i) the packaging genes (*gag-pol*) and genes coding for accessory (*vif*, *vpu*, *vpr* and *nef*) and regulatory (*tat, rev*) proteins (ii) the transfer vector and the packaging signal flanked by long terminal repeats and (iii) the replacement of the endogenous envelope *env* protein with protein G of the vesicular stomatitis virus, that has a wider tropism[283]. Originally, endogenous gp41/120 envelope protein limited HIV-derived vector tropism to CD4 expressing cells. Other less toxic pseudotypes such as the feline endogenous virus RD114, cocal and Gibon ape leukaemia virus (GALV) envelope glycoproteins or chimeras using the cytoplasmic domain of the amphotrophic MLV-A glycoprotein (RD114/TR, GALV/TR) have shown efficient transduction of CD34+ cells. However, their titers are lower than VSV-G and they are not established for GMP production. Ross River virus (RRV) and Semliki Forest virus (SFV) envelope glycoproteins also reported lower titers[284]. The development of other pseudotypes such as and Venezuelan Equine Encephalitis and Rabies virus envelope widened the tropism to neuronal tissues[285].

Second-generation lentiviral vectors eliminated viral accessory genes *vpr, vif, vpu, nef* from the packaging plasmid. The removal of accessory genes whose functions are not essential for early phases of vector transduction was found not to be detrimental for transduction efficiency[286].

However, the potential formation of replication competent retroviruses (RCR or lentiviruses, RCL)[146] still raised safety concerns about the use of LVV for clinical applications. Sequences as short as 8 nucleotides have been shown to increase the chances of recombination leading to RCR events[287]. A third generation lentiviral vectors, devoid of *tat* and providing of *rev* gene in *trans* as a fourth plasmid, was developed in order to further split the vector genome and reduce the chances of RCR formation[146]. As a consequence of the removal of *tat*, a promoter upstream of the 5'LTR driving the expression of the vector is necessary. Typically, the Rous Sarcoma Virus (RSV)[146] or the cytomegalovirus (CMV)[288] promoters are used in this chimeric promoter configuration (RSV- or CMV–HIV 5'LTR) also known as pRRL or pCCL, respectively.

In order to reduce the risk of proto-oncogene activation due to the positional effects of insertional mutagenesis caused by the 3'LTR U3 promoter/enhancer activity[289] observed in gene therapy clinical trials in the early 2000s a vector modification was implemented. Self-inactivating (SIN) vectors have reduced the LTR enhancer/promoter sequences resulting from the deletion of a 399bp DNA fragment (including the TATA box) in the 3´LTR U3 region (-418 to -18 nucleotides relative to the beginning of the R region, namely SIN-18). Other SIN vectors have been developed by deleting different fragments of the 3'LTR U3[289,290]. The deletion of LTR promoter and enhancer sequences encoded in the 3'LTR is transferred to the 5'LTR and requires the presence of an external promoter, which is lost after the first round of infection. Additional elements such as enhancers[291] of inducible elements such as the tetracycline 7tetO sequence[292] can also be added giving rise to conditional SIN lentiviral vectors (cSIN LVV). Additionally, this modification reduced the percentage of homology between the transfer vector and the viral genome down to 10%.

External/heterologous elements have also been useful in order to improve gene expression and titers at different levels. Although not essential, reintroduction of a central poly purine tract/central termination sequence (cPPT/CTS) is thought to improve nuclear import dynamics and improve lentiviral production[147]. This element has also been suggested to enhance reverse transcription as mutations in its sequence did not abolish nuclear import but decreased virus infectivity[293].

Post-transcriptional regulatory elements (PRE) are intronless sequences located downstream of the *env* gene -firstly described in hepatitis B virus (HBV)[294]- that have been proposed as an alternative to introns to enhance transcript stabilization and gene expression. Such effect is thought to be due to *cis*-acting regulatory element that actively enables RNA export and cytoplasmic localization regardless of its relative position (but not its orientation)[295]. However, Higashimoto *et al*., showed that the PRE sequence from the Woodchuck Hepatitis Virus (WPRE) increases viral titers by reducing the readthrough and improving the transcript termination[296]. The WPRE harbours an additional subelement that showed higher efficacy in lenti- and gamma-retroviral vector backbones irrespective of the promoter driving the expression and the presence of introns in the RNA[297]. LVV containing WPRE reported titers 5-7-fold higher to standard lentiviral productions[298,299]. Although the presence of an enhancer element (We1), the WHV X-protein promoter and 180bp of sequence coding for the X-protein within the WPRE raised some concerns regarding its safety[300]. This enhancer activity is not likely to pose a problem since a second enhancer (We2) is needed to drive gene expression[301], also in lentivirus[302]. However, X-proteins, and in particular those truncated in the COOH-terminal[303,304], might be indirectly (as a cofactor) involved in cellular proliferation and anti-apoptotic activity[305]. Such concerns were mitigated with the validation of an equally functional mutated derivative with abrogated WHV X-protein translation[306].

Further improvements after the SIN vector configuration comprise optimisation of packaging signal in the transfer vector and the *gag* packaging gene to decrease their homology[307] or via codon optimisation of packaging genes, which has also been reported to reduce the chances of RCL[308]. The *rev*-independence of codon optimised HIV-1 *gag-pol* gene enables the removal of the Rev-responsive element (RRE) from packaging plasmids. However, its nuclear export is compromised in the absence of *rev*. Kotsopoulou *et al*., compensated this with an overexpression of gag-pol[309]. However, most 3rd generation vectors still incorporate *rev*/RRE in *trans* as a fourth plasmid since *rev* independence resulted in lower titers. Other sequences such as constitutive transport elements (CTEs)[310,311] have shown not to be as efficient as *rev*[312] although less cytotoxic. Clontech claimed to have developed the fourth generation by further splitting the *gag-pol* cassette onto two cassettes (*gag-pro* and *vpr-pol*) although their system is not *tat* independent[313]. Precursor Pr55Gag polyprotein has been shown to play an important role in several processes of viral replication. Replacement of the Matrix protein (p17) myristoylation signal with phospholipase C-d1 pleckstrin homology (PH) domain has been shown to enhance viral production[314].

The choice of promoter to drive the expression of the internal cassette has also been optimised depending on the stemness of the target cell population and the genetic context. Cytomegalovirus (CMV) promoter is stronger than Spleen focus-forming virus (SFFV), phosphoglycerate kinase (PGK) and EF-1α in differentiated cells[315]. Nonetheless, all the aforementioned provide robust stable gene expression and have been used in clinical trials.

The chicken beta-globin insulator sequences (cSH4) have been introduced in order to neutralise the potential positional effect resulting in silencing of the expression cassette or activation of neighbouring genes. Inclusion of highly repetitive insulating sequences results in a reduction in titer and transgene expression and do not seem to compensate a questionable reduced risk of genotoxic effects[316,317]. The bovine growth hormone polyadenylation sequence has showed improved efficiency compared to the 3' LTR U5 region[318].

Other modifications include the mutation of the catalytic domain of lentiviral IN (D64V) to prevent integration. Episomal expression from integration deficient/ non-integrating lentiviral vectors (IDLV/NILV) offers an alternative as a delivery vehicle to support transient transgene expression. Despite IDLV titers are comparable to those of integrating lentiviral vectors, expression per copy has been shown to be 10-fold lower[319,320]. Background integrations were detected at a low frequency (10^{-4}) due to non-homologous end joining of lentiviral episomes into chromosomal double strand breakage sites[321]. Nonetheless, IDLVs may represent a safer transient viral mediate gene delivery system for transient expression[322]. When integration is required, hybrid vectors benefiting from the Sleeping Beauty transposon system provide a more neutral and thus potentially safer integration pattern[323,324].

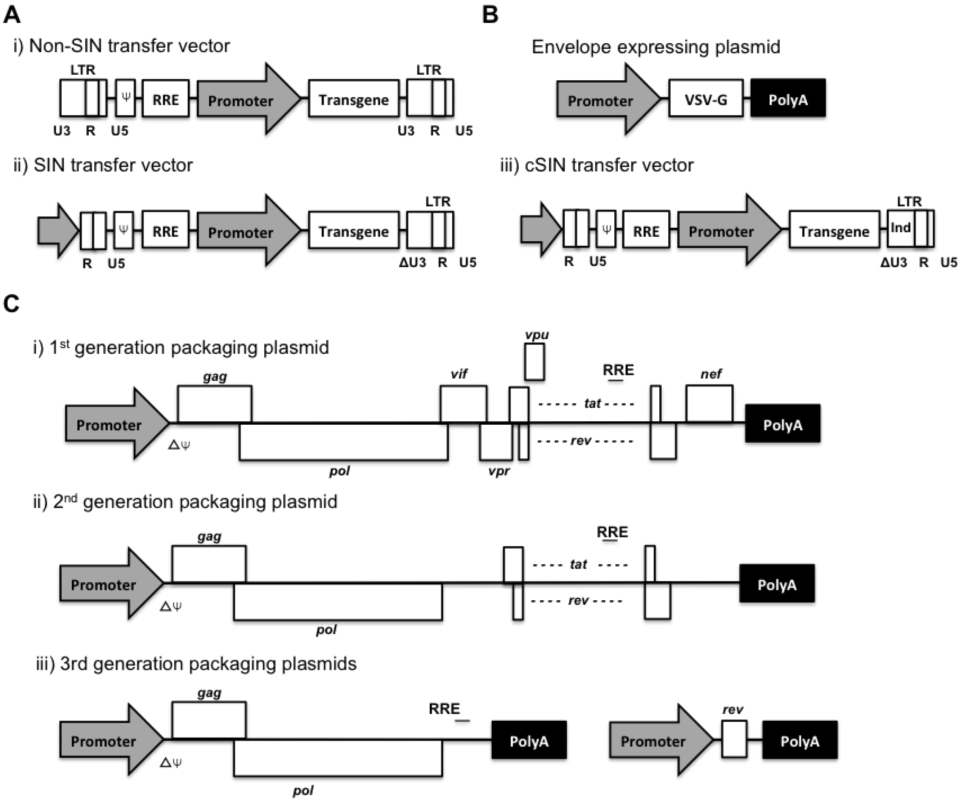


Figure 1.5. Lentiviral generations of packaging constructs and transfer vectors.

(A) (i) Non-SIN lentiviral transfer vector. (ii) Self-inactivating (SIN) lentiviral transfer vector; U3 is replaced by the Tat-independent promoter (usually CMV, cytomegalovirus or RSV Rous Sarcoma virus promoters) (iii) Conditional self-inactivating (SIN) lentiviral transfer vector: expression is dependent on the presence or lack of inducer molecule. (B) Envelope plasmid containing the VSV-G pseudotype. (C) Different generations of packaging constructs (i) first generation: viral accessory and regulatory proteins are present. (ii) Second generation packaging constructs: genes encoding for accessory proteins are depleted. (iii) Third generation. Gag-pol and rev are split into two different plasmids. VSV-G, vesicular stomatitis virus glycoprotein (VSVg), usually driven by the CMV promoter; Ψ, packaging signal; polyA, polyadenilation signal, RRE, rev sesponsive element. LTR, long terminal repeat; U3, unique in 3'; R, repeat; U5, unique in 5'; tat, trans-activator of transcription; rev, regulator of virion; nef, negative regulatory factor; vpr, viral protein R; vpu, viral protein U; gag, group-specific antigen; pol, polymerase; env, envelope.

1.1.6 Lentiviral vector production

Retroviral vectors have been used in the clinic for therapies against HIV/AIDS[325] and rare primary immunodeficiencies[32]. More recently, cancer immunotherapies have shown promising results from lentiviral-modified autologous T cells[326]. Production of infective lentiviral vector particles has been historically achieved by transfection of packaging genes and transfer vector for transient expression and is still today the most used method. Titers around 10^7 TU/mL are typically achieved with most protocols and can be increased to 10^9 TU/mL after ultracentrifugation.

Transient transfection

Transient transfection is the main method to produce LVV and relies on the delivery of gag-pol, rev (and tat if not using 3rd generation packaging systems), envelope protein and transfer vector to yield non-sustained expression of viral components and produce vector. Transient transfection saves the time-consuming cell line development of producer cell lines and allows expression of viral cytotoxic proteins. Transfection efficiency depends on the quality of the DNA, the method employed, the target cells and the size of the transfer vector (dropping significatively from 9 to 13kb)[327].

However, the main limitations of lentiviral production using transient transfection come from (i) the low adaptability of transient systems for large-scale production[328], (ii) the high cost of GMP plasmids (iii) contamination of the harvested vector with transfection plasmids[329] (iv) the difficulty in the optimisation of transfection conditions. Lentiviral vector production is still limited to a research setting using cell factories rather than large volume industrial bioreactors. Current transient transfection batches yield sufficient vector to treat one or a few patients (10^9-10^{11} TU[330,331]) limiting the reproducibility between patients in large clinical trials. Large-scale industrial batches (listed in Merten *et al.*,[332]) have achieved the 10^{11} TU threshold by optimising cell type, density media supplementation, and plasmid delivery. 293T cells are normally used for large-scale LVV production given their faster growth and enhanced productivity although 293 have also been used when traceability was not certain for GMP industrial productions[333]. Some protocols use adherent cultures, FBS and calcium phosphate[334], others use serum free-media with $Ca_3(PO_4)_2$[335] or PEI[336]. Segura *et al.*, used HEK 293 suspension cells transfected with PEI and protein-free media supplemented with Pluronic®[337]. Broussau *et al.*, and Côte *et al.*, isolated and established a HEK293 clone (HEK 293SF-3F6) for suspension culture[338,339]. 10^8 TU/mL have also been achieved using 3L stirred tank bioreactors using HEK 293SF-3F6 suspension adapted cells and serum-free media[340].

Packaging and producer cell lines

The development of packaging cell lines to produce lentiviral vectors potentially solves the aforementioned disadvantages, may produce higher amounts of vector and reduces the batch-to-batch variability. A lentiviral packaging (or helper) cell line (PCL) stably expresses the packaging and/or the envelope genes (*gag-pol*, *env* and *rev*) in *trans*. The packaging cell line becomes a producer cell line when the transfer vector is also provided to produce the packaged lentiviral vector particle. The ideal packaging (or producer) cell line should be stable in growth and vector production, produce high amounts of infective lentiviral vector and be adapted to serum-free and suspension conditions. Historical limitations of packaging cell lines are low vector titer (10^5 -10^7 TU/mL) and cytotoxicity of the lentiviral proteins leading to reduced stability over generations. Standard cell line development processes from initial cloning until master cell banking including the sequential integration and selection of all the vector components typically take from 6 to 12 months.

Choice of host cell line

More than 70% of the biopharmaceutical products in the market are produced in a few host cell lines: Chinese hamster ovary (CHO) cells, human embryonic kidney (HEK 293) cells, baby hamster kidney cells (BHK) and cells derived from mouse myeloma (NS0).

CHO cells were isolated by Theodore Puck in 1957 from a biopsy from *Cricetulus griseus*[341] and are currently the most widespread and well understood cell line in the industry for protein production[342]. CHO DG44, and DUK-B11 cells lack DHFR and are used in combination with MTX, which enables gene amplification of expression cassettes[343]. Instead, CHO-K1 cells use the GS/MSX system with methionine sulphoximine (MSX) concentrations above 3μM, enough to inhibit endogenous glutamine synthetase (GS)[344]. CHO cells are good secretors and produce proteins with a human-like glycan profile[345].

NS0 are murine myeloma cell lines derived from BALB/c mouse plasmacytoma cell line and due to their origin they are able to synthesize high levels of Ig[346]. NS0 express low levels of GS and consequently the GS/MSX system is mostly used for gene amplification[347]. Baby hamster kidney (BHK) cells were derived from subclone 13 of a parental cell line fibroblast from 5 unsexed 1-day old Syrian hamsters (*Mesocricetus auratus*) by Macpherson and Stoker in 1962[348]. All these cell lines have been adapted to grow in serum-free and suspension conditions at high densities and their glycosylation pattern is compatible with humans. However, lentiviral vectors have not been transiently produced in cell lines with rodent origin given that they are less susceptible to infection by human viruses due to the restriction factors[349].

Initially, retroviral vectors derived from MLV were produced in NIH 3T3 murine cell lines[350], which are strictly adherent. However, endogenous retrovirus could lead to potential mobilisation of vector genomes and generation of replication competent retroviruses raised safety concerns[287]. Another reason for the transition towards human cell lines is the fact that murine cell lines add a nonhuman sugar residue onto N-glycans (Galα1-3Gal) of the envelope protein and other membrane proteins[351]. Retroviral particles with this glycosylation pattern are neutralised by the human complement system in 20 minutes post-injection[352,353]. However, (Galα1-3Gal)-positive retroviral vectors (produced by a brain derived cell line, Mustela putorius furo, Mpf) are immune to human immunity indicating that other epitopes could participate in the immune system recognition[354]. Therefore, vector produced for *in vivo* applications must be produced in monkey or human-derived cell lines.

PerC6 cells were originally derived from healthy human embryonic retina and were immortalised by Crucell via transfection of the Ad E1 gene instead of viral transduction for biopharmaceutical production of proteins and Ad vectors[355]. PerC6 cell lines were explicitly designed for biopharmaceutical production and its traceability is extensively documented. The main advantage of PerC6 cell lines is their ability to grow to high cell densities in suspension, which results in higher product titers. Although their glycosylation pattern is slightly different from that of humans (fewer mannoses and hybrid structures) it is not immunogenic[356]. No amplification systems are needed since stable levels of expression are obtained from low copy number transfection[355]. In 2011, the acquisition of Crucell by Johnson&Johnson caused the discontinuation of the distribution of commercial licenses for biopharmaceutical manufacture.

Human embryonic kidney cell lines were originally isolated from a healthy aborted female foetus in 1973 in the laboratory of Alex van der Eb in Leiden[357]. During his 293rd experiment (which gave them their current name), Frank Graham transformed them using the mechanically sheared fragments of human adenovirus 5 (hAd5) using the calcium phosphate co-precipitation transfection technique[358]. Analysis of the HEK 293 genome by Louis *et al.*, 4,344 bp showed the DNA fragment corresponded to the 11% far 5' end of the Ad genome. This fragment contains the E1A, E1B and IX early hAd5 genes, which were integrated in the human pregnancy-specific beta-1-glycoprotein 4 (*PSG4*) gene located in chr19q13.2[359]. A genomic study of the cell line by Lin *et al.*, revealed that the hAd5 integrated fragment had undergone genome amplification resulting in 5-6 copies[360].

HEK 293 cells (available at the non-profit American Type Culture Collection repository under the catalog number CRL-1573) are dynamic cells and changes have been observed over passages. Growth rate increases more than twice in 40 passages (from 0.29/day at passage 43 to 0.74/day at passage 70-80) and cell size diminishes[361]. Their tumorigenicity also varies from being negligible in early passages (up to 21) to cause solid tumours (after passage 65) in two weeks when injected in nude mice[362]. Therefore, it is critical to maintain cells in early passages to limit their variability for any application. Although they were termed human embryonic kidney cells and were originally thought to have fibroblastic, endothelial or epithelial origin, their response to neuronal signalling, the presence of neuron-specific voltage channels and susceptibility to infection by neurotropic viruses suggest that they belong to neuronal lineage in the kidney[363]. HEK 293 cells are cultured in adherent cultures typically with DMEM and supplemented FBS but can adapt to suspension cultures in the absence of serum in low calcium ion concentration media[350]. HEK 293 cells are pseudo (or hypo) triploid, meaning that their genome has less than three sets of chromosomes with a modal number of 64 chromosomes. However, their abnormalities include four copies of chromosome 17 and 22 three copies of chromosome X and no traces of Y chromosome (the latter as expected)[360].

HEK 293 cells have given rise to several derivatives such as HEK 293T, HEK 293E and HEK 293FT and have been used for AAV, Ad, MLV and LVV production[364]. HEK 293T cells (originally referred to as 293tsA1609neo) (ATCC CRL-3216) were obtained by DuBridge *et al.*,[365] in the laboratory of Michele Calos upon stable transfection of standard HEK 293 with 2 plasmids a pRSV-1609 plasmid[366] containing a temperature sensitive SV40 T-antigen coding sequence driven by the RSV promoter. The other plasmid, which is no traceable in the literature, contains a neomycin resistance gene as a selectable marker for stable integration, thus 293T are resistant to neomycin. They are easily transfectable and grow faster than standard HEK 293 as T antigen interacts with several proteins and inhibits replication control[367].

HEK 293FT adherent cells (Invitrogen R700-07) were obtained after transfection of standard HEK 293 with pCMVSPORT6Tag.neo by Life Technologies in 1988 and are traceable since then. These cells also constitutively express the SV40 T-antigen (Ag) under the CMV promoter and have a similar growth rate to HEK 293T. SV40 T antigen allows amplification of transfected plasmids with a compatible origin of replication[365], which leads to higher production rates[368]. However, as with HEK 293T cells, the association of SV40 with cancer (T Ag complexes and p53, which inhibits its tumour suppressor function)[369] raises concerns on the utilisation of these cell lines for biopharmaceutical production. Nevertheless, the adenoviral E1 region was used to immortalise HEK 293 cell line and such cell line has been validated for clinical grade biopharmaceutical products[370] and no adverse events associated with the T antigen have been ever reported. Moreover, millions of people were accidentally inoculated with SV40 detected as a contaminant of the polio vaccine between during the 1950s in the USA and Denmark; follow up studies found no increase in the cancer incidence[371].

HEK 293EBNA-1 (or HEK293-E cells, ATCC CRL-10852, R620-07, Life Technologies) were established by inserting the Epstein Barr virus (EBV) nuclear antigen-1 (EBNA-1), from pCMV/EBNA. EBNA acts as a transcriptional enhancer[372] and allows episomal replication and maintenance of plasmids containing the EBV oriP origen of replication (*ori*P) in *cis*[373], which increases protein yield[374]. HEK293-E cells are also neomycin resistance as a result of the stable expression of the neomycin resistance gene driven by the Rous Sarcoma Virus long terminal repeat promoter from pRSV4neo[375].

HEK 293EBNA-1 6E cells (originally from the National Research Council of Canada, NRC file 11565) also termed or 293-6E cells[376] stably express a truncated version of the EBNA-1, lacking Gly-Gly-Ala domain. Expression of this truncated form is more stable and less cytotoxic and cell lines show higher growth rates and increased transient gene expression compared to full length EBNA1[377].

Construct design and mode of expression

Stability of production is one of the main hurdles due to cytotoxicity derived from prolonged expression of toxic vector components. The toxicity of the envelope protein (and other potential viral proteins) dictates the strategy employed for its expression. Packaging cell lines can be divided in two categories (constitutive and inducible) depending on the mode of expression of its viral proteins.

Constitutive expression

Several groups have generated packaging cell lines that constitutively express viral proteins[378–383] on occasions reaching titers >10^7 TU/mL. Given the inherent toxicity of VSV-G protein and some elements of p24Gag, constitutive packaging cell lines use other non–cytotoxic envelope pseudotypes and thus can constitutively express viral proteins for prolonged periods. Feline endogenous virus envelope RD114 protein has wide range of cell tropism but shows preferential tropism for hematopoietic CD34+ stem cells[384,385], a therapeutic target for gene therapy; Gibbon ape leukaemia virus has preferential tropism for hematopoietic progenitor cells and peripheral blood lymphocytes[386,387] and has been used in clinical trials carried out with γ-retroviral vectors[388]. The STAR[389] packaging cell line developed by Ikeda *et al.*, was the first using this mode of expression. In that study, three potential host cell lines (HT1080, HeLa and 293T) were tested with three different envelope proteins (RD114-Pro with a protease site at the R cleavage site; MLV 4070A and GALV with a cytoplasmic MLV domain). In addition, Ikeda's approach was novel because viral genes were introduced by transduction (instead of transfection) using a 2nd generation MLV vector. However, although titers were >10^7 TU/mL for 12 weeks, the use of non-SIN vectors to generate the PCL specifically impeded STAR cell lines could progress for clinical applications[385]. WinPac[381] packaging cell line used the same principle to insert *gag-pol* but provided a modular approach (already used in γ-retroviral PCLs[390–392]) in which viral transduction was used as a platform to deliver/retarget other viral components via recombinase mediated cassette exchange (RMCE). However, site-specific insertion does not completely eliminate interclonal variation in expression and therefore screening of clones is still required. In addition, the lower titers (>10^6 TU/mL) could be associated to the toxicity from integration of *gag-pol* into high transcribing sites.

RD2-MolPack developed by MolMed is another example of a constitutive packaging cell line expressing RD114-TR envelope protein (containing the cytoplasmic domain (TR) of MLV-ampho 4070)[393]. Interestingly, the packaging genes were transduced using a baculo-AAV vector previously transfected with AAV Rep78 to target their integration. However, the safety profile of this cell line was not optimal since co-expression of *gag-pol* and *rev* were driven from the same plasmid and the transfer vector was not self-inactivating. RD3-MolPack corrected the issue using SIN vectors but titers remained at approximately 10^6 TU/mL in both cases[382]. Despite its cytotoxicity, the use of VSV-G is still generalised due to its multiple advantages (broad tropism, stable particles upon ultracentrifugation[283]) and none of the other pseudotypes has gained FDA clearance for gene therapy with lentiviral vectors[394]. In any case, an ideal producer cell line platform should be able to support any pseudotype.

Inducible expression: Inducer-Off

Inducible expression systems are meant to regulate expression of cytotoxic viral proteins in cells e.g. VSV-G and protease amongst others. The VSV-G pseudotype has been extensively used due to its wide tropism (mammalian and non-mammalian cells) and its stability against ultracentrifugation shearing forces. However, it has been shown to be toxic through the formation of syncytia and subsequent cell death when constitutively expressed in packaging cell lines[283]. Vpr accessory protein, although dispensable in SIN-LVV have also been shown to be toxic[395–397]. In order to overcome these limitations, inducible systems using the Tetracycline (Tet) regulatory system[292,398–404] or the ecdysone[405,406] regulatory system have been assessed. When using the Tet-Off system, tetracycline (or its analogue doxycycline) is used to block the binding of the tetracycline transactivator (tTa) to the Tet responsive element (TRE), which supresses the expression of the gene. The tetracycline transactivator (tTa) is a chimeric protein resulting from the fusion between the DNA binding domain (N-ter) of the tetracycline repressor (TetR) and the activation domain (C-ter) of the Herpes Virus V16 protein (HV-VP16) transactivator[407]. Initially, Yu *et al.*, used the Tet-Off system in HeLa cells that constitutively expressed tTa (HtTA-1 cell line) to regulate the expression of *rev* and the latter indirectly that of packaging genes[408]. Kafri *et al.*, developed a 1st generation LVV cell line with Tet-off inducible expression of *VSV-G* yielding acceptable titers (>10^6 TU/mL) for 3-4 days[398]. Second generation lentiviral vectors demonstrated that accessory proteins are redundant for lentiviral vector transduction[286,312,409]. Consequently, in the following years, packaging cell line designs did not include *vpr*, *vif*, *vpu* and *nef* complementing in *trans*[399,400,402,405,408]. Kaul *et al.*, still reintroduced viral accessory proteins in *trans* with the objective of boosting the titers[402]. Following Kafri's approach, Farson *et al.*, independently developed 2nd generation packaging cell lines and achieved similar titers[399,400]. However, the design of this cell line (2nd generation LVV) was not considered sufficiently safe. Klages *et al.*, achieved >10^6 TU/mL in the absence of *tat* and *rev* was stably cotransfected as a fourth plasmid under the control of TRE, increasing system biosafety and reducing percentage of homology with the HIV genome to 40%[399]. This way *gag-pol* mRNA nuclear export was regulated by *rev* in a 2-step regulation system.

With the arrival of the third generation of lentiviral vectors[146], the removal of *rev* and *tat* regulatory genes was attempted by replacing them with complementary systems. The regulatory protein Tat is responsible for transcription of the full-length vector genomic RNA in HIV (and up to 2nd generation LVV). Elimination of this dependence and reduction of the viral homology was achieved by replacing the HIV 5' U3 region with a constitutive heterologous promoter e.g CMV or RSV promoter[405].

Xu *et al.*, introduced a novelty in the PCL design consisting of the replacement of the third generation LVV 5' LTR U3 promoter/enhancer region with seven copies of the Tet responsive element (TRE) giving rise to conditional self-inactivating lentiviral vector system (cSIN)[404]. Unlike standard SIN vectors, cSIN design allows delivery of transfer vector via transduction and yielded titers of >10^6 TU/mL.

In 2006, Cockrell *et al.*, combined the Tet-Off system with further splitting of the gag-pol construct into gag–pro and vpr–RT–IN[401] and a standard SIN LVV configuration to reduce the risk of production of replication competent lentivirus and yet obtain relatively high titers (>10^7 TU/mL). Despite being a transient packaging system but in line with this concept, Westerman *et al.*, used a 7-plasmid (non-cSIN) system where, besides plasmids encoding VSV-G, *tat*, *rev* and the transfer vector plasmid, up to three constructs for the *gag-pol* (Gag + Vpr-PR + Vpr-RT/IN) were used. However, titers dropped from the >10^6 TU/mL using the 5 plasmid system to the >10^5 TU/mL using the 6 and 7 plasmid system[410].

Ni *et al.*, at Virxsys developed a 3-step regulation system that avoids constitutive expression of cytotoxic viral proteins and also toxicity present in the tetracycline transactivator (tTA). In their work, the tTA is under an inducible system, which upregulates itself upon induction. This system also included expression of codon optimised *tat* and *rev*, which in turn regulates codon optimised gag-pol and VSV-G transcription[403]. Their strategy yielded 3.5x10^7 TU/mL for 11 days but leakiness of p24Gag expression resulted in silencing after 2-3 months. Gene silencing was confirmed not to be caused by gene loss but at an expression level. Methylation of ERVs, transposons or even at a post-transcriptional level has been observed among other mechanisms[411] in eukaryotic cells as a defence mechanism to the expression of foreign DNA[412,413].

Throm *et al.* used SIN-MLV to deliver *gap-pol* and *rev* (and *tat*) genes into a 3rd (and 2nd) generation GPRG (and GPRGT) packaging cell line regulated by the Tet-Off system. Unlike Ikeda's work in which LTR-MLV were used to deliver packaging genes[414], the use of SIN-MLV reduced the risk of cross-packaging of MLV genomes in lentiviral particles and allowed clinical applications. Interestingly, Throm *et al.*, also used a concatemeric array of vector genomes to enhance the expression of vector genomes, which yielded 5x10^7 TU/mL[292].

The Tet-Off system represents an advantage from the downstream processing point of view compared to the Tet-On system, as no inductor molecule is present in the culture during vector production. However, complete elimination of repressor to promote the induction of the system requires a full media change and represents a challenge for large volume bioreactors. Morover, cells need several days to reach peak of production, which makes the Tet-Off system not optimal for large-scale production. The regulation of these inducible systems is not always tight and stability is compromised due to the leakiness of VSV-G expression in the off-state. As a result, genetic and transcriptional instability was shown after 2-3 months of culture when using this method[400,403]. In addition, there is a delay between the removal of doxycycline and the induction.

Inducible expression: Inducer-On

In the Tet-On mechanism, the binding of the inducer molecule to the tetracycline transactivator (rtTA) promotes the binding of the TRE and thus switching on expression of the gene. Stewart *et al.*, (Oxford Biomedica) developed a EIAV-based packaging cell line to generate *ProSavin*® for the treatment of Parkinson's disease[415]. The cell line was based on codon optimised TetR (coTetR) enhancing its expression and the two obtained clones (PS46.2 and PS5.8) achieved a tight regulation and stable titers (<10^6 TU/mL) for 7 weeks[63]. Further modifications of the Tet-On system include the Tet repressor devoid of the VP16 protein, which has been suggested to be toxic for cellular transcription[407]. Location of the two copies in tandem of the TetO 10bp downstream of the CMV immediate early promoter TATA box allow blockage of expression by TetR homodimers in absence of inducing agent[416]. Using this strategy, Stewart *et al.*, reported stable (although low titers, mid <10^6 TU/mL) for 16 weeks in absence of selective pressure[63,415]. Recently, other Tet On inducible systems have been developed by Clontech. The Tet On 3G system achieves a x25,000 induction factor by constitutively expressing the transactivator 3G molecule under the PGK promoter, which activates transcription (*in trans*) of the gene of interest, downstream the TRE3G promoter in the presence of doxycycline[417].

Similarly, regulation of viral protein expression using the ecdysone system was used as an alternative to the Tet system since it is less leaky and more rapid in the induction (3-5 days instead 14 days) and clearance[418]. The insect hormone ecdysone (or its analogue ponasterone A) promotes the binding of this molecule to the ecdysone receptor (VgEcR)-retinoid X receptor (RXR) heterodimer and thus activation of transcription. Since this protein is not endogenous, basal levels are considered negligible. However, packaging genes under the control of one[405] or separate[406] ecdysone promoters only yielded 10^5 TU/mL (before concentration).

Another example of inducible Tet-On system is the work of Broussau *et al*., in 2008, who combined a reverse transactivator (rtTA2S-M2) of the tetracycline system with a cumate switch (CymR from *Pseudomonas putida*) for *rev* and VSV-G expression (and constitutive expression of *gag-pol*) achieving stable suspension cell lines for 18 weeks and induction/production cycles of 7 days in absence of selective pressure[338] and promising titers (3.4x10^7 TU/mL). Cumate and doxycycline can be removed after ultracentrifugation[419]. However, despite controlled expression of cytotoxic genes upon induction, their effects cannot be mitigated in long-term cultures.

Nonetheless, cytotoxicity is not the only limitation that impedes high titers. A correlation between titers and the amount of transfer vector copies introduced in producer cell lines was identified by Sheridan *et al*.,[420]. As previously mentioned, Throm *et al*., corroborated the insufficient expression of SIN LVV transfer vector genome as a limitation for vector production, already confirmed by Ikeda *et al*.,[389] and proposed a new approach based on the transfection of >200 copies in tandem (as a concatemeric array) of transfer vector[292] yielding >10^7 TU/mL.

A different strategy to approach the limiting factor is that followed by Sanber *et al*. In their study, MLV vectors were used to target recombinase recognition sites into actively transcribed sites in a controlled way. Gag-pol genes were then retargeted using recombinase-mediated cassette exchange (RMCE) and the *env* and *rev* genes were finally stably transfected yielding titers higher than 10^6 TU/mL[381].

Finally, inducible systems based on light-switchable promoters have been suggested as innovative mechanisms to regulate LVV production. The change of conformation (*trans* to *cis*) of azobenzene upon reversible induction with UV light (300-400nm) allows activation with short pulses of light[421,422]. Similarly, when excited with far-red light, photoresponsive phytoreceptor interacts with PIF3 phytochrome and mediates transcription of downstream genes[421,423].

To date, only two lentiviral packaging cell lines have been exploited for production of SIN-LVV for clinical trials. Two of them are derived from GRPG/T cell line generated by Throm *et al.*, with titers >10^7 TU/mL: (i) GPRGT-derived 650MNDhWASp1 packaging cell line by Wielgosz *et al.*, expressing WAS protein for the clinical trial treating Wiskott Aldrich Syndrome at the St. Jude Children's Research Hospital[424] and (ii) GPRG-CL204i-EF1α-hγcOPT by Greene *et al.*, expressing IL-2Rγc for SCID-X1[292,425].

Stoichiometry

As addressed in previous sections, lentiviral vector genes are split into several expression cassettes to avoid the generation of RCL. However, as a consequence, the stoichiometry is disrupted as each vector gene is expressed separately. In addition, accessory proteins, removed after the second generation of LVV, cannot participate in its modulation. Viral gene expression and splicing dictate the efficiency of assembly in lentiviral vectors. Unspliced RNA gives raise to gag-pol polyprotein and full length viral RNA and spliced RNA giving rise to envelope protein among others. Katz *et al.,* showed the amount of unspliced:spliced RNAs follows a 1:2 ratio in physiological conditions[426]. In turn, such unspliced RNA proportions must be in conjunction with the gag:gagpol frameshift rate (20:1)[427]. The introduction of mechanisms to finely control viral gene expression via stable transfection and (MOI-regulated) transduction, often used in PCL development, is cumbersome. Therefore, achieving such proportions of RNA species becomes a challenge to replicate the normal physiology of the virus.

Yap *et al*., demonstrated that the effect of gag-pol overexpression depends on the amount of envelope protein used in viral particle formation. The interaction between the premature envelope protein and its receptor within the cell constitutes a limitation and compromises its availability for the packaging of viral particles[428]. In normal circumstances, vpu viral accessory protein prevents the interaction between the viral envelope protein and the premature receptor by down-regulating the expression of CD4 receptor. In gamma-retroviral vectors, devoid of vpu, envelope protein is overexpressed to compensate for the interaction. This, together with the existence of a threshold level of env protein for packaging of MLV viral articles highlights the relevance of the env protein in viral production[429]. Low levels of receptor could have a positive effect and prevent the entry of viruses. Despite studies of stoichiometry by Katz *et al*., and Yap *et al*., the issue had never been considered in the context of a packaging cell line. However, when expression of envelope protein was not a limitation, they observed a large number of empty particles suggesting the expression and/or packaging of vector genome was compromised. This limitation, already identified by Sheridan *et al*., was also corroborated in part by Lei and Andreadis[430]. In their study, ecotropic envelope producer cell lines showed a large number of empty non-infectious viral particles, while this trend was not seen with amphotropic producers. In 2007, Carrondo *et al*., used the Flp/FRT system to mediate cassette replacement and assess the influence of each of these components on vector production. Interestingly, they found gag-pol expression is pivotal since a 2-fold variation in its content could impact titers by one or two orders of magnitude[431]. Its balance with env expression was also shown to be critical for viral infectivity, identifying a 100-fold margin between balanced and unbalanced gag-pol/env expression. Their study also showed stability of the infective particles remains unaltered if conditions are suboptimal although transduction efficiencies can dramatically drop. But more interestingly, by firstly integrating the transfer vector and selecting clones with no limitation in the amount of RNA expression and secondly preserving an optimal of gag-pol/env balance on those clones, they showed titers were not significantly increased. These findings confirm that there is room for improvement in encapsidation of viral genomes when the stoichiometry is balanced and expression of the viral genome is not a limitation.

In conclusion, each packaging cell line presents different limitations depending on differential expression of viral components, which rely on integration site, copy number or silencing. In the event of correct balance between gag-pol and env, limitation coming from vector genome can reside either on insufficient expression or on packaging efficiency.

Delivery and selection of lentiviral components

Transduction

Delivery of packaging plasmids can be achieved using either transfection or viral transduction. Transduction has shown to produce more stable expression and higher titers than transfection using second generation packaging plasmids, despite having only one copy of the transgene when a low multiplicity of infection is applied (MOI, ratio of viral particles to cells)[389]. That is explained by the ability of retroviral vectors to integrate genes into the host cell genome providing more stable expression. Integration catalysed by the viral integrase may contribute to genetic stability compared to stable plasmid transfection, which relies on double strand breaks and thus a potential selection of genetically unstable loci. Many groups have used this method to deliver packaging genes[338,398–401,403–406]. Interestingly, Ikeda *et al*., also reported enhanced probability of high-producing clones when using this method[389]. Second generation LTR- γRV and third generation SIN γRV have been used for permanent delivery of packaging genes for example in STAR and GPRG-TL-20 packaging cell lines (with the exception of *env*, which was delivered using transfection[389]). However, delivery of full SIN-LVV transfer vector (containing the ΔU3 deletion) is not recommended. This is because once transcribed, the 5’LTR no longer has promoter/enhancer activity and the transferred U3 from the 5’ LTR during reverse transcription is inactivated as it contains the ΔU3 deletion. Therefore, plasmid delivery for generation of SIN transfer vector has to be done by transient DNA transfection.

Transfection

Plasmid transfection protocol typically involves delivery of an stoichiometrically optimised mix of plasmids to a monolayer of cells, change of media a few hours/one day post-transfection (if transfection method is cytotoxic) and media harvest two days post-transfection followed by 0.45 μm filtration prior to an optional ultracentrifugation. Physical and chemical transfection methods are generally versatile, rapid, non-cell type dependent and reproducible.

Physical methods require sophisticated equipment although they avoid many of the undesirable effects of chemical and viral transduction. Physical methods include methods like high-velocity biolistic transfection of nucleic acid tungsten or gold-coated microparticles[432], laserfection-mediated permeabilisation of membranes (also known as optofection or phototransfection)[433] or magnetofection, which is particularly attractive for primary cell lines given its mildness and can be performed in the presence of serum[434]. However, electroporation, first used by Wong and Neumann *et al.*, in 1982[435], is the physical method par excellence. It is based on the application of electric fields to cells and tissues, which causes the appearance of transient aqueous pores, which results in an increase of the permeability of the cell membranes and tissue to extracellular DNA[435]. Interestingly, electroporation technology has been improved to enable continuous transfection of large volumes of flowing high density cultures[436]. This method is compliant with current regulations[437], requires 33% less DNA than other transfection methods and can be used at a bioprocessing scale for lentiviral vector production with titers 8.8x10^7 TU/mL[437].

In general, chemical methods are inexpensive, non-mutagenic, and adaptable to high-throughput applications. In addition, unlike viral delivery they do not present limitations on the amount of nucleic acid loaded and can be easily used in many cell types with varying efficiencies. Among chemical methods, calcium phosphate is sensitive to pH variations and cells require low concentration of BSA or FBS to reduce the cytotoxicity. Liposomes and cationic lipids or polymers are relatively expensive to scale up although they require less DNA than calcium phosphate[438]. Polyamidoamine dendrimers (like PEI) are less expensive, their efficiency is similar to that of calcium phosphate and they require less DNA. However, DNA-PEI complexes are relatively cytotoxic and require change of media a few hours post-transfection[439]. Key parameters to be optimised not to hamper their efficiency include the nucleic acid:chemical agent ratio, serum concentration, pH, exposure to the transfection or permeabilising reagent. There is no ideal method suitable for all application. The choice of a delivery method is dictated by the inherent properties, its manufacturability and target application.

An inherent disadvantage of stable plasmid transfection is the introduction of antibiotic resistance in the packaging cell line to select for stable expressers in each round of transfection. When not co-transfected along with the transgene in separate plasmids, antibiotic resistances are advised not to be in the same expression cassette for clinical applications[440] as their unmethylated CpG islands can induce innate immune response via the Toll-like receptor 9[441]**.**

Suspension adaptation, scale up and upstream process improvements in lentiviral production

Unlike murine NIH 3T3 cell lines, used as gamma-retroviral vector producer cells, HEK 293 cells can adapt to suspension conditions media devoid of serum and containing low Ca^{+2} concentrations[442]. Calcium is involved in cell to cell adhesion through cadherins, transmembrane calcium-dependent proteins[443]. This makes their scalability much easier, as they can be cultivated in different suspension systems such as spinner flasks, fixed bed, fluidized bed or stirred tank bioreactor (with optional perfusion). For cell lines that require cell adhesion, such as 293Ts, lentiviral productions can be scaled-up to using cell factories or stacks, units of up to 40 layers of plates or chambers providing a culture surface of 25,280 cm^2. Using this method, large-scale LVV production (250mL with $2x10^9$ TU/mL) have been achieved for the treatment of *ex vivo* immunodeficiencies[444]. HYPERflasks have shown a >9x10^6 TU/cm^2 improvement in surface productivity compared to normal T-flasks, possibly attributed to a better gas exchange[445]. Other alternatives comprise fixed bed reactors, hollow fiber reactors and micro- or macrocarriers (also used for gamma-retroviral or adenoviral production) although titers do not show an increase in productivity[446].

LVV production using suspension HEK 293E cell lines was first achieved in 3L stirred tank bioreactors yielding >10^6 TU/mL[447]. More recently, Witting *et al.*, reported titers >10^8 TU/mL using bag bioreactors, GMP-compliant closed systems that are easily scalable[448].

Increases in productivity of producer cell lines have been attempted at different levels. In upstream process, addition of sodium butyrate (NaBu) has been one of the most widely used strategy[449]. NaBu inhibits HDACs and promotes hyperacetylation of histones and other nuclear proteins[450], which translates into an increase in chromatin accessibility remodelling and transcriptional stimulation leading to higher titers. Lei *et al.*, reported 2-3-fold increase of retroviral p30 protein with 2-20mM of NaBu[430]. However, NaBu effects seem to be linked to the envelope protein pseudotype[451]; the enhancement of LVV titers pseudotyped with VSV-G is controversial[63]. Other authors have shown an increase in titers using chloroquine[452] and caffeine[453]. The former acts by increasing the pH of lysosomes and thus preventing degradation of transfected DNA (although it is highly dependent on the delivery method with which it is combined[454]) while the mechanism of action of the latter remains unclear.

Another phenomenon observed during culture of lentiviral packaging cell lines using VSV-G is autotransduction, as producer cells do not usually have superinfection interference[455]. Reverse transcriptase inhibitors such as azidothymidine[455] and tenofovir[456] have been used to prevent autotransduction of packaging cell lines and their consequent increase in vector copy number. However, this components need to be removed during downstream processing, which presents further complications. Table 1.2 summarises the main feature of all the packaging cell lines reported in the literature to date.

Table 1.2. Packaging and/or producer cell lines (PCL) for lentiviral vector production developed and published to date.

Author	Year	Vector/SIN?	Envelope protein	Pack/ Prod	Mode of expression	PCL name	Parental cell line	Construction method (and comments)	Titre (TU/mL)	Ref
Carrol	1994	HIV non-SIN	HIV-1 *env*	Pack	Constitutive	D3.2/B4.7	Vero	Stable transfection of a plasmid containing all the packaging genes and hygro resistance (only 5′LTR). Ψ, PPT and 3′LTR provided in *trans*	10^2(SupT1 cells)	457
Poeschla	1996	HIV non-SIN	HIV-1 *env*/VSV-G	Pack	Constitutive	n.s.	HeLa T4	Cotransfection of one ΔΨ plasmid containing all the packaging genes, *env* and LTR and a transfer vector.	>10^4 (HeLa T4 cells)	378
Corbeau	1996	HIV non -SIN	HIV-1 *env*	Pack	Constitutive	n.s.	phi422	Co-transfection of one ΔΨ plasmid containing all the packaging genes, *env* and LTR and a transfer vector.	10^5(CD4+ cells)	379
Yu	1996	HIV non -SIN	HIV-1 *env*	Pack	Inducible (Tet-Off)	#69 (HtTA-1)	HtT4 (HeLa)	Sequential transfection of 2 plasmids containing *gag-pol* and *rev+env* and eventually the transfer vector.	7.3x10^3 (HeLa T4 cells)	408
Srinivasa-kumar	1997	HIV non-SIN	HIV-1 *env*	Pack	Constitutive	2A.22 B4.14 5BD.1	CMT3	Cotransfection of *gag-pol+rev*, then *env* and eventually transfer vector. Evaluation of the Mason-Pfizer monkey virus (MPMV) constitutive transport element (CTE) instead of *rev*	10^3-10^4 (HeLa CD4 cells)	380
Haselhorst	1998	HIV non -SIN	HIV-1 *env*	Pack	Constitutive	n.s	MDS/ SW480	Sequential transfection of *gag-pol, rev*, then *env* and finally transfer vector	10^1-10^2 (HeLa T4 cells)	458
Kafri	1999	HIV non -SIN	VSV-G	Prod	Inducible (Tet-Off)	SODK1CG1	293T.tA (SODK0)	Co-transfection of 2 plasmids containing *VSV-G* and *gag-pol*, *rev* and subsequent transduction of transfer vector.	3x10^6	398
Kaul	1998	HIV non -SIN	HIV-1 *env*	Prod	Inducible (Tet-Off)	B16 clone	HeLa.tT4	Sequential transfection with 3 plasmids: *rev-env*, then *gag-pol-tat* and finally transfer vector	2.9x10^4 (HeLa T4 cells)	402
Klages	2000	HIV non-SIN	VSV-G	Prod	Inducible (Tet-Off)	LV^G-1/GFP	293 (TRE.VSV-G.tTA)	293 sequentially cot-ransfected with 4 plasmids: tTA+tet/*VSV-G* and then *Gag-pol/Rev*. Resulting LV^G packaging cell line then transduced with a transfer vector.	3.5x10^6 (HeLa cells)	399
Farson	2001	HIV non-SIN	VSV-G	Prod	Inducible (Tet-Off)	Lenti*kat*	293G	293G cells were transfected with aplasmid containing *gag-pol, rev, tat* and then sequentially transduced with an inducible *VSV-G* cassette and a transfer vector.	5x10^6	400
Pacchia	2001	HIV non-SIN	VSV-G	Prod	Inducible (Ecdysone)	REr1.35	293T	Sequential transfection with3 plasmids*: gag-pol-rev* (deletions in other accessory genes) or CTE, *VSV-G* and finally the transfer vector	1.2x10^5	405
Sparacio	2001	HIV non-SIN	VSV-G	Pack	Inducible (Ecdysone)	293-Rev /Gag/Pol	293	Sequential co-transfection with *tat-rev* or *rev* and *gag-pol*. 293-*gag-pol* transiently transfected with transfer vector and *VSV-G*	3.0x10^5 (HeLa cells)	406
Xu	2001	HIV cSIN	VSV-G	Prod	Inducible (Tet-Off)	SODk1 cSCG	293T.tTA (SODk1)	Transduction of transfer vector using cSIN and subsequent co-transfection of 2 plasmids: *gag-pol, tat, rev* (no *nef vif, vpr*) and *VSV-G*	2.0x10^6 (293T cells)	404
Kuate	2002	SIV non-SIN	VSV-G	Prod	Inducible (Ecdysone)	SgpG109	293	Sequential transfection with *VSV-G* and *gag-pol* and finally transduced with transfer vector (containing *tat* and *rev*)	2x10^5	459
Ikeda	2003	HIV non-SIN	MLV 4070A, GaLV, RD114-PR	Prod	Constitutive	STAR	293T	LTR-γRV transduction of *gag-pol* genes. *RD114 env* and *rev* genes are integrated by plasmid transfection.	1.0x10^7 (SIN-LVV)	389
Ni	2005	HIV non-SIN	VSV-G	Pack	Inducible (Tet-Off)	17B-5	293	Co-transfection of 3 plasmids containing *VSV-G*+RRE, *gag-pol*+TAR and TRE-tTA and TRE-*rev*+*tat*. Transduction of transfer vector	3.5x10^7 (on HeLa-tat cells)	403
Strang	2004	HIV non-SIN	HIV *env*	Prod	Constitutive	SFV E2E1 RRV E2E1	STAR (293T from[389])	STAR cells (expressing Gag, Pol, Rev, and Tat) were transfected with HIV env and transduced with a transfer vector	>10^5 (293T + polybrene)	460

Cockrell	2006	HIV cSIN	VSV-G	Prod	Inducible (Tet-Off)	SODk3	SODk0 or 293T	Transfection with TRE-VSV-G. Subsequent cot-ransfection with *gag–pro* and *vpr–RT–IN*. Transduction of cSIN transfer vector	1.0x10^{7} (293T)	401
Muratori	2006	HIV non-SIN?	VSV-G	Prod	Inducible (Ecdysone)	293-Rev /Gag/Pol	18-4[406] (293T)	293-Rev/Gag/Pol express gag-pol and rev separately and were transfected with transfer vector and subsequently with *VSV-G*	n.s	461
Broussau	2008	HIV SIN	VSV-G	Prod	Inducible (Tet-On/cumate switch)	293SF-PacLV	293SF	2 strategies. One-shot: Co-transfection of 3 plasmids: *gag-pol, rev* and *VSV-G*. Second transfection with transfer vector. Two step: first co-transfection with 1 plasmid *rev*, *gag-pol*; second with 2 plasmids *rev* and *VSV-G*	3.4x10^{7}	338
Throm	2009	HIV SIN	VSV-G	Prod	Inducible (Tet-Off)	GPRG-TL20-GFP	293T/17	Serial transduction *gag-pol*, *rev*+tTA, *VSV-G* and finally a concatemers of transfer vector	5.0x10^{7}	292,425
Stewart	2009	EIAV SIN	VSV-G	Prod	Inducible (Tet-On)	PS5.8, and PS46.2	293T	Sequential transfection with coTetR, gag-pol, VSV-G, and transfer vector	<10^{6}	63
Lee	2012	HIV SIN	SVGmu	Prod	Inducible (Tet-Off)	DC-LV	GPR cells[292] (293T)	GPR (expressing *gag-pol* and *rev*) transduction with Tet-off/SVGmu (env). Transfection of a concatemers of transfer vector	>10^{7} (293T cells)	462
Storna-iuolo	2013	HIV SIN	RD114-TR	Prod	Constitutive	RD2-MolPack-Chim3	293T	Serial load of HIV *gag-pol*, *rev* with baculo/AAV vector to give rise to PK-7 cell line *tat* and *RD114-TR* genes introduced by VSV-G pseudotyped SIN LVV.	1.0x10^{6}	393
Wielgosz	2015	HIV SIN	VSV-G	Prod	Inducible (Tet-Off)	650MNDh WASp1	HEK 293T/17	Transfection of GRPG or GRPT-G from Throm *et al.*, 2009[292] with transfer vector concatemer	>1.0x10^{7} (HeLa cells)	424
Hu	2015	HIV (IDLV) SIN	VSV-G	Prod	Inducible (Tet-Off)	n.s	PVG3 (293)	Cells constitutively express tTA and inducible VSV-G expression. Stable transfection with *gag-pol*, *tat*/*rev*. Transfer vector transduction with cSIN vectors* and transfection**.	*5x10^{6} **2x10^{8}	463
Sanber	2015	HIV SIN	RD114-PR	Prod	Constitutive	WinPac-RD	293FT	293T MLV transduction tagging = 2G; 2G RMCE with gag-pol Subsequent transfection of *rev*, *RD114* and transfer vector	1.0x10^{7}	381
Marin	2016	HIV SIN	RD114-TR	Prod	Constitutive	RD3-MolPack	PK-7 from[393] (293T)	Transduction of PK7 cell lines (expressing *gag-pol* and *rev*) with SIN LVV with *RD114-TR* and subsequently transfection of SIN transfer vector	1.8x10^{6} (CEM A3.01 cells)	382
Humbert	2016	HIV SIN	Cocal	Prod	Constitutive	eGFP2-12 and C4 1-9	HEK 293T	Serial stable transfection of 293T cells with cocal envelope, *gag-pol*, rev and LV transfer vector (hygro, puro, blast, zeo resistance genes, respectively)	>10^{6} (HT1080 cells)	383

n.s. non-specified; Ref, reference; Pack, packaging cell line; Prod, producer cell line. Titers before concentration.

1.2 Cell line development and genome editing

1.2.1 The conventional cell line development workflow

Since the approval of the first recombinant therapeutic protein (tissue plasminogen activator, tPA) in 1986, ‘biologicals’ have gained presence in the pharmaceutical market. According to the FDA, biologicals are medical products obtained from natural sources; they are typically large complex molecules compared to chemical drugs. Vaccines, therapeutic proteins, tissues or organs and cell and gene therapiey products are examples of biologicals. Most of these products are typically manufactured in cells deriving from the same cell and with a homogeneous phenotype that can be sustained for prolonged periods in culture with denominated cell lines. Cell line development consists of the optimisation of each of the steps involved in the production of a biological in order to achieve scalable, stable (in growth rate, genetically and protein levels) and high yield production processes. The cell line development strategy for biopharmaceutical production follows an established workflow[464]: an expression cassette containing a gene of interest is introduced into (preferably suspension adapted, serum free) a suitable host cell line together with a selectable marker that will confer advantage to cells expressing the transgene. After that, selection is applied to avoid growth of cells that have not up taken any DNA. Gene amplification strategies are often introduced at this stage to increase the number of transgene copies. Selected and amplified clones are isolated and its specific productivity evaluated using high through put systems. The best performing clones are then scaled-up to fed-batch cultures and monitored for long-term productivity and stability as well as other factors (proliferation, viability, folding, and secretion) prior to cell banking. In this section of the Introduction chapter, potential problems typically encountered during the cell line development process will be extensively covered.

1.2.2 Problems in cell line development and their potential solutions

Low expression of the transgene of interest

High levels of expression of heterologous proteins suppose a burden for the host cell metabolism. Expression cassettes used for biopharmaceutical production usually contain a strong cellular or viral promoter to drive the expression of the cDNA of the gene of interest[465], terminated with strong polyA signals. In order for the mRNA to be more stable and exported to the nucleus for transcription, an intronic sequence is normally included between the promoter and the beginning of the coding sequence[466]. Other common modifications include the codon optimisation of the DNA sequence to enhance the use of tRNA codons abundant in the species[467], removal of cryptic splice sites or a more balanced GC content[468]. The selectable marker, can either be expressed under a different promoter or under the same promoter as a polycistronic mRNA using an internal ribosomal entry site (IRES)[469]; this way selection or amplification is linked to the transgene expression. However, expression of downstream genes in gene fusions separated by IRES is lower. Alternatively, 2A peptide sequences allow expression of different proteins from a single ORF separated by a picornavirus auto-proteolytic[1] 18aa motif[470].

Silencing of transgene expression

Although often attributed to cytotoxicity derived from viral proteins, instability of vector production has been a common problem associated with transfection, mainly due to gene loss or gene silencing[460]. Some authors claim gene loss becomes a problem in sustained cultures[471]. Nonetheless, other studies have shown both instability of expression despite stable copy numbers thus attributing instability to gene silencing[403], an eukaryotic mechanism to defend from foreign DNA[471]. Selection and maintenance of cells with packaging function can be accomplished by expressing packaging genes alongside with a selectable

[1] Ribosomal skip mechanism (*cis*-acting hydrolase elements) has been proposed instead of autocleavage[813].

marker gene. In the development of FLY retroviral packaging cell lines, Cosset *et* al., optimised the distance (74nt devoid of any ATG) between the stop codon of the stop codon of the *pol* gene and the start codon selectable marker in order to allow reinitiation of translation. This enabled higher expression of mRNA and a better selection of cells expressing viral proteins, leading to higher-titer vector[472]. Optimisation of expression systems can be further achieved by adding elements that protect the expression cassette from the effects derived from genetic elements located in regions proximal to the integration site. These positional effects are often associated with gene silencing occurring as a consequence of the methylation of the DNA in heterochromatic regions. *Cis*-acting elements such as chicken lysozyme[473], beta-globin[474], beta-interferon[475], scaffold/matrix attachment regions (S/MAR), insulators or ubiquitous chromatin opening elements (UCOE)[476] can be employed to maintain active chromatin. S/MARs are genomic DNA attachment points to the nuclear matrix[477], which also act as binding sites for CCCTC-transcription factor and nuclear matrix proteins[478,479]. S/MARs create a loop that maintains chromatin transcriptionally active. UCOE consists of a methylation free CpG island that keeps chromatin open in housekeeping genes[480] and showed increased levels of antibody production[481] and restore wild type phenotype when used in SIN LVV in mouse models of SCID-X1[482]. UCOE (commercialised by Merck-Millipore) not only increased protein titers up to 5-fold[483] but also the proportion of high producing clones[481].

Limitations in the cell metabolism

Increasing demands in protein production pose a metabolic and viability limitation for the host cells[484]. Protein and vector yields have also been increased through engineering host cell line homeostatic processes at different levels. Such changes have been applied to CHO cell lines for antibody production but could feasibly be applied to HEK 293 cell lines for lentiviral production. Anti/pro-apoptotic regulating factors such as Bcl2 family proteins have been expressed in host cells to delay apoptosis[485]. Type II programmed cell death (or autophagy) can also be delayed by overexpression of Bcl-xL or constitutive expressing Akt in conditions of nutrient exhaustion[486,487]. These modifications extend the productive phases of the cell cycle increasing the yield of protein production. Overexpression of $p27^{KIP1}$ and $p21^{CIP1}$ have shown to arrest the CHO cell cycle in the most prolific phase (G1)[488] although lentiviral vector production does not depend on cell cycle[449]. Metabolic engineering has also been explored to reduce the amounts of ammonia and lactate accumulated in culture, which are toxic for the cell[489]. Cells expressing high levels of glutamine synthetase can convert ammonia into glutamine in the presence of glutamate[490]. Modifications in the TCA cycle such as overexpression of pyruvate carboxylase[491] or knock-down of LDH-A with iRNA[492] have also been attempted. Folding, secretion and glycosylation profiles can also be optimised although their application is more focused to antibody development.

Other strategies such as directed evolution consist of the application of selective pressure to force selection and mutation (mimicking Darwinian engines of evolution) and ultimately improve the performance of host cell lines. Cell culture at lower temperatures (32°C) has shown increased cell volumes and higher productivities (also in HEK 293 cells)[493]. Prolonged exposure to hydrogen peroxide can be applied to enhance the tolerance to genomic instability effect[494].

Low efficiency of integration

The non-viral integration of heterologous DNA in the host cell genome can be achieved by plasmid DNA transfection. Upon exposure of foreign DNA, 10^{-3} cells (depending on the cell type) will insert that into their genome via homology dependent or independent mechanisms[495]. For an ideal DNA transfection, cells are generally recommended to be low passage (<20), high viability and mid-high confluency (40-80%). Once inside the cell, most nucleic acid molecules are degraded in the cytoplasm and only 10% of them reach the nucleus[496]. Microtubules seem to play a role in the intracellular trafficking of plasmids to the nucleus[497] but the mechanism is not clearly understood. Transfection leads to random integration or non-homologous recombination and generation of stable transfectants with a frequency of 10^{-3}-10^{-5} cells[498]. DNA remains mainly episomal[467]. Under the same conditions of density and media, expression levels were reported to vary over time depending on the number of copies together, the site of integration, silencing due to chromosomal rearrangements or methylation and the phase of cell cycle can provide heterogeneity in the clonal fitness and expression of each clone. The cell cycle is also relevant; cells transfected during the S-phase were reported to have maximum uptake and expression[499]. Regarding the quality and type of DNA, contamination of prokaryote DNA with lipopolysaccharide endotoxin carryover has been shown to be toxic for the cell[500]. Similarly, integrated DNA can also influence expression of neighbouring sequences[254].

Transfection of linearised plasmid is often advised for stable transfection of plasmid DNA as free ends are more recombinogenic[501]. However, this depends on the site of linearisation. In addition, linearisation of the plasmids adds digestion and inactivation steps with further purifying complications.

Alternatively to transfection, another non-sequence specific way to efficiently integrate plasmid DNA into host cells is using transposon systems. DNA transposons are natural genetic elements residing in the genome as repetitive sequences that translocate from a specific chromosomal location to another through a direct ‘cut-and-paste’ non-replicative mechanism. This mechanism maintains a stable copy number, is independent of cellular repair pathways, displays low immunogenicity and gene silencing and makes DNA transposons very attractive as delivery tools for gene therapy. Transposons naturally contain the transposase gene flanked by inverted terminal repeat sequences (ITR). A two-plasmid system containing the gene of interest (GOI) and selectable markers flanked by ITRs and a separate transposase is necessary to avoid uncontrolled lateral transfer of the GOI. Several systems have been used for transgenesis and mutagenesis across a wide variety of organisms from yeast to mammals: Tc1/mariner-like element, Sleeping Beauty[502]; the Medaka fish-derived system Tol2, a member of the hAT family[503] and the PiggyBac system[504]. Transposon based system present a more neutral (and safer) integration profile with a slight preference for active genes. For this reason, this technology has been combined with efficient viral delivery to generate hybrid tools for gene therapy[324,505].

Cell line development has benefited from transposon technology. The PiggyBac system reported high frequency of stable integration and enhanced productivity in CHO cells compared to conventional transfection[506]. Balasumramanian *et al.,* reported 3-4-fold increase in volumetric productivities for a Fc-fusion and a monoclonal antibody using the same system and cells[507]. Inducible systems have also been developed with PB systems to minimise the effect ot protein overexpression on cell stress and growth[508]. Ley *et al.,* showed that MAR and transposons could be combined to improve transgene expression also in CHO cells, which turns useful in low copy number expression cassettes or in cassettes lacking selectable marker[509].

Transduction is considered the most efficient (95-100% efficiency, calculated as the proportion of cells expressing the gene delivered by the virus) means of stable gene transfer as vectors possess the inherent ability to deliver the transgene into the nucleus. The presence of large portions of human genome occupied by human endogenous retroviral vectors (hERVS) reflects that viruses have successfully evolved to stably integrate into genomic positions suitable for their propagation. Far from the safety concerns such as the potential insertional mutagenesis leading to cellular transformation or the patient immune response seen in the clinic[510], in cell line development the drawbacks are theoretically limited to permissiveness of the cell line and the potential cytotoxicity. However, tropism can be modified by pseudotyping vectors with proteins to target a specific subset of cells. Several alternatives are available for different approaches depending on the tropism, intended duration of the expression and gene size.

Positional effects derived from illegitimate integration

Illegitimate integration

The majortity of exogenous DNA integrated into the host chromosome will follow a non-homology–based mechanism, also known as illegitimate integration. Upon the occurrence of a double strand break, the ratio between homology directed repair to non-homologous end joining ratios range from 4:1 to 1:10^6, being typically around 1:10^3-10^4 [511], although these ratios are subject to cell type and (possibly related) cell cycle. As previously mentioned, and their outcome in terms of sequence is less predictable than in homologous directed repair given that there is no repair template. Several studies have attempted to clarify the nature of this process as a better understanding of the factors that govern this process can provide insight for more efficient intergation. DNA that does not degrade in the cytoplasm can be modified extrachromosomally either via homologous recombination with sequences that share homology[512], mutated (indels or rearrangement)[513] or concatamerised by NHEJ mechanisms (which mutate the last 25 nucleotides of each side)[514].

The Non-Homologous End Joining (NHEJ) repair system is more common and ligates both ends of the DSB. NHEJ repair enzymes act in any order, and can function independently of one another at each of the two DNA ends being joined. NHEJ is likely to introduce indels (insertions and deletions), which can sometimes impact gene expression. For gene editing purposes, the impreciseness of NHEJ is often used to generate a frameshift mutation that disrupts gene expression and knockout (or knockdown) genes for the study of their function.

Modification can also occur once DNA has been integrated. Typically, foreign DNA integrates into a few sites displaying a 1-6bp microhomology region between copies of transgene in random orientation integrated in tandem by NHEJ[515]. The quantity depends on the genetic instability of the cell type. For example, transformed cell types show more complex integration patterns than normal cell lines or human cell lines are 30-100 times less likely to integrate exogenous DNA. Accessibility to chromatin is another key factor; 15% of illegitimate integration were reported in coding sequences, which represent not more than 2-3% of the human genome[516,517]. Interestingly, a study showed that the vast majority of illegitimate integration events occurred in AT-rich regions and close to topoisomerase recognitions sites, indicating bent regions are integration 'hotspots'[518]. After the integration event, the recipient DNA sequence has also been found to be modified[516,517,519]. Generally the consequences of integration comprise the disruption of recipient gene expression but incorporation of telomeric regions that could potentially induce chromosomal rearrangements have been described after integration of linearised DNA[520]. Methylation patterns can be also altered although their consequences are not well understood[521]. Instability can also be induced in recipient genomic loci if pericentromeric regions are introduced[522]. In addition, genomic stability is feature specific and thus is not constant accross the human genome, varies among cell types and external stimuli can also interact with unstable sites[523]. The probability of a DNA fragment to integrate into a locus that already harbours an illegitimate integration event was shown to be 100-450 times higher than in a random genomic site[524].

The idea of introducing genetic modifications *in situ* offers unique benefits: not only allows modification/restoration of the phenotype[525] but also eliminates the concerns regarding dose effect and the regulation of expression[526]. Genome editing, defined as the precise nucleotide modification of the genome, provides several distinct means for addressing the limitations of previous gene therapy approaches. Genome editing is a means of controlled mutagenesis of the genome, whether it is done through non-homologous end-joining or homologous recombination. This technology can be employed for therapeutic use by efficiently disrupting and inactivating a gene[525], precisely fixing a detrimental point mutation[527], or integrating a correct or useful genetic sequence into the cell genome[528].

The ideal gene-editing tool should feature the following characteristics: (i) high frequency of desired sequence changes in the target cell population; (ii) no off-target mutations; (iii) rapid and efficient engineering and assembly of molecules that target any site in the genome at low cost; (iv) capability for fine-tuning and regulation and (v) amenable to a packaging and delivery approach that will allow therapeutic dosing of cells and target tissues both *ex vivo* and *in vivo*.

Homologous recombination

Upon introduction of DNA into the cell, this can be integrated either by homologous recombination or illegitimate integration. Besides canonical HR, small fragment homologous recombination techniques[529] are based on homologous recombination, allow knock-out or knock-in of DNA fragments. Although targeting of specific region of the genome has been achieved for many years through homologous recombination, its frequency is low (10^{-5}-10^{-7} events per cell)[530] since it relies on naturally occurring double strand breaks and it had not been contemplated as a therapeutic alternative[531].

Homologous recombination, also called homologous (directed) repair (HR, HDR) is a less common mechanism than non-homologous end joining (NHEJ) in mammalian cells (although very common in yeast) given that the repair template (a sister chromatid) is only available during mitosis. Unlike HDR, NHEJ is not cell cycle dependent. Unlike NHEJ repair mechanism, which introduces insertions and deletions (indels), HDR maintains the sequence fidelity; the repaired DNA sequence is identical to that before the double strand break[532,533]. Normally, the sister chromatid (and rarely the homolog chromosome) is used as a template to repair the DSB but gene targeting exploits the use of external DNA to serve as a template and incorporate external or corrected DNA into the cell. Therefore, gene addition can be accomplished at in a site-specific manner if donor DNA is provided upon the generation of a DSB. In this modality, recombination of a cassette (flanked by homology arms) into desired loci of interest, typically safe harbours[534–536] enables functional gene correction, heterologous transgene knock-in or targeted transgene insertion without target gene disruption. Some groups have used this strategy to introduce tags when the cohesive sequence generated by the nuclease were known[537,538]. Chen *et al.*, also used this mechanism to generate animal models by precisely introducing point mutations[539]. Genome editing tools have also been shown to enable large chromosomal rearrangements[540,541]. However, serious concerns surround this approach since off-target DSB are susceptible to causing cancer[542].

In the context of cell line development, successful events must be favoured using antibiotic selection until stable pools can be further screened for expression. Homologous recombination has been extensively used to modify the genome of CHO cell lines to overexpress endogenous genes by the insertion of promoter sequences[543], produce antibodies devoid of fucose residues[544] and site-specifically integrate MAR to enhance protein production[545]. Capecchi and Smithies won the Nobel Prize in 2007 for their work on homologous recombination, which enabled the generation of transgenic mice, indispensable as models for medical research[546].

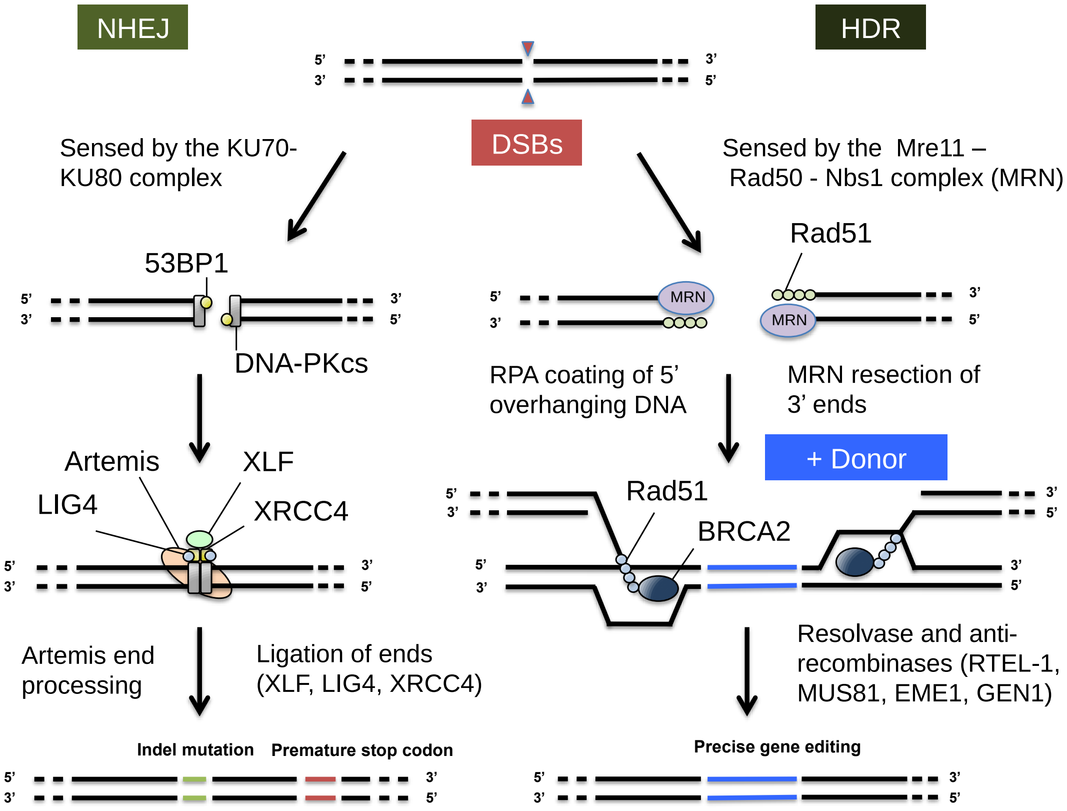

NHEJ
HDR
5'
3'
3'
5'
DSBs
Sensed by the KU70-KU80 complex
Sensed by the Mre11 – Rad50 - Nbs1 complex (MRN)
53BP1
DNA-PKcs
Rad51
MRN
MRN
RPA coating of 5' overhanging DNA
MRN resection of 3' ends
+ Donor
Artemis
XLF
LIG4
XRCC4
Rad51
BRCA2
Artemis end processing
Ligation of ends (XLF, LIG4, XRCC4)
Resolvase and anti-recombinases (RTEL-1, MUS81, EME1, GEN1)
Indel mutation
Premature stop codon
Precise gene editing

Figure 1.6. Schematic of the proteins and processes involved in DNA repair pathways.

HDR, Homologous directed repair. Initially, DNA that has suffered from a double strand break (DSB) is sensed by the Mre11-Rad50-Nbs1 (MRN) complex, which recruits host DNA repair factors through the action of ATM mediators (as in NHEJ). Then, Mre11 processes the 3' DNA ends to generate cohesive ends, and ssDNA fragments are temporarily protected from degradation by RPA coating (replication protein A). In the strand invasion step, Rad51 (together with Rad52, Rad54 and BRCA2) invades the undamaged homologous DNA creating a displacement loop, which serves as a template to synthesise the missing DNA from the 3' end. Finally, the structure is unravelled and resolved by the action of anti-recombinases and resolvases such as RTEL-1 (regulator of telomere elongation helicase 1), MUS81 (crossover junction endonuclease), EME1 (Essential Meiotic Structure-Specific Endonuclease 1) and GEN1 (Holliday Junction 5' Flap Endonuclease) and the generated strand gets replicated to generate the second strand.

NHEJ, Non-homologous end joining. The DSB is sensed by KU70–KU80 heterodimer complex which recruits p53-binding protein 1 (53BP1) and DNA-dependent protein kinase (DNA-PKcs)[547,548]. These proteins as well as Ataxia telangiectasia mutated (ATM)-mediated DNA stabilisation via phosphorylation of histone H2A.X (also recruited in the HDR pathway)[549] prevents degradation of DNA ends. End processing is by Artemis and subsequently DNA ligase 4 (LIG4), X-ray repair cross-complementing protein 4 (XRCC4) and XRCC4-like factor (XLF) participate in the final ligation of both ends.

Nuclease based approaches

Nuclease based approaches pursue the introduction of DSBs, which increases the efficiency of recombination by more than two orders of magnitude[550]. After nucleases cut the DNA, the host's DNA repair machinery (NHEJ and HDR, always active in eukaryotic cells) is employed to enact repairs that can be combined with integration of donor genetic material. In the last years, a number of easily accessible, relatively simple and highly specific tools have emerged to enable genome engineering in many ways for multiple applications. Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR), Zinc Fingers Nucleases (ZFN)[551,552], MegaNucleases (MN)[553,554], Transcription Activator-Like Effector Nucleases (TALENs)[555] are novel technologies that have changed the paradigm of genome editing/engineering for multiple applications. In the frame of cell line development, ZFNs have been used to double knockout of the DHFR[556], GS and FUT8[557] gene to maximise genetic amplification or for the disruption of Bak and Bax pro-apoptotoc proteins[558]. Cell line development applications using TALEN genome editing technology are less numerous due to its relatively complex assembly and longer timelines. Knockout of α(1,3)-fucosyltransferase (FucT) and the two β(1,2)-xylosyltransferase (XylT) plant genes is an example of potential applications of this technology to generate proteins with a more mammalian glycosylation pattern[559]. However, In this project we have used the CRISPR system (that will be discussed in more detail but the reader is referred to the following reviews on the other genome engineering technologies[560,561].

- *<u>Clustered Regularly Interspersed Palindromic Repeats / CRISPR-associated protein (CRISPR/Cas)</u>*

CRISPR/Cas proteins naturally function as an adaptive immunity system in bacteria and archaea, to defend the organism against foreign nucleic acid sequences[562–566]**.** The bacterial immunity function observed in the CRISPR/Cas systems can be divided in two phases: (i) adaptation, where a segment of the foreign DNA is excised and incorporated in the CRISPR array as a protospacer (mainly carried out by Cas1 and Cas2)[567,568] and (ii) effector, where the pre-crRNA (Crispr RNA) is expressed, processed to mature crRNA[569–572] and mediates complex with the Cas9 protein that will cleave foreign sequences[573,574]. This second phase encompasses several CRISPR systems according to the composition of their effector function[575,576]. Type I and type III (as well as the putative type IV) belong to class 1 CRISPR systems and mediate their interference through multiple proteins and Csm/Cmr effector complexes[575,577]. Unlike class 1 systems, class 2 systems employ a single effector (Cas) protein to cut foreign DNA and include type II and type V CRISPR Cas systems[570,578,579]. Cas9 protein (950-1,600aa depending on the species) is the effector protein of class 2 type II CRISPR system and possesses a RuvC-like and a HNH nuclease domains[580].

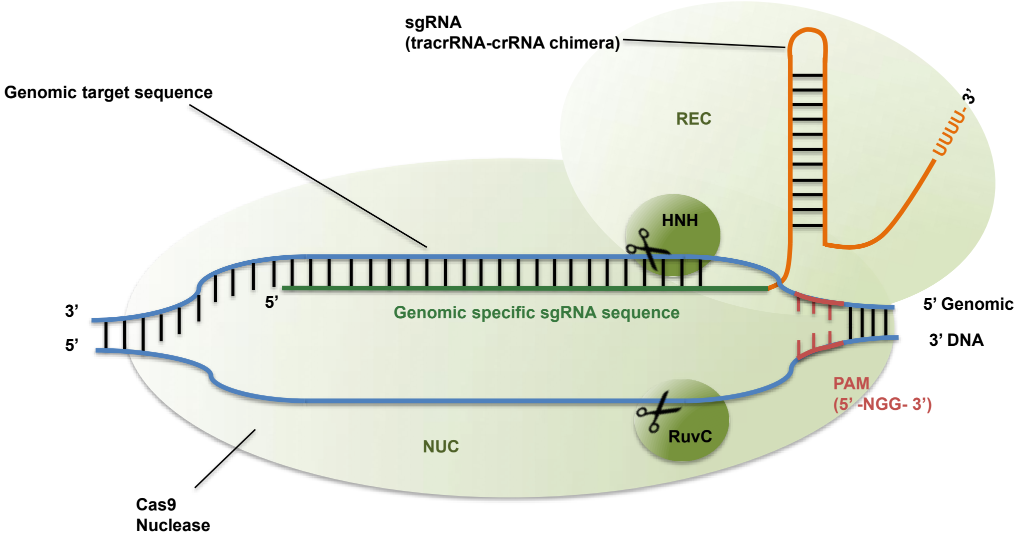


Figure 1.7. Schematic of the RNA-guided genome editing CRISPR-Cas9 nuclease system.

REC, recognition subunit; NUC, nuclease subunit; PAM, protospacer adjacent motif; HNH and RuvC are endonuclease domains named due to the critical His-Asn-His residues and *E.coli* DNA repair protein, respectively.

S.pyogenes CRISPR-Cas9 system was engineered to induce site-specific DSB into the host genome and enable DNA edition using the host cell DNA repair machinery. The SpCRISPR-Cas system consists of the Cas9 nuclease and a single guide RNA (gRNA). Guide RNA is composed of a tracrRNA (transactivating) and the crRNA. In 2012, Jinek *et al.*, showed successful fusion of these two components into a single-chain guide RNA (sgRNA), which simplified the system even further[579]. The sgRNA hybridises with a complementary 20bp sequence of the sgRNA (known as the protospacer), which is preceding a species-specific protospacer adjacent motif (PAM) and directs the nuclease to introduce a DSB between the 17th and 18th position of the sgRNA complementary sequence. This will trigger the host cell DNA to edit the DNA via NHEJ (error prone, introducing indels) or HDR (precise repair, used to integrate sequences).

In terms of design, CRISPR-Cas9 system allows targeting of virtually any sequence in the host genome. The ability to redirect the CRISPR/Cas system to new target sites by only swapping the 20 base pair targeting sequence of the gRNA is a significant advantage compared to MN, MegaTAL, ZFN and TALEN systems due to its simplicity of design, inexpensiveness and multiplexing potential. The only constraint is that the desired cleavage site must be located immediately upstream from PAM (protospacer adjacent motif). These three nucleotides are specific to each bacterial species from which the Cas9 and gRNA are derived. In the case of the standard CRISPR-Cas9 system, derived from *Streptococcus pyogenes*, the sequence typically used is 'NGG'. SpCas9 also cuts upstream a 'NAG' PAM, although cutting efficiency is reduced to one fifth[581].

Despite being a recent editing system, the CRISPR-Cas9 has been used for multiple application such as high through put functional screening[582], labelling of several loci (CRISPRainbow)[583]. Knock-in approaches have been used to add foreign DNA into the genome of mouse and human embryonic stem cells[584,585], mouse embryos[586,587] or stem cells to generate transgenic animal models[588] or even CHO cells for cell line development[589]. Interestingly, CRISPR-Cas9 system has also been used to prevent HIV proviral replication. However, while some indels block viral replication others mutate the sgRNA recognition of target sequence and allow the virus escape contributing to the generation of resistant viruses[590]. Like other nucleases, CRISPR-Cas9 is able to promote genome rearrangement and could be used as a tool to study cancer[540].

In June 2016, the first CRISPR clinical study received approval by the NIH advisory committee for treatment of cancer. The *ex-vivo* treatment, manufactured by University of Pennsylvania, consists of a triple edition of patient T cell genome[591].

Examples of CRISPR-Cas9 in cell line development include the knock-in of a 3.7kb reporter cassette containing mCherry gene and a neomycin selectable marker into the COSMC locus[589] or a FUT8 knockout case study by WuXi Biologics[592].

Recombinase based approaches

Alternatively to HDR or nuclease-based methods, which are triggered by the presence of DSB in the DNA, recombinases catalyse cleavage and reunion between specific sequences or recognition sites of the target DNA molecule and can lead to insertion, deletion or inversion of DNA fragments depending on the orientation of the recombinase recognition site[593]. According to whether their active nucleophilic aminoacid residue in the catalytic domain is a Tyr or a Ser, recombinases can be classified in two families (serine and tyrosine recombinases). Serine recombinases (or resolvases) catalyse irreversible reactions and can be further splitted into small (excision, identical recognition sites) and large (excision, inversions, integration, non-identical recognition sites). Tyrosine recombinases can be uni- or bi-directional and both can catalyse excision, inversion and integration although the former acts on identical recognition sites and its recombination is reversible and the reaction catalysed by the latter is irreversible and acts on non-identical recognition sites[593].

Although the basis of the recombination is a series of transesterification reactions, the mechanism of recombination differs between serine and tyrosine recombinases[594]. Serine recombinases cleavage and strand transfer occurs at the same time with all four ends bound to the protein. Cleavage of DNA strands by tyrosine recombinases occurs and intermediate structures are resolved via the Holliday junction pathway[595]. Most of the recombinase-mediated cassette exchange (RMCE) strategies performed in the last years employ one of the following systems:

The ΦC31 (or R4) /attB-attP system is an example of a large serine recombinase and mediates recombination between distinct recombinase recognition sites. Attachment sites (att) in phage and bacterial sequences (or 34bp *attB* and 39bp *attP*)[596] allow unidirectional irreversible recombination by the *Streptomyces* phage ΦC31 or the R4 integrases (serine recombinases[597]) generating *attL (*left) and *attR* (right) sites in human and mouse cells[598]. In addition, the existence of endogenous attP-like sequences or 'pseudo' attP sites[599,600] opened the door to gene therapy although it also raises concerns for their potential genomic rearrangement.

The Cre/loxP system derived from the bacteriophage P1[601,602] is composed of a 38kDa protein that 'causes recombination' and a 34bp target sequence (locus of X-over of P1) consisting of a 8bp spacer sequence flanked by 13bp inverted repeats (or palindromic arms). This system has been exploited for recombinase-mediated cassette exchange (RMCE), which follows two steps: (i) random integration or site-specific HR-mediated introduction of the recombinase recognition site (a landing platform) and (ii) retargeting of the initially targeted site with the cassette of interest flanked with recombinase recognition sites. Efficient integration was achieved using this strategy[603,604]. Cre and ΦC31 recombinases, toxic for the cell[605,606], were only expressed transiently. Nevertheless, the first generation of these strategies suffered from reversibility and incorporation of bacterial sequences in the host chromosome. The composition of the recombinase recognition site was studied to give rise to mutants that allow unidirectional and irreversible[607–609]. The most common Cre/loxP spacer mutants (lox511 and lox2272[610,611]) and arm mutants (lox61[612] and lox71[607]) supress the ability of the recombinase to revert (excise) the integration. Not only mouse transgenesis[613] but also antibody (anti-RhD in CHO)[614] and retroviral vector production (2x10^{7} TU/mL in HEK 293)[391,392] also have benefited from the second generation RMCE. The *S.cerevisiae* Flp/FRT system[615] (or its thermostable improved version Flpe/FRT[616]) is also a bidirectional tyrosine recombinase. Similarly to the Cre/loxP system, the second generation RMCE using F_3 and F_5 showed reduced excision of recombined sequences[617].

Kameyama et al., and later Obayashi et al., perfected the Cre/loxP system by allowing irreversible serial accumulative and unidirectional retargeting of cassettes after targeting of two and one (respectively) loxP sequences[614,615]. However, these heterologous recombinases rely on the pre-existing recombinase recognition sites (previously introduced via HR) and as a consequence, the number of potentially targeted genomic positions is limited.

Custom hybrid recombinases can be generated by combining a catalytic domain with invertase/resolvase function with a ZFP or a TALE domain[618]. The catalytic domain of the recombinase recognises a 20bp core sequence that is flanked by 2 ZFP or TALE binding sites. Given the cooperative nature of the enzyme specificity, different recombinase catalytic domain variants (Gin α, β, γ, δ, ε, and ζ, with different specificities[619]) from bacteriophage Mu DNA invertase Gin that contact with DNA dimers have been identified and can be modularly 'mixed and matched' (using plasmid assembly systems such as OPEN[620] or CoDA[621]) to yield recombinase variants with distinct specificities.

Wide range of productivities of selected cells

Growth rates and specific productivities of clones resulting from random integration are highly heterogeneous. Intrinsic interclonal variation combined with acquired drift after isolation are thought to explain this phenomenon[622]. A higher stringency can augment the effect of selectable marker and reduce the number of subsequent screening. Stringency is the degree of selective pressure applied to cells post-transfection (or transduction) with the objective of killing clones that have not taken up any copies of transgene DNA. Stringency can be also understood as the ratio between antibiotic uptaken vs detoxifying counteracting measure taken by the cell. Thus, low producers can be also eliminated by increasing the stringency. However, despite increasing productivity, an increase in the stringency of selection can affect cell growth if antibiotic concentrations are too high[623]. Stringency can be increased during antibiotic selection if the selectable marker is attenuated. This way, the expression of the GOI and selectable marker has to be higher to overcome the selection process and stringency can be increased at lower concentrations. This can be achieved by using AU-rich elements (AREs) to promote selectable marker mRNA degradation or polypeptide regions rich in proline (P), glutamic acid (E), serine (S), and threonine (T) (PEST) regions (from the C-ter murine ornithine decarboxylase), which destabilise selectable marker proteins[624]. Selection stringency can be also improved by using a weak promoter such as the SV40 early or non-optimal IRES can be used in combination with the selectable marker. Alternatively, codon deoptimisation[625] or mutation of the selectable marker have also been attempted reporting the latter a 10-fold increase in recombinant protein titer[626]. For example, clones mutated neomycin phosphotransferase II display reduced affinity for the antibiotic, which promotes overexpression of the selectable marker to survive[627]. In addition, the probability of isolating a high producer has also been demonstrated to be higher if the selectable marker is attenuated. Positioning of the selectable marker downstream of the gene of interest is also critical to colocalise selection events and mitigate the effects of gene fragmentation of bicistronic cassettes, a phenomenon that occurs upon stable transfection[628].

Low copy number of integrated vector

As a result of illegitimate integration derived from transfection, expression of the gene of interest driven by an uncontrolled low number of copies may be unsufficient for production goals. Alternatively, low levels of expression can be due to recombination-mediated reduction of the number of 'head-to-tail' integrated copies of transgene. Mammalian cell lines have the ability to undergo genomic rearrangements and increase the copy number of resistance genes upon increasing concentrations of selectable marker. This phenomenon, known as genetic amplification, was first observed in cancer cells treated with increasing concentrations of a chemotherapeutic drug (methotrexate, MTX)[629] and can be used in cell line development to increase the copy number of integrated vector. Gene amplification strategies are based on the co-transfection of the gene of interest alongside a selectable marker i.e. dihydrofolate reductase (DHFR). DHFR is an endogenously produced protein, which can be blocked by MTX[630,631]. However, after 2-3 weeks of culture with the drug in auxotrophic conditions (GHT-minus media) that prevent synthesis of thymidylic acid and purines through alternative pathways, surviving cells contain several hundreds of copies of the DHFR gene integrated in the chromosomes alongside the gene of interest[632], resulting in 10-20 fold increase in the specific productivity[633]. Adenosine deaminase (ADA)/ 2-deoxycoformycin system[634] or the glutamine synthetase/methionine sulfoximine (GS/MSX) system work in a similar fashion[347]. Knock-out of endogenous copies of these genes has been successfully attempted to enhance the effect of the gene amplification[635].

High analytical burden of screening clones

The frequency of high producing clones is rare (10^{-3})[636], as insertion into a high transcribing site is not common and usually expression imposes a metabolic burden for the clones. Therefore, hundreds or thousands of clones are typically screened, which involves a labour-intensive process that can take up to 6-12 months[355,637]. Often clones with higher productivity also show slower growing rates[638,639]. Ideally, high productivity and growth but also stable production is desired for scale up.

Traditional screening methods

Limiting dilution cloning (LDC) is typically used to isolate clones because of its simplicity, reliability and cost[640]. A few hundreds individual clones are individually seeded at low densities so that each well only contains a single cell. The best producer clones are assayed for specific productivity by ELISA and a second round of cloning is required to ensure clonality since statistical analysis indicates it may not be guaranteed[641]. Imaging systems can assist with this on that matter. Therefore, a process like this is highly resource- and time-consuming given that it can take several months and only a few hundred clones can be screened.

FACS-based methods

Flow cytometry and cell sorting is a relatively inexpensive high-throughput alternative that can help increase the number of screened clones and can be combined with LDC. Briefly, a laser interrogates the florescence of the protein or cell and an electric charge is applied to the droplet to segregate different subpopulation. Cell sorters can reach speeds up to 10^8 cells per hour[642]. Clone productivity can be screened by analysing the expression of a surface antibody or proteins on the cell membrane [643]. Alternatively, bicistronic expression of the gene of interest and CD20 ligand (a protein that is not normally expressed in the cell membrane) separated by IRES can be used to assess gene expression. When independently translated, fluorescent antiCD20 allowed accurate correlation of antibody production[644]. Nevertheless, the ligand production is a metabolic burden for the cell and heterogeneity in the fluorescence levels. If the product is not expressed in the membrane, GFP can be used as a reporter gene. Meng *et al*., and Mancia *et al*., and others demonstrated that a correlation exists between the level of GFP fluorescence and the expression of heterologous recombinant protein when co-expressed using two promoters[645,646]. These findings led to the use of GFP as a reporter gene for generating stable, high-expressing cell lines and proteins by gating for high eGFP producers by FACS[644–646]. Yoshikawa *et al.,* described a method based on intracellular fluorescently labelled methotrexate that can quantitatively penetrate the membrane and bind DHFR[647]. Resistance to MTX is proportional to *dhfr* copy number, which was expressed along with the gene of interest.

Secretion-based assays

Unlike previously described methods, gel microdrop technology directly evaluates protein expression at a single cell level. A cell sorter distributes individual cells into a gel microdrop of biotinylated agarose that is linked using avidin bridge to an antibody that specifically recognises the secreted protein. Once the protein of interest is bound to the capture antibody, a fluorescently labelled antibody binds it and allows detection of the secreted protein[648]. The main pitfall is the reduced capacity of this method as only 10% of the droplets contain a single cell[649].

Affinity matrix-based reaction assays follow the same principle as the gel microdrop technology although cells are biotinylated and bridged to a capture antibody. Secreted proteins are then sandwiched between capture and detection antibody (with a FITC molecule conjugated). Cells are cultured in high-viscosity low permeability medium and high producers can be subsequently sorted and isolated. Reduced timelines and 5-fold increase in titers were reported using this method[650].

Automated systems

Laser-enabled analysis and processing LEAP™ (Cyntellect) is a high-throughput automated screening technology that picks adherent and suspension cells[651] immobilised in a capture matrix based on secretion using a specific antibody. The laser is used to eliminate undesired neighbouring cells and reduce heterogeneity resulting in a 20-fold increase in productivity[652]. The main drawback of this method besides its cost is the potential damage occasioned to the cell[636]. The Cell Xpress system combines the LEAP technology with multicolour live imaging, specific detection reagents and a fully automated close contained environment to minimise the risk of contamination[652].

Other systems such as CellCelector™ (Aviso) and Clone Pix™ (Genetix) also use semi-solid medium to immobilise cells and limit diffusion of the secreted protein, which immunoprecipitates with the fluorescently labelled capture antibody and forms a halo[653]. The Cello™ closed system (TAP Biosystems) integrates automatic cell culture, microscopy and analytic devices (Sonata, Piccolo, CElloSellecT) combined with high-throughput robotics and advanced software. Transfected cells are introduced into the system, which automatically seeds them into plates. Although the sophistication of this system is translated into expensive equipment, up to 800 plates can be processed in parallel[636].

Most high through put selection methods reviewed in this section rely on the use of fluorescent antibody and semisolid media combined with sophisticated automated systems, which increases screening costs. The publication of the human[229] and the CHO genome[654] boosted comparative genomics to identify genes associated with cell growth and productivity. Predesigned cDNA microarray CHIPS[655,656] allow high through put determination of expression levels although commercial arrays are only able to detect certain standard genes. Recently, whole transcriptome analysis opened the door to high-throughput screening of expression levels at a genomic level. However, expression values are gene-specific and the position effect of the integration of the transgene is not always reflected in this analysis. Therefore, there is a need for a simple high-throughput system with higher accuracy than FACS that can account for expression and site of integration and correlate this to transgene expression levels.

As stated before, in the context of cell line development, the analytical burden is intense given the heterogeneity of the clones even though selection strategies are employed. In an academic setting, where most gene therapy projects originate, large amounts of screening resources may not be available. The integration properties of lentiviral vector, enhancing site-specific integration into pre-defined loci could also help to bring production of high-performing clones within the reach of academic laboratories. This project presents a potential solution to this problem using barcode-mediated selection of high producing clones.

1.2.3 A solution for simple high-throughput screening: a barcode-based method

Screening based on genetic tags (i.e. ESTs) was originally accomplished using DNA microarrays[657,658], but became progressively replaced with next-generation sequencing[657,659–662]. The latter allows for more quantitative, high-throughput and accurate data analysis. However, at that time, high through put sequencing technology also presented inherent limitations in terms of costs and number of low complexity samples sequenced in parallel (capacity). Physical segregation of samples into different lanes allows limited multiplexing and limits the number of conditions to be tested and involves higher processing times and costs. Parameswaran and Meyer introduced the barcode technology (also known as bar codes or bar-codes) that enables efficient tagging multiple samples run in parallel[663]. DNA barcoding is an innovative technology that consists of marking samples with a specific DNA sequence tag so that they can be pooled together in the same sequencing lane/run. Although barcoding was initially applied to enable routine parallel processing of multiple sequencing datasets, multiple groups implemented the barcode system to specifically tag and retrieve samples of different biological origin (cells, lineages, tissues) maximising the multiplexing potential of this technology. In this study, sequencing barcoding will be referred as indexing, in order to distinguish it from clonal cell marking barcodes or *cellular barcoding*.

In the last years, cellular barcoding has been used for efficient and quantitative monitoring of clonal dynamics and spatial distribution of integration sites during gene correction of hematopoietic stem cells in clinical trials[661,664]. In this context, vector-host chromosome junctions are retrieved using integration site analysis techniques such as ligation-mediated PCR (LM-PCR). DNA barcoding has been consolidated as an inexpensive, relatively simple and powerful method that allows sample multiplexing and has been used in multiple applications including not only characterisation of clonal dynamics of hematopoietic, bacterial populations[665] and discernment of cell lineages but also to label different sources of RNA for Cap Analysis of Gene Expression (CAGE)[666] or even for the identification of rare HIV drug resistance mutations[667].

Barcode library design, quantification of complexity and error correction

Although the construction of DNA libraries is reported to be a slow and laborious process due to the relative inefficiency of ligation based methods, the library can be used for several applications[668–671]. The construction of a nucleotide library appears to be a simple and inexpensive task consisting of subcloning a string of "N"s into a vector backbone. Nonetheless, several aspects need to be taken into account when generating a randomized sequence tag.

The number of possible combinations makes multiplexing practically limitless, if one considers the number of variants equivalent to the fourth power of each random nucleotide in the DNA stretch. However, the length of the variants also influences the complexity or theoretical diversity of the library. The number of edits needed to transverse or transit from a barcode variant to its nearest neighbour is a critical. This parameter is known as the Hamming distance (if insertions and deletions are not considered, in which case would be Leveinshtein distances). In the case of libraries with low Leveinshtein distance, false barcodes can be more easily generated from polymerase or next-generation sequencing errors, leading to misrepresentation of the actual library complexity. The mean number of dissimilarities between variants is a trade-off and can be modulated when playing with two of parameters: sequencing depth and length of the barcode. Higher sequencing depths enable the analysis of longer barcodes and, in theory, higher complexity libraries. However, they are directly linked with a larger number of misreads, which contributes towards less dissimilarities, making the clusters less differentiated and distinguishable and thus decreasing the library complexity. The length of the barcode can also play an important role in the library design and is dependent on the throughput of the application. Shorter barcode lentghs reduce the entropy introduced in the system and diminish the mean number of dissimilarities observed between the variants of a library.

Different transduction protocols have been described to minimise the chance of biasing the fate of progenitor cells. While cell fate does not pose a problem in this study, multiple integration of barcodes into one cell and repeat usage should be considered. Integration of multiple barcodes into one cell can be minimised by using lower transduction efficiencies/multiplicity of infection and evaluated by vector copy number qPCR. In general, it does not influence the outcome since each integration site can be individually retrieved even in the same cell.

A more significant concern arises if multiple cells or integration sites have the same barcodes, which is known as repeat usage of barcode variant. This issue might have been originated during the library preparation, transduction or intrinsically in the number of cells or the library size. This will result in ambiguous assignments and the loss of a biological relevance and constitutes an experimental parameter that requires to be optimised. In theory, the probability of a barcode variant tagging two different cells is negligible. However, it is generally considered best to limit the number of cells to be tagged to be 10% of the library sample space[672].

Similar approaches

Semiquantitative information can be extracted from barcode quantification derived from sequencing or microarray data. Contribution to lineages can be extrapolated from the read representation of a particular barcode variant, although those variants with less counts can be underestimated[662]. However, quantification of expression derived from barcode integration in a specific locus constitutes a different approach. Filion's laboratory recently introduced the TRiP technology, which combines Sleeping Beauty transposase to drive the integration of plasmid library (generated by barcoding PCR) containing a reporter gene[673]. Co-transfected cells express the reporter gene and barcode at different levels subject to their integration site and the barcode allows for correlation of barcode counts with their position in the genome. Barcoded plasmid integration sites are analysed by inverse-PCR and coupled to mRNA transcripts, which contain the barcode within the GFP ORF to measure their individual expression by quantitative high-throughput sequencing. The application of this technology in their lab is to study expression in different chromatin context.

The CellTracker® technology released in 2013 by Cellecta, Inc (when this project was initiated) also shares some of the same principles with this project. A library consisting of 50 million barcodes is stably integrated into a starter founder population of cells using lentiviral vectors[674]. The tag is passed onto the cell progeny upon cell replication; and a red fluorescent protein (RFP) and puromycin marker are also included in the lentiviral vector to help maintain selection of barcoded cells. This system has been applied to cancer cells to monitor differentiation over the course of drug treatment[216, 217]. However, the location of the barcode (in the middle of the lentiviral vector backbone, Appendix A) does not enable any association of barcode counts to a particular integration site.

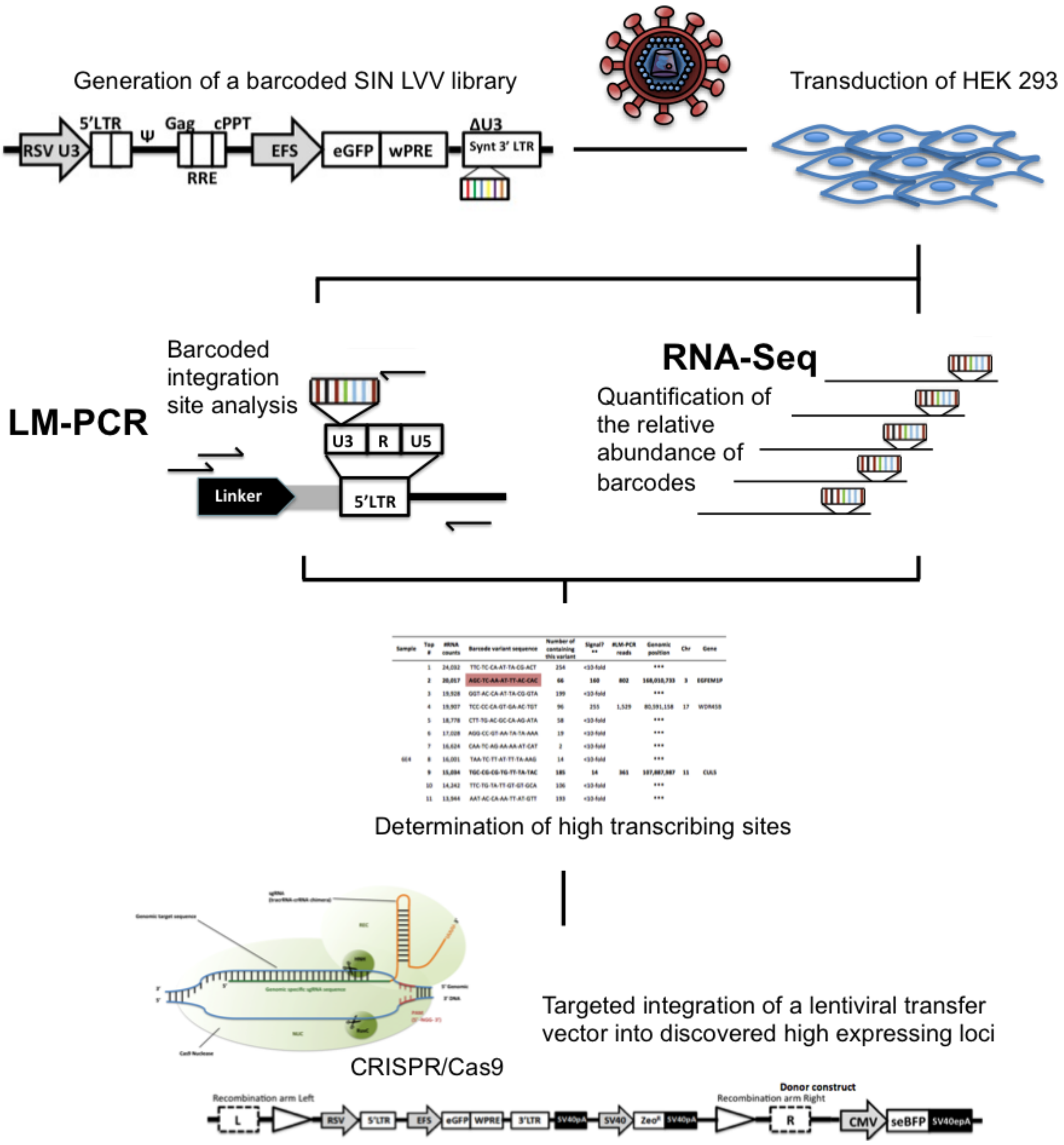
Hypothesis and objectives

This project intends to prove the following scientific hypothesis:

> *"It is possible to use cellular barcoding as tools to identify high transcribing regions derived from lentiviral integration in host cell lines. Such loci can prove useful to insert lentiviral components for packaging cell line development."*

In order to address this hypothesis, a lentiviral vector library containing multiple DNA sequence tags (barcodes) will be generated. HEK 293 host cell lines will be transduced at a low MOI to allow integration of the barcoded provirus into their genome. Chromosome-vector junctions containing the barcode will be identified using ligation-mediated PCR (LM-PCR) and next-generation sequencing. In parallel, barcoded RNA reads from transduced cells will be analysed via RNA-Seq. The RNA reads with higher number of barcode counts will be correlated to a genomic position. Finally, a donor plasmid containing a lentiviral transfer vector will be specifically targeted into such loci using CRISPR-Cas9 genome editing technology and titers of the resulting packaging cell lines assessed.

Graphic abstract


Generation of a barcoded SIN LVV library
RSV U3
5'LTR
Ψ
Gag
cPPT
RRE
EFS
eGFP
wPRE
ΔU3
Synt 3' LTR
Transduction of HEK 293
LM-PCR
Barcoded integration site analysis
U3
R
U5
Linker
5'LTR
RNA-Seq
Quantification of the relative abundance of barcodes
Determination of high transcribing sites
CRISPR/Cas9
Targeted integration of a lentiviral transfer vector into discovered high expressing loci
Donor construct
Recombination arm Left
Recombination arm Right
L
RSV
5'LTR
EFS
eGFP
WPRE
3'LTR
SV40pA
SV40
Zeo^R
SV40pA
R
CMV
seBFP
SV40epA

Sample	Top #	#RNA counts	Barcode variant sequence	Number of containing this variant	Signal? **	#LM-PCR reads	Genomic position	Chr	Gene
	1	24,032	TTC-TC-CA-AT-TA-CG-ACT	254	<10-fold		***		
	2	20,017	AGC-TC-AA-AT-TT-AC-CAC	66	160	802	168,010,733	3	EGFEM1P
	3	19,928	GGT-AC-CA-AT-TA-CG-GTA	199	<10-fold		***		
	4	19,907	TCC-CC-CA-GT-GA-AC-TGT	96	255	1,529	80,591,158	17	WDR45B
	5	18,778	CTT-TG-AC-GC-CA-AG-ATA	58	<10-fold		***		
	6	17,028	AGG-CC-GT-AA-TA-TA-AAA	19	<10-fold		***		
	7	16,624	CAA-TC-AG-AA-AA-AT-CAT	2	<10-fold		***		
GE4	8	16,001	TAA-TC-TT-AT-TT-TA-AAG	14	<10-fold		***		
	9	15,094	TGC-CG-CG-TG-TT-TA-TAC	185	14	361	107,887,987	11	CUL5
	10	14,242	TTC-TG-TA-TT-GT-GT-GCA	106	<10-fold		***		
	11	13,944	AAT-AC-CA-AA-TT-AT-GTT	193	<10-fold		***		

Chapter 2

MATERIALS and METHODS

2.1 Materials

2.1.1 General reagents

Table 2.1. List of general reagents used in this study.

Reagent	Manufacturer
Restriction enzymes	Thermo Fisher Scientific (Waltham, MA, USA)
1kb Plus DNA ladder	Thermo Fisher Scientific
SYBR Safe DNA gel stain	Thermo Fisher Scientific
Agar	Sigma-Aldrich (Dorset, UK)
Vegitone lysogeny broth	Sigma-Aldrich
Tryptone	Sigma-Aldrich
Yeast extract	Sigma-Aldrich
Agarose	Sigma-Aldrich
Molecular biology grade water	Sigma-Aldrich
10x Orange G DNA loading buffer	BioVision (Milpitas, CA, USA)
Ultra Pure 10x TAE buffer	Thermo Fisher Scientific

DNA polymerase I large (Klenow) fragment	New England Biolabs (Ipswich, MA, USA)
dNTP Mix	Promega (Manheim, Germany)
Dimethyl sulfoxide (DMSO)	Sigma-Aldrich
Ampicillin	Stratagene (La Jolla, CA, USA)
Kanamycin	Sigma-Aldrich
FastAP thermosensitive alkaline Phosphatase	Thermo Fisher Scientific
Gey's balanced salt solution	Sigma-Aldrich
Hank's balanced salt solution	Sigma-Aldrich
Trypan Blue	Sigma-Aldrich
S.O.C medium	Thermo Fisher Scientific
Platinum quantitative PCR Supermix–UDG with ROX	Thermo Fisher Scientific
T4 DNA ligase	Thermo Fisher Scientific
Polyethylene glycol (PEG) 4000	Thermo Fisher Scientific
Glycerol (≥99%)	Sigma-Aldrich
Ethanol (≥99%)	Sigma-Aldrich
Isopropanol (≥99%)	Sigma-Aldrich
Polyethylenimine (PEI)	Sigma-Aldrich
T4 Polynucleotide kinase (PNK)	Thermo Fisher Scientific
Disposable scalpels	Swann-Morton (Sheffield, UK)
Cloning rings	Sigma-Aldrich
PIPES[2] (≥99%)	Sigma-Aldrich
Calcium chloride ($CaCl_2$, ≥93%)	Sigma-Aldrich
Potassium chloride (KCl, ≥99%)	Sigma-Aldrich
Sodium azide (NaN_3, ≥99%)	Sigma-Aldrich
Manganese dichloride ($MnCl_2$, ≥99%)	Sigma-Aldrich

[2] 1,4-Piperazinediethanesulfonic acid, piperazine-1,4-bis (2-ethanesulfonic acid), piperazine-N,N′-bis (2-ethanesulfonic acid)

Potassium hydroxide (KOH, ≥85%)	Sigma-Aldrich
Sodium chloride (NaCl, ≥99%)	Sigma-Aldrich
Magnesium chloride ($MgCl_2$, ≥99%)	Sigma-Aldrich
Magnesium sulphate (Mg_2SO_4, ≥99%)	Sigma-Aldrich

2.1.2 Reagents for PCR and qPCR

Table 2.2. List of reagents used for PCR and qPCR in this study.

Reagent	Manufacturer
Q5 High-Fidelity DNA polymerase	New England Biolabs
One-Taq DNA polymerase	New England Biolabs
Oligonucleotide primers	Thermo Fisher Scientific
Oligonucleotides probes	Thermo Fisher Scientific

2.1.3 Kits

Table 2.3. List of kits used in this study.

Kit	Manufacturer
Cell line Nucleofector Kit V	Lonza (Slough, UK)
DNeasy Blood and Tissue Kit	Qiagen (Manchester, UK)
EndoFree Plasmid Maxi Kit	Qiagen
QIAprep Spin Mini prep Kit	Qiagen
QIAquick Gel Extraction Kit	Qiagen
QIAEXII Gel Extraction Kit	Qiagen
QIAquick PCR Purification Kit	Qiagen
RNeasy Plus Mini Kit	Qiagen
Topo-TA PCR Cloning Kit	Thermo Fisher Scientific
Profection Mammalian Transfection Calcium Phosphate	Promega (Madison, USA)
DNA-*free DNA* removal Kit	Thermo Fisher Scientific
RNA Clean & Concentrator-5	Zymo Research (Cambridge, UK)

2.1.4 Parental plasmids

Third generation lentiviral vector pRRL.SIN.cPPT.PEW was originally obtained from Didier Trono's laboratory[677]. The third generation lentiviral plasmid originally included the enhanced GFP reporter gene (eGFP) under the control of the phosphoglycerate kinase (PGK) promoter and contained the Woodchuck hepatitis virus post-transcriptional regulatory element (WPRE) downstream of eGFP. Plasmid map is shown in Appendix A.

The EF-1 alpha short promoter (EFS) was obtained from a third generation lentiviral vector expressing human alpha-iduronidase, pCCL EFS hIDUA, from Axel Schambach and Chris Baum's group. Plasmid map is shown in Appendix A.

The zeocin resistance cassette cloned in the AvrII site of the pRRL.SIN. SyntLTR.cPPT.EEW was obtained from pcDNA3.1Zeo(+) (Thermo Fisher Scientific, V860-20). Plasmid map is shown in Appendix A.

GeneArt constructs for CRISPR-Cas9-mediated knock-in contain two 800bp homology arms (for genomic positions described in Chapter 5) and recombinase recognition sites (attB, loxP and FRT) flanking a multi-cloning site. Unlike the parental plasmid for EMX1 donor, the plasmid that gives rise to the EGFEM1P and CUL5 donor constructs contains a blue fluorescent protein gene under the control of the CMV promoter and upstream the SV40 early polyA signal. Plasmid map is shown in Figure 5.1.

pTelo plasmid was obtained from GSK Cell and Gene Therapy Lab and was used as the plasmid only contains pBR322, Kanamycin resistance gene, a lacZ reporter gene and a short non-coding human telomerase amplicon was used as a qPCR standard. The plasmid map is shown in Appendix A.

CRISPR-Cas9 plasmids were obtained from Sigma-Aldrich. For EMX1 genomic position, two separate plasmids (pCMV-cas9 and pU6-gRNA, both with Kanamycin resistance) encoded the expression of the Cas9 protein and the EMX1 sgRNA, respectively. In the case of EGFEM1P and CUL5 genomic positions, the sgRNA+Cas9 were expressed from the same plasmid pCMV-Cas9-RFP (the Cas9 downstream of the sgRNA). Expression of Cas9 can be assessed by RFP fluorescence as this marker is fused with the Cas9 protein using a 2A element. Plasmid maps are shown in Figure 5.2.

2.1.5 Bacterial strains (*E. coli)*

Stbl3 cells Thermo Fisher Scientific

Genotype: *mcrB mrr hsdS20 (rB-, mB-) recA13 supE44 ara-14 galK2 lacY1 proA2* rpsL20(StrR) xyl-5 λ- leu mtl-1 F

DH10-beta cells (Dong Hanahan laboratory) New England Biolabs

Genotype: *Δ(ara-leu) 7697 araD139 fhuA ΔlacX74 galK16 galE15 e14- ϕ80dlacZΔM15 recA1 relA1 endA1 nupG rpsL (Str^R) rph spoT1 Δ(mrr-hsdRMS-mcrBC)*

2.1.6 Mammalian cell lines

Table 2.4. List of host cell lines used in this study.

Cell line	Description
HEK 293T	Human Embryonic Kidney cell line (adherent culture)
HEK 293 SA RIX	Human Embryonic Kidney cell line (suspension culture)
HEK293-6E	Human Embryonic Kidney cell line (suspension culture)

HEK 293T cells[678] were obtained from the Institute of Child Health/UCL, London, UK. HEK293 SA RIX cells were obtained from GSK vaccines, Rixensart (Belgium). HEK293 6E cells were originally obtained from National Research Council of Canada (#L-11266)[376].

2.1.7 Media and supplements

Table 2.5. List of media and supplements used in this study.

Medium or supplement	Manufacturer
OPTI-MEM	Thermo Fisher Scientific
Dulbecco’s Modified Eagle Medium (DMEM)	Thermo Fisher Scientific
CD293	Thermo Fisher Scientific
Freestyle 293	Thermo Fisher Scientific
GlutaMAX	Thermo Fisher Scientific
Pluronic F-68 (100X)	Thermo Fisher Scientific
HyClone G418 solution	Thermo Fisher Scientific
Dulbecco’s Phosphate Buffered Saline (DPBS)	Thermo Fisher Scientific
TrypLE Express Enzyme, no phenol red	Thermo Fisher Scientific
Fetal bovine serum (FBS)	Thermo Fisher Scientific
Zeocin	Thermo Fisher Scientific
Antibiotic-Antimycotic (100x)	Thermo Fisher Scientific
Dimethyl sulfoxide (DMSO)	Sigma-Aldrich
0.05% Trypsin-EDTA (1X) (TE)	Thermo Fisher Scientific

2.1.8 Equipment

Table 2.6. List of equipment used at GSK in this study.

Piece of equipment (GSK)	Manufacturer
Nanodrop ND-2000 spectrophotometer	Thermo Fisher Scientific
BioDoc-It Imaging system	UVP (Cambridge, UK)
JB1 Unstirred Waterbath	Grant Instruments
Infors HT Minitron	Infors HT (Bottingem, Switzerland)
Heraeus Biofuge pico (molecular biology)	DJB Labcare (Milton Keynes, UK)
Heraeus Multifuge X3R (molecular biology)	Thermo Fisher Scientific
Heraeus Multifuge 3S(tissue culture)	Thermo Fisher Scientific
C1000 Touch Thermal Cycler	BioRad (Hemel Hempstead, UK)

PowerPac Power Supply	BioRad
Heraeus HeraCell Air-Jacketed CO_2 Incubator	Thermo Fisher Scientific
Evos FL Cell Imaging System	Thermo Fisher Scientific
BD Accuri c6	BD Biosciences (San Diego, CA, USA)
Lab-Therm LT-XC Shaker-incubator	Kuhner (Birsfelden, Switzerland)
Axiovert 25 inverted bright field microscope	Zeiss (Cambridge, UK)
TK100 cryostorage Unit	Taylor-Wharton (Elstree, UK)
Nucleofector 2b	Lonza
INCell 2000	GE Healthcare (Hatfield, UK)

Table 2.7. List of equipment used at UCL/ICH/MCI in this study.

Piece of equipment (UCL/ICH/MCI)	Manufacturer
Nanodrop ND-1000 spectrophotometer	Thermo Fisher Scientific
Uvitec DOC-CF08-TFT. Gel Documentation System	UviTec
Eppendorf Mastercycler® Pro Thermal Cycler	Eppendorf
Gene Pulser II Electroporation System	BioRad
IX70 inverted bright field microscope	Olympus
FACSAria III	BD Bioscience
Cyan ADP Analyzer	BD Bioscience
ABI Prism 7000 Sequence Detection System	Thermo Fisher Scientific
Sorvall Discovery 100SE Ultracentrifuge	Thermo Fisher Scientific
Sorvall Legend Mach 1.6R Tabletop centrifuge	Thermo Fisher Scientific

2.1.9 Bioinformatic software and scripts

The analysis pipeline described in this study include several Perl and Bash scripts and should work on standard UNIX-based operating systems. The procedures described in this section were written by an external collaborator, Yilong Li, under specified criteria and require the following software to be installed and executable from the terminal (for example through aliases or through symbolic links in a directory that is in $PATH). The scripts listed below can be consulted in the Appendix B.

Table 2.8. List of programs required to run bioinformatic pipelines in this study.

Program	Version	Download link
BWA	0.7.12-r1039	http://sourceforge.net/projects/bio-bwa/files/
BEDtools	2.20.1	https://github.com/arq5x/bedtools2/releases
ssearch36	36.3.8b	http://faculty.virginia.edu/wrpearson/fasta/CURRENT/
FastQC	0.11.4	http://www.bioinformatics.babraham.ac.uk/projects/download.html#fastqc
Blat	36x1	http://hgdownload.cse.ucsc.edu/admin/exe/macOSX.x86_64/blat/blat
R	3.2.1	https://cran.r-project.org/src/base/R-3/
Starcode	1.0	https://github.com/gui11aume/starcode

2.2 Methods

2.2.1 Maintenance and long term storage of *E. coli*

Escherichia coli (*E. coli*) cells were cultured in vegitone lysogeny broth (vLB) at 37°C with shaking at 250rpm (in an Infors HT Minitron incubator) containing the appropriate antibiotic. Bacterial colonies were selected on vLB agar plates (1.5% (w/w) agar dissolved in vLB by heating) with the corresponding antibiotic (100 µg/mL of ampicillin or 50 µg/mL of kanamycin). For long-term storage of bacterial stocks, liquid cultures of bacteria in the exponential growth phase were resuspended in vLB with 15% (v/v) of glycerol and stored at -80°C.

2.2.2 Plasmid DNA preparation

E. coli cells containing the desired plasmid were cultivated in vLB with the appropriated antibiotic for selection (usually 100µg/mL ampicillin or 50µg/mL kanamycin) at 37°C with shaking at 250 rpm overnight. DNA was extracted using a plasmid DNA Miniprep Kit or Endofree plasmid Maxi Kit following manufacturer’s instructions and resuspended in molecular grade water. DNA concentration was measured on a Nanodrop ND-1000 or ND-2000 spectrophotometer and samples were stored at -20°C.

2.2.3 Restriction digests of plasmid DNA

For plasmid digestion, 0.5-2µg of plasmid DNA was incubated with 5U/µg of the appropriate restriction enzymes and 10% (v/v) of suitable 10X enzyme buffer in a final volume of 20µL per reaction. Double digests were performed sequentially, purifying DNA between reactions using QIAquick PCR purification column (Section 2.2.7), when a compatible buffer was not available. When not present in the enzyme buffer, BSA was added to a final concentration of 0.1mg/mL in order to stabilize enzymes during incubation. Reaction mixture was incubated for 1-2 hours at the recommended temperature.

2.2.4 Phosphorylation and dephosphorylation of DNA fragments

Dephosphorylation was carried out by adding 10% (v/v) of 10X FastAP buffer and 1U FastAP thermosensitive alkaline phosphatase to linear DNA (up to 1µg); The reaction was spun briefly, incubated for 10 minutes at 37°C and stopped by heating 5 minutes at 75°C.

For phosphorylation of annealed oligonucleotides, 10-50 pmol of 5’ termini were incubated with 10% (v/v) of 10X T4 polynucleotide kinase buffer, 10mM ATP and 10U of T4 polynucleotide kinase at 37°C for 20 minutes. The reaction was heat inactivated at 75°C for 10 minutes.

2.2.5 DNA elongation and blunting of overhanging DNA ends

Digested plasmid DNA (1-2 µg) was combined with 10% (v/v) 10X Klenow DNA polymerase buffer and 1U/µg Klenow enzyme together with 50µM of each dNTP in order to fill in 5’-overhangs and resect 3’-overhangs. Reaction components were mixed and incubated at room temperature for 10 minutes. The reaction was stopped by heat inactivation for 10 minutes at 75°C in a PCR machine and fragments were purified as indicated in Section 2.2.7.

2.2.6 PCR amplification of DNA fragments

Amplification of DNA fragments was performed using 0.625 units/reaction of One*Taq* polymerase, 1X One*Taq* standard reaction buffer, 200µM of dNTPs, 0.2µM of forward and reverse primers, a variable amount of DNA template and nuclease-free water in a total volume of 25µL.

Thermocycling conditions comprised an initial denaturation step at 94°C for 30 seconds followed by 20-35 amplification cycles consisting of denaturation at 94°C for 30 seconds, primer annealing at 50-60°C for 30 seconds and extension at 68°C for 1min/kb and a final extension step of 5minutes at 68°C.

When high fidelity was required for the amplification of subcloning, fragments were amplified using Q5 High-Fidelity DNA polymerase. In that case, reactions were set up using 0.5µM of primers and 0.5U/reaction of Q5 High-Fidelity DNA polymerase.

Thermocycling conditions for Q5 High-Fidelity DNA Polymerase consisted of an initial denaturation step at 98°C for 30 seconds followed by 20-25 amplification cycles consisting of denaturation at 98°C for 10 seconds, primer annealing at 50-60°C for 20 seconds and extension at 72°C for 2 minutes and a final extension step of 2 minutes at 72°C.

2.2.7 Purification of DNA fragments

When separated by restriction digest and gel electrophoresis, DNA bands were rapidly marked under low intensity UV light and isolated from agarose gels using a scalpel. The specific identified DNA bands were extracted from the gel using a QIAquick Gel Extraction Kit according to the manufacturer's instructions. When the size of the insert or plasmid backbone was not between 70bp to 10kb, a

QIAEXII Gel extraction Kit was used because it allows purification of a wider range of DNA fragment sizes (40bp to 50kb).

QIAquick PCR Purification Kit columns were used in order to purify DNA after ligations or digestions when suitable as per manufacturer's instructions.

2.2.8 Agarose gel electrophoresis

0.5-2% (w/v) agarose was added to 1X TAE buffer (40mM Tris acetate, 5mM EDTA) and heated until the mixture was completely dissolved. 0.5µg/mL ethidium bromide or SYBR Safe Gel Stain (0.1µL/mL of gel) was added when the agarose was cooled to approximately 55°C and then poured into the appropriate gel casting mould with an inserted comb.

10x Orange G DNA loading buffer was used as a loading dye for DNA samples and 1kb Plus DNA ladder was loaded as a molecular weight marker. Samples were run together with a 1kb plus DNA ladder (for determination of DNA fragment size) in a gel electrophoresis tank in 1X TAE buffer using a voltage of 80-120V. The DNA was visualized under UV light and photographed with a UV-gel documentation imaging system.

2.2.9 Ligation of DNA fragments

Ligations were performed with 1U of T4 DNA ligase, 1X T4 DNA ligase buffer (≤10% v/v), 100ng of plasmid backbone and 1:1 to 1:10 molar ratios of insert:vector in a final volume of 10-20µL. Reactions were incubated at 16°C, 4°C or room temperature (25°C) for 1-2 hours, 6-8 hours or overnight (depending on the conditions tested) and were transformed into chemically or electro-competent *E.coli* bacterial cells.

When required, 10% (v/v) of 50% PEG 4000 Solution was added to the ligation mix in order to promote intermolecular binding (only if followed by chemical transformation).

Alternatively, temperature-cycle ligations (TCL) described by Lund[679] were carried out for 12–16 hours in Eppendorf Mastercycler® Pro Thermal Cycler programmed indefinitely to cycle between 30 seconds at 10°C and 30 seconds at 30°C.

2.2.10 Topo TA ligation-independent cloning

According to Thermo Fisher Scientific instructions, 1µL of A-tailed PCR product was mixed with 1µL of water, 0.5µL of Salt Solution and 0.5µL of pCR4 TOPO TA vector (volumes per single reaction) and incubated for 1 hour at room temperature to undergo topoisomerase-mediated ligation of DNA termini. The reaction was then transformed into *Stbl3* competent cells (Section 2.2.12).

2.2.11 Preparation of Chemically Competent cells

Chemically competent *E. coli* cells were prepared according to the Inoue method[680]. A single colony *Stbl3 E. coli* was used to prepare an initial 3ml inoculum in vLB broth at 37°C with vigorous shaking at 250rpm overnight. 250mL SOB were inoculated (2% tryptone, 0.5% yeast extract, 10mM NaCl, 2.5mM KCl autoclaved and 10mM $MgCl_2$ and 10mM Mg_2SO_4 sterile-filtered and added before use) and cultured with moderate shaking at RT to an OD_{600} of 0.6. The culture was chilled on ice for 15 minutes and centrifuged at 2500*g* for 5 minutes at 4°C. Cells were gently resuspended in 80mL ice-cold Inoue transformation buffer (10mM PIPES, 15mM $CaCl_2$, 250mM KCl, 55mM $MnCl_2$, 1M KOH to pH6.7) for 10 minutes, and centrifuged as before. The resulting pellet was resuspended in 20mL 7% DMSO in Inoue Transformation Buffer, placed on wet ice for 10 minutes, and dispensed into aliquots previously snap frozen in liquid nitrogen and stored at -80°C.

2.2.12 Transformation of competent *E. coli*

When using chemically competent cells, 1-5µL of ligation mix was added to 100µL of competent cells previously thawed on ice. The transformation mix was incubated for 30 minutes on ice before undergoing heat shock for 30 seconds at

42°C (depending on manufacturer recommendations) and placed on ice for 2-5 minutes afterwards.

When using electroporation, 50µL of electrocompetent cells were transferred into a pre-chilled 0.2mm gap electroporation cuvette. Electroporations were performed using a Gene Pulser II Electroporation System at 25µF, 2.5kV, 200Ω.

In both cases, up to 1mL of pre-warmed (37°C) SOC medium (SOB medium with 10mM $MgCl_2$ or 20mM $MgSO_4$ and 20mM glucose) was added to the cells and incubated for 1 hour at 37°C with vigorous shaking (200-250rpm). Bacteria were then plated out to pre-warmed (37°C) vLB agar plates containing the appropriate antibiotic and incubated overnight at 37°C.

The next day, 1mm colonies were picked using sterile pipette tips and cultured in 3-5mL of vLB with the appropriate antibiotic overnight at 37°C with agitation (200-250rpm).

2.2.13 Sanger-sequencing of PCR amplicons and subcloned plasmids

Sanger sequencing of PCR amplicons or subcloned plasmids (and libraries) was performed in-house at GSK, by UCL DNA Sequencing services or externally by Source Bioscience.

2.2.14 Construction of barcoded vector libraries

1% (v/v) single strand oligonucleotides (100µM) obtained from Invitrogen at 25nmol, desalted purity (5'-TATGAGTAANNNATCNSGATNNAAANNGGTNWAACNNTGANNNTGGTAACACCGACTAGGATCCTGAT-3'; 5'-CTAGATCAGGATCCTAGTCGGTGTTACCANNNTCANNGTTWNACCNNTTTNNATCSNGATNNNTTACTCA-3'; note they create overhangs compatible with *XbaI* and *NdeI*) were combined in water with 1X ligase buffer and allowed to gradually cool (1°C/min) to 16°C from after an initial 5-10 min temperature hold at 95°C. Double stranded DNA adapters were then carefully resuspended, aliquoted and stored at -20°C for subsequent use.

50-150ng of vector backbone was diluted to 20μL with 400U T4 DNA ligase and the annealed oligonucleotides were added to 10-, 100-, 1000 or 10000-fold molar excesses to the vector (Section 2.2.9). A reaction without oligonucleotides was included for comparison. After 1-2 hours, 6-8 hours or overnight ligation at 16°C 4°C or room temperature, 1-5μL of ligation mix were transformed into competent cells (Section 2.2.12).

Dephosphorylation of backbone, phosphorylation of the paired oligonucleotides (Section 2.2.4), addition of PEG 4000 (5% (v/v) ligation mix volume), heat inactivation after ligation (ligation mix incubated at 65°C for 20 minutes in order to inactivate T4 DNA ligase) or other insert:backbone molar ratios were also tested parameters in the optimization process for the creation of a barcoded plasmid library.

2.2.15 Propagation of cell lines

Adherent human embryonic kidney (HEK) 293T cells were seeded into 175cm^2 T-flasks and maintained as a monolayer in Dulbecco's modified eagle medium (DMEM) supplemented with 10% foetal calf serum (FCS).

Suspension human embryonic kidney 293 SA RIX were cultured in CD293 medium and supplemented with 4mM GlutaMAX™. Cell lines were incubated in 75cm^2 T-75 flasks (upright position) in static incubators, although shaking incubators (not available at UCL/ICH/MCI) enable higher viabilities.

Human embryonic kidney 293 cells 6E (HEK 293 6E) were cultured in Freestyle 293 medium and supplemented with 10mL of 10% Pluronic F-68 (100x) and 500 μL of 50mg/mL HyClone G418.

All cultures were incubated at 37°C in the presence of 5% CO_2, in 70% relative humidity conditions. Suspension cell lines were incubated in conical flasks in shaking incubators at 140 rpm in a Kuhner Lab-Therm LT-XC. Suspension cells were passaged by diluting 1:10 with fresh medium and transferring them to a new flask. Adhered cells were detached by incubating them with TrypLE

Express™ (1X) or 0.05% (v/v) Trypsin-EDTA (1X) for 5 minutes at 37°C (after a wash step with PBS). Cells were diluted down to the desired concentration with fresh medium when they started to detach. Cells were passaged twice a week down to a $0.3x10^6$ cells/mL cell density. Cell density and viability were quantified by exclusion method using Trypan blue.

2.2.16 Production of integrating lentiviral vectors

Lentiviral vectors were produced using polyethylenimine (PEI, Sigma-Aldrich, USA), HEK 293T cells and the four-plasmid system that a third generation lentiviral vector requires, following the procedure described by Naldini *et al.*, 1996[681]. $1.5x10^7$ cells were seeded 24 hours prior to transfection in a T-175 flask in 15-20mL DMEM with 10% FBS. The next day cells were co-transfected with sterile-filtered plasmids in Opti-MEM® media: 45µg/plate vector construct, 17.5µg/plate of pMD.G2 (containing VSV-G *env* gene[682]), pMDLg/pRRE 32.5µg/plate (containing *gag* and *pol* genes[677]) and 12.5µg/plate of pRSV Rev (containing *Rev* gene[146]) and 1µL of 10mM polyethylenimine (PEI) solution. Medium was replaced 4-6 hours post-transfection to remove PEI and cells were incubated at 37°C 5% CO_2 for 48 hours.

When transfected with calcium phosphate technique, 5µL of pMDL/RRE, 2.5µL of pRev, 3.5µL of VSV-G and 14µL of transfer vector (or pTelo) plasmids (all of them at a 1µg/µL concentration) were mixed with 62µL of $CaCl_2$ 2M and 413 µL of water (amounts required for transfection in 10cm^2 plates). 500µL of 2x Hepes phosphate buffer saline (HBS) was added in 15mL tubes and air was bubbled through the solution with a 1mL plastic pippete attached to a pippete pump to form the DNA precipitates while the DNA mix is progressively added dropwise. Complexes were then incubated 20 minutes at room temperature and subsequently added to the cells. Medium was replaced 14-16 hours post-transfection and cells were incubated at 37°C 5% CO_2 for 48 hours.

In both cases, the medium was then harvested and flasks were replenished with fresh medium. Harvested media was cleared using a 0.22µm filter to remove cell debris and was centrifuged at 98,000*g* in a Sorvall Discovery 100SE

Ultracentrifuge for 2 hours at 4°C. Supernatant was discarded after centrifugation and viral particles were resuspended in Opti-MEM® media, aliquoted and stored at -80°C. Centrifugation was repeated 72 hours post-transfection for a second harvest following the same procedure.

2.2.17 Flow cytometry analysis of transduced cell lines

The proportion of transduced cells positive for eGFP fluorescence was analysed after being cultivated for 72 hours at 37°C and 5% CO_2. Cells were cultured to a semi-confluent phase in 6-well plates and the medium was removed, cells washed with PBS, treated with 250μL of TrypLE Express™ (1X) or 0.05% (v/v) Trypsin-EDTA (1X) in PBS for about 3 min at room temperature, and pelleted by centrifugation at 1200x*g* for 5 minutes in a Thermo Fisher Scientific Sorvall Legend Mach 1.6R Tabletop centrifuge or a Heraeus Multifuge 3S. The cell pellet with 0.1-1x10^6 cells as then resuspended in 300μL of ice cold PBS and kept on ice in polypropylene tubes until analysis.

Flow cytometry analyses were done at Great Ormond Street Hospital Flow Cytometry Core Facility (UCL-Institute of Child Health/Great Ormond Street Hospital for Children) using a CyAn ADP Analyzer (Dako, Glostrup, Denmark) with an argon laser (excitation at 488nm; filter at 491nm and emission at 530nm; filter at 530±20nm) according to the procedure recommended by the manufacturer. Ten thousand cells were analysed for each sample. Data was analysed using FlowJo version 7.6.5 (TreeStar, Stanford University). When performed at GSK, a BD Accuri c6 was used (excitation at 488nm; filter at 491nm and emission at 530nm; filter at 533±30nm). Data was analysed using BD CSampler version 1.0.264. 21.

2.2.18 Determination of viral vector titer by flow cytometry

Functional lentiviral vector titer can be assessed by flow cytometric evaluation of transduction rates (Section 2.2.17). At 24 hours prior to infection 1x10^5 cells (HEK 293T, HEK 293 SA RIX or HEK 293 6E) were seeded in a 24-well plate and in 1mL of media. The next day the media was replaced by 1mL of 10-fold serial dilutions starting with 10µL of virus in the appropriate media. At day 3 post-infection, 250µL of TrypLE Express™ (1X) or 0.05% (v/v) Trypsin-EDTA (1X) was added to cells and the dissociation reagent was inactivated by resuspending them in 1mL of DMEM with 10% FBS. Cells were pelleted and resuspended in 500µL of FACS buffer (1% FCS, 0.05% sodium azide in PBS) to perform flow cytometric analysis to determine the percentage of GFP positive cells. The infecting or transducing capacity (titers) of the lentiviral vector produced was calculated by multiplying the number of seeded cells by the percentage of GFP positive cells and dividing by the volume of lentiviral vector used for the infection.

Titrations were also performed on different HEK 293 cells since the number of transducing units per millilitre (TU/mL) can vary among cell lines[284,398,683]. Cells were cultured in 24-well plates and were trypsinized with TrypLE Express™ (1X) or 0.05% (v/v) Trypsin-EDTA (1X). Suspension cells were spun in a 24-well plate at 1200x*g* (in a Thermo Fisher Scientific Sorvall Legend Mach 1.6R Tabletop centrifuge) for 5 minutes in order to change the media.

2.2.19 Transduction of host cell lines

HEK293 cells were seeded at a density of 0.3x10^6 cells/mL per condition in a 6-well plate with the appropriate culture medium (as previously described). Alternatively, 1,000 and 10,000 HEK293 6E cells were seeded in a 96 well plate (100,000 cells were transduced in a 24well plate). Cells were infected with different desired MOIs according to the transducing units of titrated lentiviral vector 24h after seeding and were incubated for 72 hours incubation at 37°C and 5% CO_2. Half of the media was replaced 24 hours post-infection.

2.2.20 Fluorescence microscopy

Images of transduced or transfected 293 cells were captured using an Olympus IX70 microscope or an EVOS FL Cell Imaging System (ocular magnification: 10x, objective magnification 4x, 10x or 20x) after 72 hours incubation at 37°C and 5% CO_2.

2.2.21 Isolation of genomic DNA

Medium from up to 5x10^6 HEK 293 cell lines was removed, the cells were washed twice with PBS and genomic DNA extraction from and HEK 293 cell lines was performed using the DNeasy Blood and Tissue Kit as per manufacturer instructions. DNA was eluted in AE buffer (10 mM Tris-Cl, 0.5 mM EDTA; pH 9.0) and stored at -20˚C.

2.2.22 Measurement of nucleic acid concentration

Concentration and purity of plasmid DNA was determined using a Nanodrop ND-1000 or ND-2000 spectrophotometer with a 0.2mm path to measure the absorbance at 260nm because nitrogen rich bases absorb light at this wavelength. Nucleic acid concentration can be determined once known that the extinction coefficient of dsDNA, ssDNA and RNA is 50, 33, 40 µg/mL, respectively. The ratio absorbance at 260nm and 280nm was used to assess the purity of DNA. A ratio of ~ 1.8-2.0 was accepted as sufficiently pure DNA.

Prior to next-generation sequencing runs, DNA and RNA sample concentration and purity were measured pre- and post-library preparation using an Agilent 2100 Bioanalyzer or a 2200 Tapestation. DNA peaks were expressed in fluorescent units between 10,380bp and 35bp high DNA sensitivity markers; 28S/18S ratios was used to assess quality of total RNA samples

2.2.23 Long term storage and revival of mammalian cell lines

For long-term storage, 1-5x10^6 cells were pelleted at 1200x*g* (in a Thermo Fisher Scientific Sorvall Legend Mach 1.6R Tabletop centrifuge or Heraeus Multifuge 3S) for 5 minutes, resuspended in 1mL of cryopreservation medium (suitable media according to Section 2.2.15 with 10% dimethyl sulfoxide, DMSO) and transferred to cryovials. Cells were gradually frozen (1°C/minute) using an isopropanol-freezing container before being transferred to liquid nitrogen.

In order to revive cells, frozen aliquots were thawed by hand, resuspended in 9mL of pre-warmed medium and pelleted at 1200x*g* for 5 minutes (in a Thermo Fisher Scientific Sorvall Legend Mach 1.6R Tabletop centrifuge or Heraeus Multifuge 3S). Supernatant (with DMSO) was removed and cells resuspended in 10mL and transferred to a 125mL shake flask or T-175 in a suitable tissue culture flask for at least 48 hours prior to be seeded for experiments. Revived host cell lines were passaged twice (Section 2.2.15) prior to experiments involving transfection (Section 2.2.39) or transduction (Section 2.2.19).

2.2.24 Cell sorting of transduced cell lines

120x10^6 HEK 293 cells were transduced in bulk with the lentiviral library (RRL.SIN.cPPT.EEW+Barcode) at an MOI of 0.5. After 1 week post-transduction, cells were pelleted, washed and resuspended in Gey's balanced salt solution at a concentration of 10-20x10^6 cells/mL. Media for cell recovery after sorting consisted of 1:1 of fresh media and conditioned media (filtered, non-exhausted previously used media which supplies growth factors and metabolites) and antibiotic-antimycotic at a 1x final concentration to prevent bacterial or fungal contamination. Cells were analysed and sorted by Clare Mudd at Labstract (Stevenage Bioscience Cayalyst) on a fluorescence-activated cell sorter FACSAriaIII using FACSDiva software (BD Biosciences). The fluorochromes were excited at 488-nm and green fluorescence was detected using a 530/20 filter. Prior to sorting, the nozzle, sheath, and sample lines were sterilized with 70% ethanol. A 100-µm ceramic nozzle (BD Biosciences), sheath pressure of 20 pounds per square inch (PSI) and an acquisition rate of 5,000-10,000 events per second was used to sort cells at 4°C into 3 different pools based on the level of eGFP-specific fluorescence (top 2.5% high, 2.5% mid and 2.5% low GFP producers) in aseptic conditions. GFP- and GFP+ populations were also sorted as a control.

When performed at Labstract (The Catalyst building, Stevenage), 50,000 cells were sorted per intensity condition (100,000 for GFP+ and GFP-) into eppendorfs containing 200µL of recovery media under the following conditions: serum free, 4°C, 100-µm nozzle, continuous agitation and a sorting speed of 2,000-4,000 events per second. In any case, cells were transferred to V-shaped 96 well plates for recovery and expanded when reached maximum confluency.

2.2.25 Determination of lentiviral vector copy number by quantitative real-time PCR (qPCR)

Lentiviral vector copy number was determined by quantitative real-time PCR (qPCR) using the absolute quantification method at Institute of Child Health, UCL.

Reactions were performed in triplicate using approximately 250ng of genomic DNA as a template per reaction, 0.9 µM of each primer, 0.2 µM of fluorescent probe, and the Platinum qPCR SuperMix-UDG with ROX mastermix.

Real time PCR was performed as follows: 1 cycle of 50°C for 2 minutes, 1 cycle of 95°C for 10 minutes followed by 40 cycles of [95 °C for 15 seconds; 60°C for 1 minute] using an ABI Prism 7000 Sequence Detection System.

pMKRQ BTW2R plasmid DNA (Appendix A) containing 10^6-10^3 copies/5 µl the Woodchuck hepatitis virus enhancer element sequence (WPRE), kindly gifted by John Counsell, although originally cloned by Conrad Vink (both Institute of Child Health/UCL, London, UK), was used as standard for lentiviral copy number quantification. WPRE qPCR primers and probes were kindly provided by John Counsell:

Forward primer: 5’-TGGATTCTGCGCGCGGGA-3’

Reverse primer: 5’-GAAGGAAGGTCCGCTGGATT-3’

Probe (5’-3’): FAM-CTTCTGCTACGTCCCTTCGGCCCT-TAMRA

Vector DNA copy number was calculated using the genome mass and the mass of DNA employed in the qPCR. Copy numbers per cell are calculated dividing the extrapolated copy numbers by the number of cells present in the sample. Cell number is calculated as the ratio between the mass of template DNA used in each reaction divided by the mass of a single host cell genome. The mass of the host cell genome is calculated using number of chromosomes, the ploidy, their length in bp and the mass of 1bp of DNA. This method of normalisation has been used in the literature for absolute quantification of transcripts[684].

2.2.26 Extraction and isolation of cellular and viral RNA

Cellular RNA was extracted from 0.1-1x10^7 cells per condition using the RNeasy Plus Mini Kit following manufacturer's instructions. When extracting RNA from lentiviral vectors, 40µL of concentrated (ultracentrifuged) viral vector were treated using a QIAamp Viral RNA Mini Kit following manufacturer's instructions. In both cases, RNA was eluted in 50µL of RNAse-free water, measured on a Nanodrop ND-1000 or ND-2000 spectrophotometer (Section 2.2.22) and samples were stored at -80°C.

2.2.27 Elimination of residual DNA in RNA samples

Residual DNA was eliminated from RNA preparations using DNA removal with DNA-*free* Removal Kit or the RNA Clean & Concentrator Kit following manufacturer's instructions. Samples were eluted in 25µL of RNAse-free water and stored at -80°C prior to generation of cDNA.

2.2.28 Generation of Barcoded cDNA from viral/ cellular RNA

Synthesis of cDNA and amplification of a specific barcoded region was achieved in a single step using the SuperScript® III One-Step RT-PCR System with Platinum® *Taq* DNA Polymerase. A cDNA/RNA hybrid was generated by the SuperScriptIII reverse transcriptase from 100ng of viral or cellular RNA. The reaction was carried out in a 30-minute incubation at 59°C. Specific amplification of barcoded regions was performed using, 1µL of 10µM forward and reverse primers (RNAbc_150ups-fwd 5'-ACGAGTCGGATCTCCCTTTG-3' (annealing at the 3' end of the WPRE and thus only amplifying barcode from the 3'LTR); Barcode-PBS-rev 5'-GGATCCTAGACGGTGTTACC-3') and reaction buffer (1x) containing

50μM each dNTP under the following cycling conditions: 1 cycle of 94°C for 2 minutes (which deactivates the reverse transcriptase and activates the *Taq* DNA polymerase) followed by 25 cycles of [94 °C for 30 seconds; 60°C for 30 seconds and 68°C for 15 seconds] and a final extension 5min step at 68°C. Reactions were further purified using the QIAquick PCR purification Kit (Section 2.2.7) prior to submission for next-generation sequencing. A comprehensive diagram showing the fragment of reverse transcribed vector RNA amplified after reverse transcription can be consulted at Figure 3.9F.

2.2.29 Integration site analysis of barcoded integrated lentiviral vectors by Ligation-mediated (LM-PCR)

Lentiviral vector – host chromosome junctions were retrieved by using a linker cassette that provides a known sequence to specifically amplify target sequences when ligated to fragmented genomic DNA.

Linker cassette was generated by mixing 12.5 pmol of each linker oligo (Linker fwd: 5'-GTAATACGACTCACTATAGGGCACGCGTGGTCGACGGCCCGGGCTGGT-3' and Linker rev 5'Phos-ACCAGCCCGGGCCGT-3'SpC3, compatible with blunt end restriction enzymes such as *DraI*) in a 50μL final volume (to a 25μM final concentration) with 10% T4 DNA ligase buffer as described in Section 2.2.9. The mixture was heated at 95°C for 2 minutes and gradually cooled to room temperature.

The linker cassette reverse oligonucleotide contains a modification in order to prevent the PCR-suppression effect[685]. This effect occurs upon extension of the linker cassette reverse oligonucleotide, which results into a full-length linker cassette that can form linker concatemer and serve as a template for end-to-end amplification. Under the annealing and extension temperature conditions, intramolecular annealing is strongly favoured over the annealing of a shorter primer to the linker cassette, which leads to a 'panhandle' structure, which hampers PCR amplification. A three-carbon spacer (C3-Spacer or 3-SpC3) was added to the 3' terminus to impede extension of the linker cassette reverse oligo and thus ligation of any molecule on the 3' end.

2.5µg of genomic DNA were digested with 80 units of restriction enzyme (in this study *DraI*, *NlaIII*, *BsuRI* and *HpyCH4v*) for 2h at 37°C, column-purified with the QIAquick PCR purification kit and ligated to 1.9µL (47.5pmol) of linker cassette at 16°C over night. The reaction was stopped at 70°C for 5 minutes and diluted 5 times with distilled water prior to 2 rounds of PCR using 5µM primers, 2mM $MgCl_2$ and 0.05 units of TrueStart Hot Start Taq DNA polymerase with the following thermocycling conditions: 7 cycles at [94°C for 25 seconds and 72°C for 3 minutes]; 32cycles at [94°C for 25 seconds and 67°C for 3 minutes] and 1 cycle at 67°C for 7 minutes.

LVVP1 (lentiviral vector primer1): 5'- GCTTCAGCAAGCCGAGTCCTGCGTCGAG -3'

LCP1 (linker cassette primer1): 5'- GTAATACGACTCACTATAGGGC -3'

LVVP1 anneals at a sequence immediately downstream the LVV 5'LTR so that amplification from both LTRs is avoided. A comprehensive scheme of the locationof the binding site of the primers can be consulted in Figure 4.1.

The second PCR round was carried out using the same master mix composition and 1/50 (diluted with molecular grade water) of the first PCR product with the following thermocycling conditions: 5 cycles at [94°C for 25 seconds and 72°C for 3 minutes] and 32cycles at [94°C for 25 seconds and 67°C for 3 min].

LVVP2 (lentiviral vector primer2): 5'- GGATCCTAGTCGGTGTTACCA -3'

LCP2 (linker cassette primer2): 5'- ACTATAGGGCACGCGTGGT -3'

1µL of the 2nd PCR product was ligation-independent cloned into a pCR4 backbone following the instructions of the using the TOPO-TA Cloning Kit for 1hour at room temperature and the resulting mixture transformed into *Stbl3* chemically competent cells section and validated by Sanger-sequencing.

Once the technique was validated on low-throughput, 2nd PCR products containing the Illumina adapter sequences and individual indexes to allow sample identification were column-purified with the QIAquick PCR purification Kit and sent to Genewiz (South Plainfield, NJ, USA) or UCL Genomics (London, UK) for Next-Generation Sequencing using the a MiSeq and 300bp Paired End strategy.

2.2.30 Next-Generation Sequencing

Generation of the sequencing libraries and sequencing runs were carried out by Genewiz (South Plainfield, US) or UCL Genomics (London, UK) using the following primers:

Plasmid (Uppercase for Illumina compatible sequences)

P5fwd-upstream-barcode:
5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCTccttcgccctcagacgagtcggatctc-3'

bio-P7rev-downstream-barcode (also used for the generation of a custom barcoded library from total RNA):
5'-[bio]GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTcaaaaagcatctagatcaggat
cctagtcggtgttacca-3'

LM-PCR

P5_FWD-LentiXASP2: 5' ACACTCTTTCCCTACACGACGCTCTTCCGATCTactataggg
cacgcgtggt

P7_REV-LentiXPBS2: 5'-GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTggatccta
gtcggtgttacca-3'

When performing integration site analysis, LM-PCR host genome-vector junctions were sequenced using Illumina MiSeq sequencer 2x300bp (paired-end) configuration and a v3 Kit in order to increase cluster density, maximise read length as well as improve quality scores.

When performing barcode expression analysis or whole transcriptome analysis, RNA-Seq expression was performed using NextSeq500 (UCL genomics) and HiSeq2500 (Genewiz) on libraries prepared from total RNA.

2.2.31 Paired-end joining (and reverse complementing sequences)

R1 and R2 datasets (per sample) were output from the sequencing in fastq format when using paired end configuration. Regarding paired-end joining of plasmid PCR reads, the read length from MiSeq 300bp PE configuration was higher than the length of the PCR product and thus the reads often tend to sequence 'into' the primers outside the amplified target product. Therefore, all reads were first trimmed down to 240bp. LM-PCR reads were trimmed down to 150bp to obtain the optimal amount of merged reads in subsequent steps.

Subsequently, the actual read pair merging was done using BWA. The appropriate value for parameter -Q, optimised through trial and error.

No merging was performed on RT-PCR samples HiSeq2500 100bp PE because the barcode was entirely found in R2. Instead, a simple script called 'rc_fastq.pl' (Appendix B) was written for reverse-complementing FASTQ R2 files. In the case of finding low quality scores towards the 3' end of the read (represented as Ns), the scrip would also remove reads with ≥5Ns.

2.2.32 Manipulation of sequences using Galaxy

Next-generation sequencing data manipulation involving sequence trimming, sorting, selection or replacement as well as combination or subtraction of datasets, operations with columns and data storage was performed using Galaxy, a web-based platform for high throughout genomic analyses[686–689].

2.2.33 Quality control

FastQC[690] (Babraham Bioinformatics) was applied on the merged fastq files to in order to assess quality statistics, GC content, sequence length distributions and duplication and overrepresentation levels.

2.2.34 Extraction of barcodes (and host integration sites)

A custom Perl script, that uses Ssearch36 software, was adapted to different read configurations in order to extract barcodes from the merged reads in different experiments.

The script 'extract_library_barcode.pl' (Appendix B) was used to extract barcodes from the plasmid library. The script specifically identifies the following pattern: 20bp sequence upstream the barcode (5'-GACAAGATCCATATGAGTAA-3')-barcode sequence (5'-NNNATCNSGATNNAANNGGTNWAACNNNTGANNN-3')-20bp downstream the barcode (5'-TGGTAACACCGACTAGGATC-3'). No mismatches or insertions/deletions are allowed in matching this pattern.

The extraction of barcodes and integration site sequences from LM-PCR merged reads (sequenced with a MiSeq 300bp PE strategy) was performed using perl script version 'extract_viral_insertion_barcodes.pl' (Appendix B) meeting the following criteria: (1) Alignment to linker ended at the last 5 bases (bases 32-36) on the reference linker sequence. (2) The sequence identity was > 80%.

The script 'extract_viral_insertion_barcodes.pl' was modified to 'extract_rt-pcr_barcodes.pl' in order to extract barcodes from viral and cellular RT-PCR amplicons sequenced under a HiSeq2500 100bp PE.

All the scripts mentioned in this Section 2.2.34 were written by Yilong Li (external collaborator, Wellcome Trust Sanger Institute, Cambridge, UK).

2.2.35 Barcode clustering

Barcode variants with insufficient dissimilarity and low relative representation were pooled together using Starcode clustering software[691]. Clustering parameters 'size absorbing ratio' (-r) and 'editing or Leveinshtein distance' (-d) were optimised in agreement with their frequency of dissimilarities.

2.2.36 Plotting barcode distributions

Cumulative density of variant frequencies pre- and post-clustering treatment were plotted using the following script in R 'plot_plasmid_library_ distributions.R' (Appendix B).

Frequencies of dissimilarities between barcoded variants were calculated using the script 'MHB08-059_check_and_assign_pSYNT' and the resulting tables were plotted in R using the 'Barcode_error_correction.R' script (Appendix B). The latter two scripts were kindly given by Martijn H Brugman (Leids Universitair Medisch Centrum (LUMC), Germany).

2.2.37 Length filtration and mapping

Host sequences were filtered to be at least 20bp long and subsequently converted into fasta format before being mapped against the human genome using Blat[692]. Integration sites were plotted using the UCSC Genome Graphs tool[693] together with CpG islands and RefSeq gene annotation tracks. *Homo sapiens* GRCh37 assembly (NCBI) Reference genome version used was Homo_sapiens (GRCh37/hg19):

ftp://ftp.ensembl.org/pub/release75/fasta/homo_sapiens/dna/Homo_sapiens.GRCh37.75.dna.primary_assembly.fa.gz

2.2.38 Feature annotation

BEDtools was used to assign intersect, merge, count and genome annotation features to genomics intervals or positions retrieved during the integration site analysis[694]. Initially, gene coordinates and symbols were obtained from Ensembl BioMart, available at the following link:

http://grch37.ensembl.org/biomart/martview/f828296ee921fd33b715b6aea8d521aa). Only genes and transcripts with a CCDS ID (Consensus Coding Sequence ID) were included.

CpG island annotations, based on *Homo sapiens* (GRCh37/hg19) assembly Feb. 2009, were downloaded from the UCSC table browser as a BED file (https://genome.ucsc.edu/cgi-bin/hgTables).

The raw downloaded annotation data were processed using the code in Bash script 'make_ucsc_gene_txs.sh' (Appendix B) in order to prepare annotation (bed) files of genomic features such as genes, transcription start sites and CpG islands that were used in the analysis.

Custom tracks containing annotation for repetitive elements were obtained from UCSC Table browser, which uses data from 'Repbase update library of repeats' from the GIRI (Genetic Information Research Institute)[695].

2.2.39 Nucleofection of host cell lines with CRISPR-Cas9 constructs

Stable integration of donor constructs into the genome of HEK 293 6E cell lines was achieved by co-transfection of separate plasmids containing sgRNA and Cas9. 2x10^6 cells per condition were pelleted and washed with Hank's balanced solution prior to nucleofection using the Amaxa Nucleofector Kit V. Cells were resuspended in 82µL of Solution V supplemented with 18µL of Supplement 1 and mixed with 2µg of each plasmid. The mixture was transferred into an electrocuvette and nucleofected using a Nucleofector 2b device (Lonza) using the program S-018. Cells were resuspended with pre-warmed media and gently transferred into a 6-well plate statically incubated at 37°C and 5% CO_2 for at 48 hours before any selective pressure was applied. Transfection efficiency was assessed 48 hours post-transfection by flow cytometry.

Reactions were performed in triplicates and negative controls (no DNA and single plasmid controls) were performed alongside.

2.2.40 Selection and isolation of host cell line colonies

HEK 293 transfected cell pools were kept in 6-well plates for 48 hours post-transfection without antibiotic selection. The transfection efficiency of donor construct and sgRNA and Cas9 plasmids was assessed by FACS (Section 2.2.17) since plasmids contained an eGFP and RFP fluorescent marker, respectively. The media was then replaced by media containing the appropriate antibiotic (300-500μg zeocin/mL) and cells were cultured for 1-2 weeks under selection (changing the media every 3-4 days) until cell colonies became visible. Individual cell colonies were transferred to individual 24-well plate using cloning cylinders and keeping selective pressure. Alternatively to cloning rings, media was replaced with PBS containing 5-10% of TrypLE Express™ (1X) or 0.05% (v/v) Trypsin-EDTA (1X) and incubated for 5 minutes at room temperature. Colonies could then be directly and individually pipetted to 24-well plate. Once cells were expanding, they were readapted to suspension cultures.

2.2.41 Scale up of host cell line colonies

HEK 293 6E eGFP positive colonies from pools successfully transfected with donor construct according to FACS (Section 2.2.17) were transferred from a 24-well plates to a 6-well plates and 125mL shake flasks when they reached 90% confluency. Media with selective pressure (300-500μg Zeocin/mL) was replaced every 3-4 days. Clones were screened for eGFP intensity by FACS, vector copy number by qPCR (Section 2.2.25), viral titer (Section 2.2.18), off-target integration (Section 2.2.42) and integrity of the junction by PCR and Sanger sequencing (Section 2.2.6 and 2.2.13).

2.2.42 Determination of off-target integration of donor constructs

In order to assess random integration of donor construct, a blue (BFP) fluorescent cassette was downstream of the right homology arm. If recombination events were successful on both homology arms, no blue signal should be detected. Since strongly enhanced blue fluorescent protein (seBFP) emission and excitation wavelengths is not within the rank of detection of the BD Accuri c6, an IN Cell Analyzer 2000 (GE Healthcare Life Sciences, UK) was used instead to detect BFP (excitation at 350nm; emission at 470nm); RFP (excitation at 596nm; emission at 515nm); GFP (excitation at 490nm; emission at 525nm). Overlapping signal was compensated with single-fluorochrome controls (pmaxGFP, pMA-RQ HA-MCS-

HA2-BFP CUL5 and pCMV-Cas9-U6-sgRNA-RFP for GFP, BFP and RFP, respectively). Images were taken at a 20x magnification and percentages of BFP, GFP and RFP positive cells were quantified with Columbus software using a custom script written by Toral Jakhria (GSK) (Appendix B) to quantify the proportion of GFP+ve/BFP-ve cells.

Alternatively, images of cells bright field or expressing BFP and/or GFP were taken using a confocal fluorescence microscope Leica TCS SPS II with an Argon laser with a Alexa Fluor 488nm detection filter for GFP and a 405nm laser and DAPI detection filter for BFP. 20x (HC PL APO 20x/0.70 CS) and 40x (CX PL APO 40x/0.85 CORR) magnification objectives were used and photomultiplier I and II, respectively for GFP and BFP.

2.2.43 Confirmation of integrity of donor construct-host genome junctions by PCR

Genomic DNA from HEK 293 6E host cell lines generated upon nucleofection of donor construct and sgRNA+Cas9 plasmids was extracted (Section 2.2.21) and junction sequences were amplified by PCR (Section 2.2.6) both sides of the insertion for each candidate using the following primers:

Control position EMX1 (right junction) (Tm=55C)

Zeonestedfwd2:	5'-gtcgagacgtacccaattcg -3'
EMXrightrev4:	5'-atcctcccctttcctctggt -3'

EGFEM1P (left junction) (Tm=64.5C)

Leftfwdnew1:	5'-cgttcccttcttcccttcct -3'
Gagrev2:	5'-gtaagaccaccgcacagcaa -3'

CUL5 (right junction) (Tm=57C)

Zeonestedfwd2:	5'- gtcgagacgtacccaattcg-3'
CUL5rightrev1:	5'-caagctcatcactgcacctc -3'

PCR amplicons were TOPO cloned into a pCR5 backbone (Section 2.2.10), transformed into *Stbl3* competent cells, plated out in LB agar plates with the appropriate antibiotic and incubated overnight at 37°C. The next day, colonies were picked and liquid bacterial cultures were set up for plasmid DNA extraction (Section 2.2.2) and Sanger sequencing using the M13 reverse primer (5'-CAGGAAACAGCTATGAC -3') (Section 2.2.13).

2.2.44 Statistical analysis

One-way Anaysis of Variance (ANOVA) was used to discern whether differences between mean±SD obtained from three replicates for each condition were significantly different. A post hoc Dunnett's test was used to compare multiple treatments to a 'fixed' negative control. A post hoc Tukey's test was used in conjunction to one-way ANOVA to determine significance between treatments, conditions or groups without a control. Groups of related samples were analysed using the Friedman's test analysis of variance.

In Chapter 4, Chi-squared tests were performed to determine the probabilities that integration sites are significantly close to RefSeq genes and other genomic annotation features compared to randomly generated integration sites using VISA[696]. The total number of observed frequencies is multiplied by the expected ratio (obtained from randomly generated IS) to determine the expected number of events assumed under the null hypothesis. The Chi-square statistics were calculated using the formula $(O\text{-}E)^2/E$ where E is the number of expected events and O is the number of observed events. The statistic obtained and the number of degrees of freedom where then used to calculate the p-values. Yate's correction was applied in the case of a 2x2 contingency table and 1 degree of freedom.

Statistical analysis was performed in GraphPad Prism version 5.0 (San Diego, CA, USA). In all cases, levels of significance were established as follows: * $p<0.05$, ** $p<0.001$, *** $p<0.0001$.

Chapter 3

RESULTS: Generation, characterisation and delivery of lentiviral barcoded vector libraries

3.1 Introduction

The development of biopharmaceutical producer cell lines requires a considerable investment of time and resources. Low-throughput methods for cloning, screening and selection despite being simple, reliable and inexpensive are time-consuming and significantly limited by the number of clones that can be feasibly screened. On the other hand, higher-throughput (HTP) strategies present the opposite characteristics: automated and sterile closed system albeit expensive and highly sophisticated. A high resolution, high-throughput, biologically driven method for selection of high expression clones applicable to multiple cell lines seems an attractive concept to shorten cell line development timelines and reduce associated costs.

Assays currently used to identify high titre clones (e.g. ELISA, qPCR) cannot be performed on polyclonal populations, which creates a requirement for arduous generation of thousands of single clones. In this project, we propose a strategy that could take advantage of the simplicity and inexpensiveness of the lower throughput methods but can be applied in a HTP scheme. The hypothesis utilises the natural ability of lentiviruses to integrate into high-transcribing regions within human cell genomes. Lentiviral vectors carrying a reporter gene are used as a tool to target and identify those defined loci. Beforehand pre-identification of genomic positions could reduce screening workloads and contribute to generate cell lines in a more reproducible manner.

In addition to this semi-targeted approach, the system is coupled to a reliable method of clonal labelling and detection based on cellular barcoding. The barcode system consists of inserting a partially random DNA sequence tag into viral vectors. Upon transduction, the tag is stably integrated into the genome of a target cell and is inherited by its progeny. Expression values derived from barcode counts upon integration are ranked and correlated to the genomic position of integration. This way, site-specific expression can be measured in parallel in polyclonal populations avoiding the need to generate and screen thousands of clones. Next, transfer vector backbone is site-specifically integrated into these well-expressed candidate loci via CRISPR-Cas9 and co-transfected with the rest of packaging plasmids to assess viral titers.

The aim of this results chapter is to develop a pool of unique third generation lentiviral vector particles, namely a vector library, with tags incorporated in a suitable position of the vector that facilitates post-integration retrieval.

3.2 Aims

The specific aims for this chapter are:

- To engineer a lentiviral transfer vector expressing eGFP carrying a unique identifiable DNA tag (barcode) that enables integration site tracking.

- To construct a barcoded plasmid and viral barcoded library with sufficient size and complexity to screen integration sites at high-throughput sequencing scale.

- To characterise the complexity and composition of such barcoded libraries at a plasmid and viral vector stage.

- To determine whether the addition of a foreign DNA sequence tag into the 3'LTR of the lentiviral vector has effect on functional titers.

- To establish a transduction protocol able to deliver a single copy of barcoded viral vector into host cell lines.

3.3 Construction of lentiviral barcoded libraries

3.3.1 Cloning of pRRL SIN cPPT EFS eGFP WPRE

Firstly, a third generation lentiviral vector[146] was engineered to contain a reporter gene under the control of an internal promoter. The self-inactivating vector pRRL SIN cPPT PEW (Appendix A), a gift from Didier Trono's laboratory, contains a chimeric 5'LTR formed by the U3 of the Rous sarcoma virus (RSV) region joined to the HIV-1 R and U5 region (Figure 3.1A). The vector also contains a central polypurine tract (cPPT) and a reporter gene (enhanced GFP) under the control of the phosphoglycerate kinase promoter (PGK). The woodchuck hepatitis virus posttranscriptional regulatory element (WPRE) present downstream of the eGFP reporter gene has been reported to improve vector titers and expression levels of the transgene[298].

The original aim of this project was to use the lentiviral barcoding strategy to identify genomic sites able to support high levels of antibody production. The EF1alpha promoter was cloned into the vector given that stable expression of high levels of protein has been achieved with this promoter in CHO cells[697]. However, an intron-deleted version of the promoter region from the human translation elongation factor 1 α subunit (EF1α) (EFS, EF1α short form) was used in order to prevent inefficient splicing of the EF1α intron from the lentiviral genomic RNA[698], leading to differentially spliced vector products. The EFS version has successfully been used in the context of SIN lentiviral vectors by a number of academic laboratories and presents a safer insertional mutagenesis safety profile[698–701]. In addition, it has been observed that levels of transgene expression associated to the EF1α promoter are higher than PGK[702].

The aforementioned vector was double-digested with *XhoI* and *BamHI* in order to replace the original PGK promoter with the intron-deleted (short) version of promoter region from the human translation elongation factor 1 α subunit (EF1α) (EFS). This was previously PCR-amplified adding both restriction sites from pCCL EFS hIDUA (Appendix A), kindly provided by Maria E Alonso-Ferrero (Institute of Child Health/UCL, London, UK), with the following primers (fwd 5´-TCAGTctcgagGATTGGCTCCGGTGCCCG-3'; rev 5'-ggatccCGCGTCACGACAC-3'; restriction sites highlighted in lowercase). The resulting plasmid, pRRL SIN cPPT EFS eGFP WPRE (Figure 3.1), as well as parental plasmids were test-digested and fully sequenced to verify the integrity of its sequence and components (data not shown).

The resulting vector (pRRL SIN cPPT EFS eGFP WPRE) was then cut with *AvrII* and *Acc65I* in order to replace the original 3'LTR fragment with one including two unique restriction sites within the U3 region in the 3'LTR (*XbaI* and *NdeI*, separated by 6bp) in a position where Somers *et al.*, had previously integrated loxP sequences with no effects in viral integration and expression in 2010[703]. This position corresponds to 1bp downstream from the 400bp U3 deletion that gave rise to the SIN generation of lentiviral vectors by Zufferey *et al.,*[288] (Figure 3.1C). Such modification was introduced to allow directional sticky-end cloning of barcodes within the U3 region of the 3'LTR. The modified or 'synthetic' LTR sequence was synthesised by GeneArt (Invitrogen, Thermo Fisher Scientific) delivered in plasmid DNA, pMK-RQ KpnI-LTR-AvrII, subsequently digested with *AvrII* and *Acc65I* and ligated into a pRRL SIN cPPT EFS eGFP WPRE backbone to give rise to pRRL SIN SyntLTR cPPT EFS eGFP WPRE (Figure 3.1B).

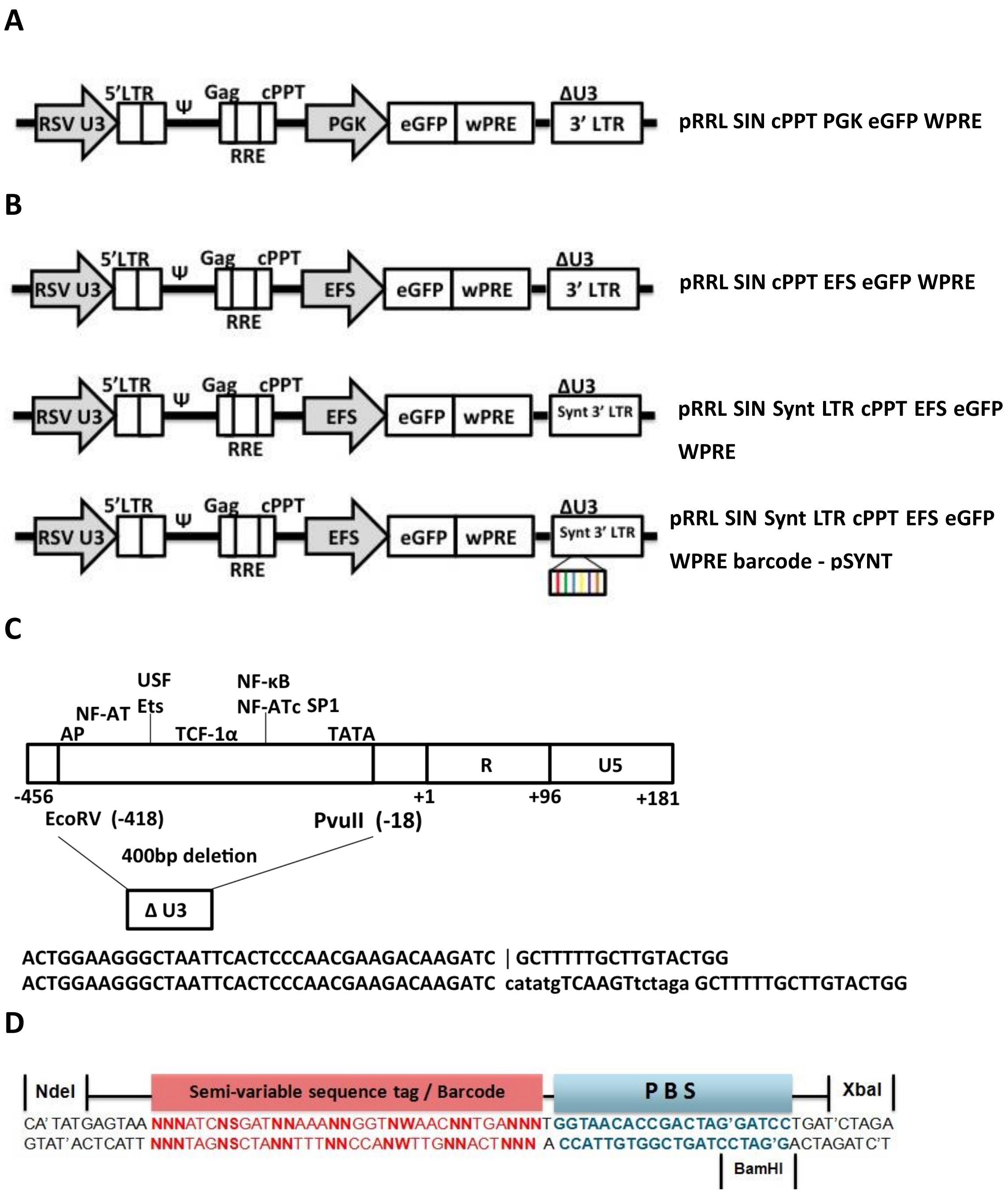

A
RSV U3
5'LTR
Ψ
Gag
cPPT
RRE
PGK
eGFP
wPRE
ΔU3
3' LTR
pRRL SIN cPPT PGK eGFP WPRE
B
RSV U3
5'LTR
Ψ
Gag
cPPT
RRE
EFS
eGFP
wPRE
ΔU3
3' LTR
pRRL SIN cPPT EFS eGFP WPRE
Synt 3' LTR
pRRL SIN Synt LTR cPPT EFS eGFP WPRE
pRRL SIN Synt LTR cPPT EFS eGFP WPRE barcode - pSYNT
C
USF
Ets
NF-AT
AP
TCF-1α
NF-κB
NF-ATc SP1
TATA
R
U5
-456
+1
+96
+181
EcoRV (-418)
PvuII (-18)
400bp deletion
Δ U3
ACTGGAAGGGCTAATTCACTCCCAACGAAGACAAGATC | GCTTTTTGCTTGTACTGG
ACTGGAAGGGCTAATTCACTCCCAACGAAGACAAGATC catatgTCAAGTtctaga GCTTTTTGCTTGTACTGG
D
NdeI
Semi-variable sequence tag / Barcode
P B S
XbaI
CA' TATGAGTAA NNNATCNSGATNNAAANNGGTNWAACNNTGANNNTGGTAACACCGACTAG'GATCCTGAT'CTAGA
GTAT'ACTCATT NNNTAGNSCTANNTTTNNCCANWTTGNNACTNNN A CCATTGTGGCTGATCCTAG'GACTAGATC'T
BamHI

Figure 3.1. Vector maps of third generation lentiviral vector plasmids expressing GFP and their subsequent modifications.

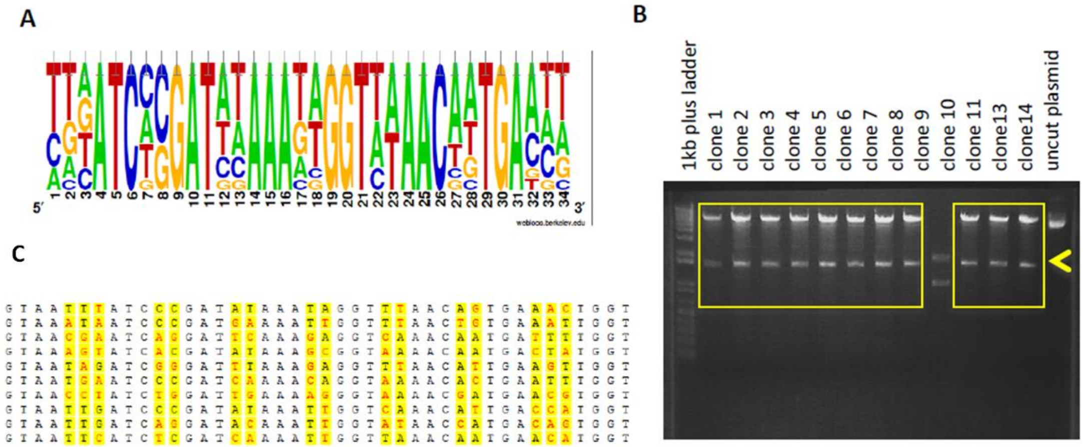
(A) Parental plasmid. RSV U3, Rous sarcoma virus U3 long terminal repeat promoter regions; 5'LTR, HIV-1 5' long terminal repeat; Ψ, HIV-1 RNA packaging signal; SIN, self-inactivating (U3-deleted) HIV-1 long terminal repeat; cPPT, central polypurine tract; Gag, HIV Gag gene; RRE, Rev responsive element; eGFP, enhanced green fluorescent protein; WPRE, woodchuck posttranscriptional regulatory element. **(B)** Intermediate plasmids generated in this study. **(C)** Diagram of the 3' LTR and the SIN lentiviral vectors 400bp deletion. Modified (synthetic) LTR with *NdeI* and *XbaI* sites indicated in lowercase **(D)** Lentiviral barcoded library containing the semi-random variable sequence tag (barcode) in the ΔU3 3'LTR. Schematic of the barcode. PBS, primer binding site. W is the nucleotide code for A/T; S is the nucleotide code for G/C.

The barcode system is based on a synthesized variable non-coding DNA sequence tag or 'barcode', formed by a semi-random 68mer non-coding DNA barcode library, with the purpose of uniquely and individually tagging and tracking integrated proviral vector genomes in host cell lines. On stable chromosomal integration, each vector will introduce a unique, identifiable and heritable mark into the host cell genome. The barcode consists of a semi-random 34bp sequence followed by a 20bp common 'anchor' sequence and it is based on a previous barcode construct by Gerrits *et al.*,[659]. The Gerrits *et al.*, design alternates triplets of known nucleotides with variable positions and labels each cell while the latter acts as primer binding site (PBS, not related to the viral primer binding site). This way, the occurrence of erroneous restriction sites within the barcode is minimised and barcode-positive and negative clones can be easily distinguished by PCR. Downstream bioinformatics analysis after integration site analysis and RNA-Seq is also facilitated by this barcodes configuration (Figure 3.1D).

The semi-random variable sequence tag design used in this study consists of 14 positions with 4 potential nucleotides and 2 positions with 2 potential nucleotides which theoretically make up to 1,073,741,824 ~ 10^9 (4^{14} x 2^2) possible combinations or variants of barcode sequences (Figure 3.1D). However, the complexity is limited in practice by other steps such as the viral titer or the size of the subcloned plasmid library (number of bacterial clones generated on transformation). This configuration was chosen in order to maintain a balanced GC content across the barcode as well as to prevent the formation of secondary structures and the accidental generation of restriction sites. The fixed triplets included within the variable nucleotides allow unambiguous sample identification and are also meant to facilitate the analysis of sequencing results by providing an internal standard to evaluate the quality of each sequence trace. In terms of nucleotide composition, the signature of each vector particle needs to be sufficiently distant to be able to distinguish it from other similar signatures or 'false' signatures originated from sequencing errors.

As an initial test to assess feasibility of oligonucleotide cloning, equimolar amounts of individual strands of barcode were dissolved in 0.5x ligation buffer at a final concentration of 10μM, heated at 95°C, annealed gradually at decreasing temperature (1°C/min), phosphorylated and ligated to column-purified *NdeI* and *XbaI*-digested pRRL SIN SyntLTR cPPT EFS eGFP WPRE backbone. The ligation product was transformed into *Stbl*3 competent cells and cultured in agar plates with the appropriate antibiotic. The resulting plasmid was named pRRL SIN SyntLTR cPPT EFS eGFP WPRE barcode (Figure 3.1B).

The following day, 97 colonies had grown and 8 and 0 colonies were observed in backbone-only and insert-only control ligations, respectively. The presence of another *BamHI* site 1.5kb away from the barcode *BamHI* site revealed >90% of the clones contained the barcode (Figure 3.2). DNA from 10 positive clones was extracted and Sanger-sequenced with the primer (5'-TGTGTTGCCACCTGGATTCTG-3') demonstrated that all clones contained a different barcode variant and all nucleotides were evenly represented in the variable positions (Figure 3.2A and Figure 3.2C). These results demonstrate a successful cloning strategy for barcode oligonucleotide cloning and barcode detection.



```
GTAATTTATCCCGATATAAATAGGTTTAACAGTGAAACTGGT
GTAAATAATCCCGATGAAAATTGGTTTAACTGTGAAATTGGT
GTAACGAATCAGGATTCAAAGAGGTCAAACAATGATTTTGGT
GTAAAGTATCACGATATAAAGCGGTAAAACAATGACTATGGT
GTAATAGATCGGGATTTAAAGAGGTTTAACATTGAAGTTGGT
GTAATGAATCCCGATCAAAACAGGTAAAACACTGAATTTGGT
GTAACCTATCTGGATTGAAAAGGGTAAAACGATGAACGTGGT
GTAATTGATCCCGATATAAATTGGTCAAACATTGACCATGGT
GTAATTGATCAGGATACAAATTGGTATAACCATGACAGTGGT
GTAATTCATCTCGATCAAAATTGGTTAAACAATGAACATGGT
```

Figure 3.2. Confirmation of barcoded oligonucleotide library cloning into a lentiviral backbone.

(A) Pictogram of relative frequencies of nucleotides in sequenced barcodes (resource available at: weblogo.berkeley.edu.logo.cgi[704]). Adenine (A) is shown in green; cytosine (C) in blue; thymidine (T) in blue and Guanine (G) in yellow. (B) Diagnostic digest of individual minipreps (clones 1-14) with *BamHI* to check the subcloning of barcode. The 1 kb plus ladder is used in agarose gels as molecular weight standards (Life Technologies, ThermoFisher Scientific, Appendix B). (C) Barcode sequences obtained by Sanger-sequencing of 10 clones. Nucleotides highlighted in yellow represent the variable positions within the barcode sequence.

3.3.2 Construction of a lentiviral barcoded library

Once the barcode was demonstrated to be cloned into the lentiviral backbone, the next step was to provide it with sufficient barcode variants in order to allow high-throughput screening of host cell line genomes. To generate the barcoded library, double stranded inserts containing the barcode (Barcode_top 5’Phos-TATGAGTAANNNATCNSGATNNAAANNGGTNWAACNNTGANNNTGGTAACACCGACTAGGATCCTGAT-3’ and barcode_bottom 5’Phos-CTAGATCAGGATCCTAGTCGGTGTTACCANNNTCANNGTTWNACCNNTTTNNATCSNGATNNNTTACTCA-3’) were synthesized by annealing two pools of phosphorylated oligonucleotides.

Although the designed barcode is theoretically capable of harbouring a barcode population of 4^{14} x 2^{2} = 1,073,741,824 > 10^{9} variants, the size of the barcoded library will be limited at different stages due to technical limitations such as the transformation efficiency of annealed barcodes into competent bacteria or the number of viral particles generated in a lentiviral vector preparation. Several conditions were tested including insert:backbone molar ratio, ligation times and temperatures in order to increase transformation efficiency (Figure 3.3).

A common issue described in the literature is the ligation of repeated copies of oligonucleotide insert forming concatemers. However, no concatemers were observed in any of the colonies sequenced after barcode cloning. The different number of overhanging base pairs in the restriction site may have contributed towards the absence of concatemers, which was also confirmed by gel electrophoresis (data not shown) and next-generation sequencing (Section 3.4.1).

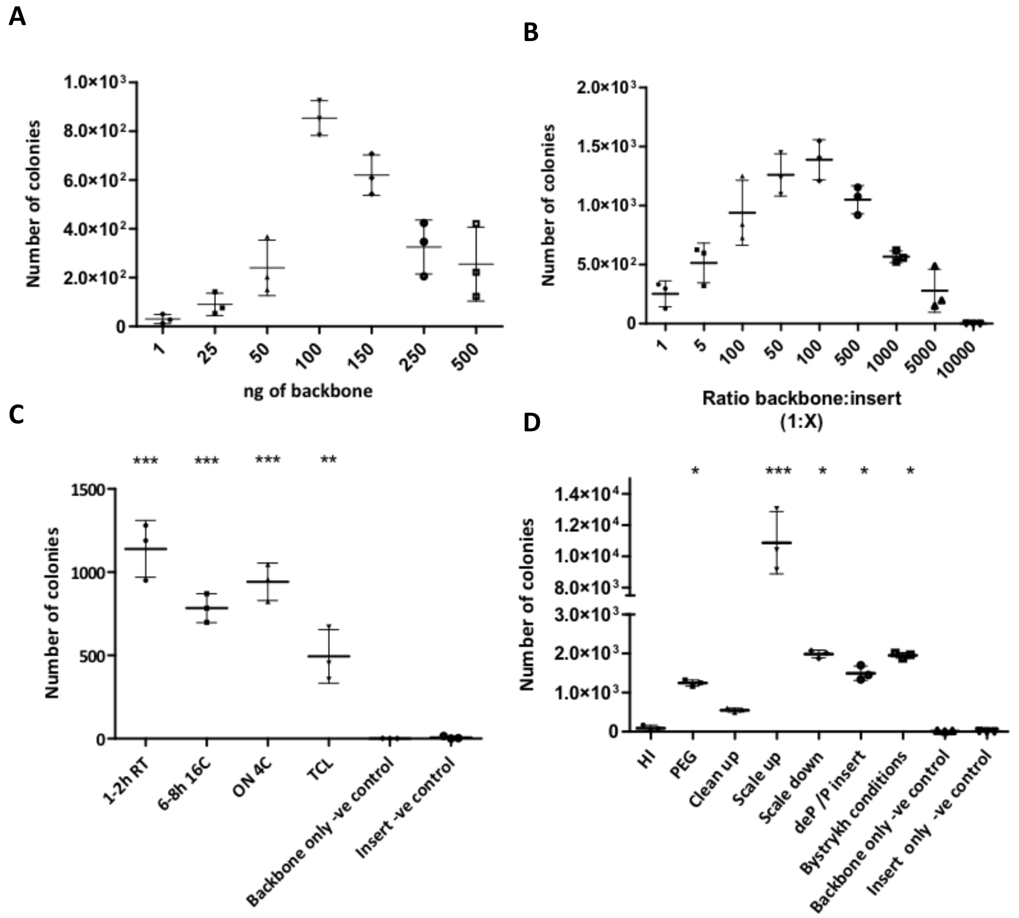


Figure 3.3. Optimisation of different parameters to generate the barcoded pSYNT library.

(A) Number of obtained clones using increasing amounts of lentiviral backbone plasmid DNA. (B) Number of obtained clones increasing insert:backbone ratios using 100ng of lentiviral backbone. (C) Number of colonies obtained using different times and temperatures for ligation (100ng backbone and 1:100 backbone:insert ratio). Temperature-cycle ligation (TCL) method described by Lund *et al.*, 1996 for oligonucleotide library cloning consists of 12-16h ligations, alternating 30 seconds at 10°C and 30 seconds at 30°C at a frequency of 10 cycles/h[679]. (D) Number of colonies obtained using varied methods reported in bibliography. Colonies were obtained after ligation 100ng of backbone with 1:100 ratio of oligonucleotide for 1-2h at room temperature. Different ligation protocols suggested in bibliography: HI, heat inactivation of T4 DNA ligase; PEG, addition of polyethylene glycol 10% (v/v) of 50% (w/v); Scale up, 10x scale up ligation reaction; Scale down, 2x scale down reaction; deP Backbone and P insert, standard ligation carried out with dephosphorylated backbone and phosphorylated oligonucleotide; conditions suggested by L. Bystrykh comprise a 2h ramp from 22°C to 18°C followed by an overnight incubation at 18°C; backbone negative control, only backbone ligation; Insert negative control, only insert ligation. All transformations were performed using homemade Stabl3 chemically competent cells (transformation efficiency 1.5x10^7) prepared following Inoue protocol[680]. All results presented (means ± SD; * p<0.05, ** p<0.001, *** p<0.0001, compared to backbone only negative control, one-way ANOVA and the Post hoc Dunnett's test) correspond to 3 technical replicates.

Optimization experiments showed that a backbone:insert ratio of 1:100 of reported the best results when ligated for 1 hour at room temperature (Figure 3.3A, B and C). Several methods and protocols found in the literature (e.g. PEG, heat inactivation, scale up, de/phosphorylation, etc.) were also tested after having established optimal amount of vector backbone and insert ratio as well as ligation conditions (Figure 3.3D).

Polyethylene glycol (PEG) is an additive that can be added to blunt ligations to improve its efficiency by acting as a condensing agent and favouring intermolecular and intramolecular binding[705]. However, most manufacturers recommend not combining PEG when using electrocompetent cells or heat inactivation unless the ligation product is dialysed. The number of colonies obtained using 10% (v/v) of 50% (w/v) PEG was equivalent to that of the standard protocol. A 2-fold concentrated ligation reaction (50% reduction in volume) was performed also with the aim to create a more condensed environment and facilitate the approach between molecules but did not report a major improvement in the number of clones. In contrast to this, a 5x scaled up ligation was performed in parallel using a final volume of 100µL and resulted in a 7-fold increase in the number of clones. Heat inactivation or purification of the ligation reaction prior to transformation did not exhibit any effect in the transformation efficiency. Alternatively, an oligonucleotide library preparation protocol suggested by Leonid Bystrykh (University of Groningen, Netherlands) (consisting of a 2 hour ramp from 22°C to 18°C followed by an overnight incubation at 18°C) was also tested.

Another factor that might limit transformation efficiency is the amount of fully double-stranded insert caused by mismatches in the variable sequence of the barcode that impair proper oligonucleotide annealing. Three strategies involving extension of the second strand over a single strand of barcode template were tested in order to address that potential problem (Figure 3.4).

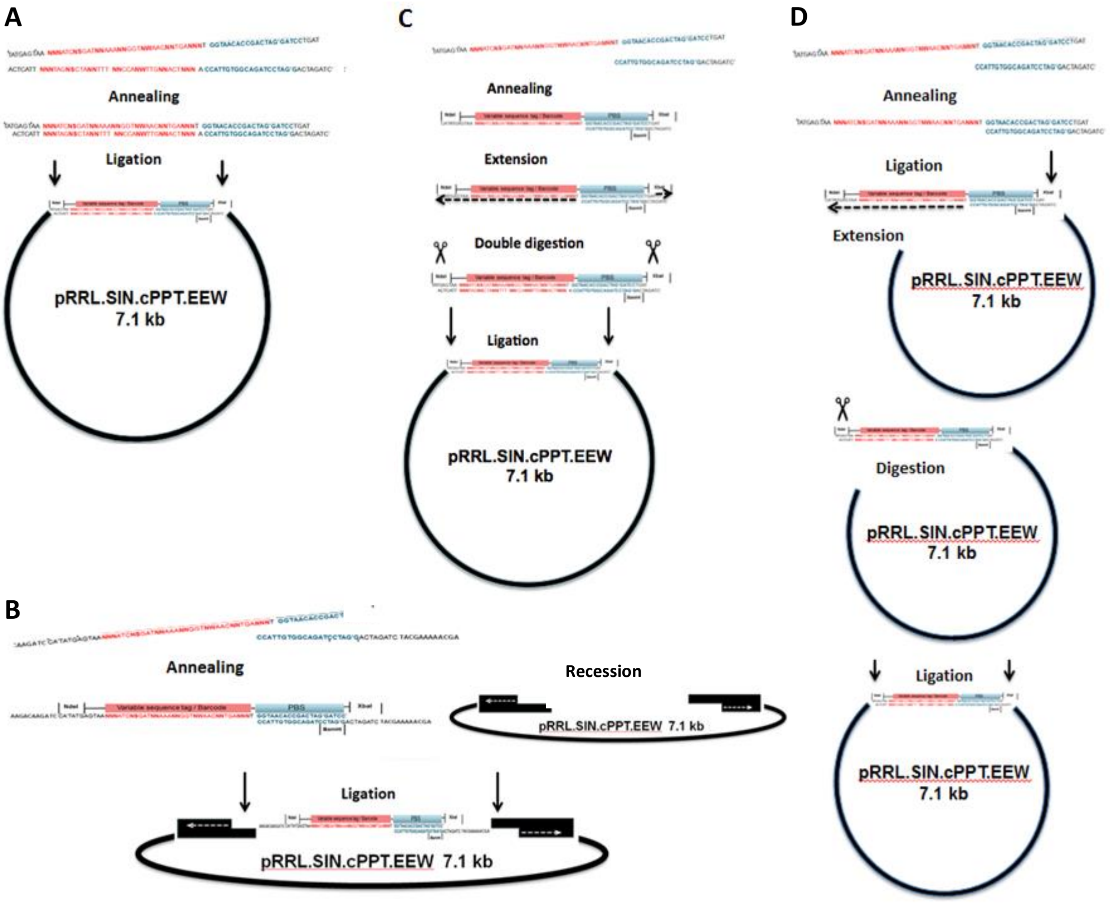


Figure 3.4. Cloning strategies for construction of the barcoded plasmid library.

(A) Standard strategy of oligonucleotide cloning. Annealing of a pair of oligonucleotides containing the complete barcode sequence resulting in a fragment with compatible ends to be ligated into the vector backbone (B) Alternative strategy for oligonucleotide library cloning consisting in the annealing of a pair of oligonucleotides not comprising the variable barcode region. The 3' ends of the vector backbone were recessed to increase the number of overlapping nucleotides and thus transformation efficiency. (C) Second alternative strategy for oligonucleotide library cloning. Annealing of a pair of oligonucleotides not comprising the variable region of the barcode followed by the extension of the barcode variable sequence and consequent digestion of both ends and ligation into the vector backbone. (D) Third alternative strategy differing from (C) only in a ligation of an annealed oligonucleotide fragment end prior to strand extension, digestion and ligation of the other compatible end.

However, none of the alternative cloning strategies previously described (Figure 3.4) reported a significant increase in the number of colonies achieved (data not shown) and thus were discarded due to their added complexity. Eventually, a 10x scaled up ligation reaction was prepared following the same stoichiometric proportions with a final volume of 200µL. The whole reaction was transformed using 2mL of *Stbl3* chemically competent cells and the estimated number of colonies was 120,000 (based on a 10^{-3} diluted culture, assuming 100% plating efficiency) representing a 10-fold increase in respect to the previous scale up reaction. In all cases, insert and backbone (recircularisation) negative controls remained low with close to 10 and 25 clones, respectively.

Clones were screened for the presence of barcode by restriction digest and sequencing resulting in 85-90% of positive clones harbouring different variants in all cases. No concatemerisation of the barcode was observed. The barcoded plasmid library was named pSYNT. Although a library size of 10^5 clones is far from the initial expectations to reach the 10^9 potential variants (theoretical sample space), it is sufficiently high to represent a better alternative to current high-throughput clone selection methods (screen capacity of 10^3-10^4 clones[636]).

The pool of colonies was grown overnight in 125mL of vLB in the presence of 50μg/mL of ampicillin in order to generate a plasmid library glycerol stock (124 x 1mL aliquots) stored at -80°C. After overnight culture, an aliquot was streaked out on agar plates containing the appropriate antibiotic and colonies were screened for the presence and composition of barcode. Plasmid DNA from each picked bacterial clone was isolated and Sanger-sequenced and analysed in order to pre-validate the library prior to next-generation sequencing. 90% of the colonies presented barcode, no concatemers were observed and all barcode variants were different. The four nucleotides were equally represented in the variable positions (data not shown).

3.4 Characterisation of barcoded lentiviral libraries

3.4.1 Barcoded plasmid library validation by next-generation sequencing

Next-generation sequencing was performed on the barcode fragment to assess the library size and complexity of the pSYNT barcoded plasmid library. Aliquots 11 and 49 of the plasmid barcoded library (124 aliquots) were randomly selected for plasmid preparation (named pSYNT11 and pSYNT49). A 187bp fragment containing the barcode sequence was amplified from pSYNT11 and pSYNT49 with primers containing Illumina compatible sequences (highlighted in uppercase): (P5fwd-upstream-barcode 5'-ACACTCTTTCCCTACACGACGCTC

TTCCGATCTccttcgccctcagacgagtcggatc tc-3’; bio-P7rev-downstream-barcode 5’-[bio]GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTcaaaaagcatctagatcaggatcctag tcggtgttacca-3’). The 220bp PCR product was submitted to UCL Genomics for next-generation sequencing with the MiSeq using a 300bp paired-end (PE) configuration.

Fastq files of the two PCR products (named Plasmid_PCR11 and 49), obtained from pSYNT11 and 49 barcoded regions, were analysed using FASTQC tool[690] in Galaxy[686] yielding 1,752,655 and 1,135,377 reads, respectively. R1 and R2 reads were merged with BWA pemerge yielding 1,578,532 (90%) and 1,004,241 (88.5%) successful merges for PlasmidPCR11 and PlasmidPCR49, respectively. Quality control of merged reads revealed an optimal mean quality of the reads across the bases of the PCR amplicon, including an average size of nucleotide (after merging), balanced GC content and a high amount of repetitive sequence as expected by the low complexity of the regions surrounding the barcode (Figure 3.5).

Barcodes were extracted from merged reads using a custom Perl script, ‘extract_library_barcode.pl’ (Appendix B). This script detects a DNA string consisting of a defined 20bp sequence upstream of the barcode followed by the barcode itself and a defined 20bp sequence downstream. The efficiency of vector barcoding (or the number of vectors with barcodes) was 88% according to deep sequencing, confirming results obtained at low throughput by enzymatic digestion and Sanger sequencing (8% unbarcoded vectors). 1,392,401 and 892,174 barcodes (88.2 and 88.8% of the sequences from the previous step, respectively) were extracted from Plasmid_PCR11 and Plasmid_PCR49, which have a total of 89,207 and 65,410 variants (6.4 and 7.3% of the sequences from the previous step respectively), of which 12,019 overlapped between the two replicates (Figure 3.6A).

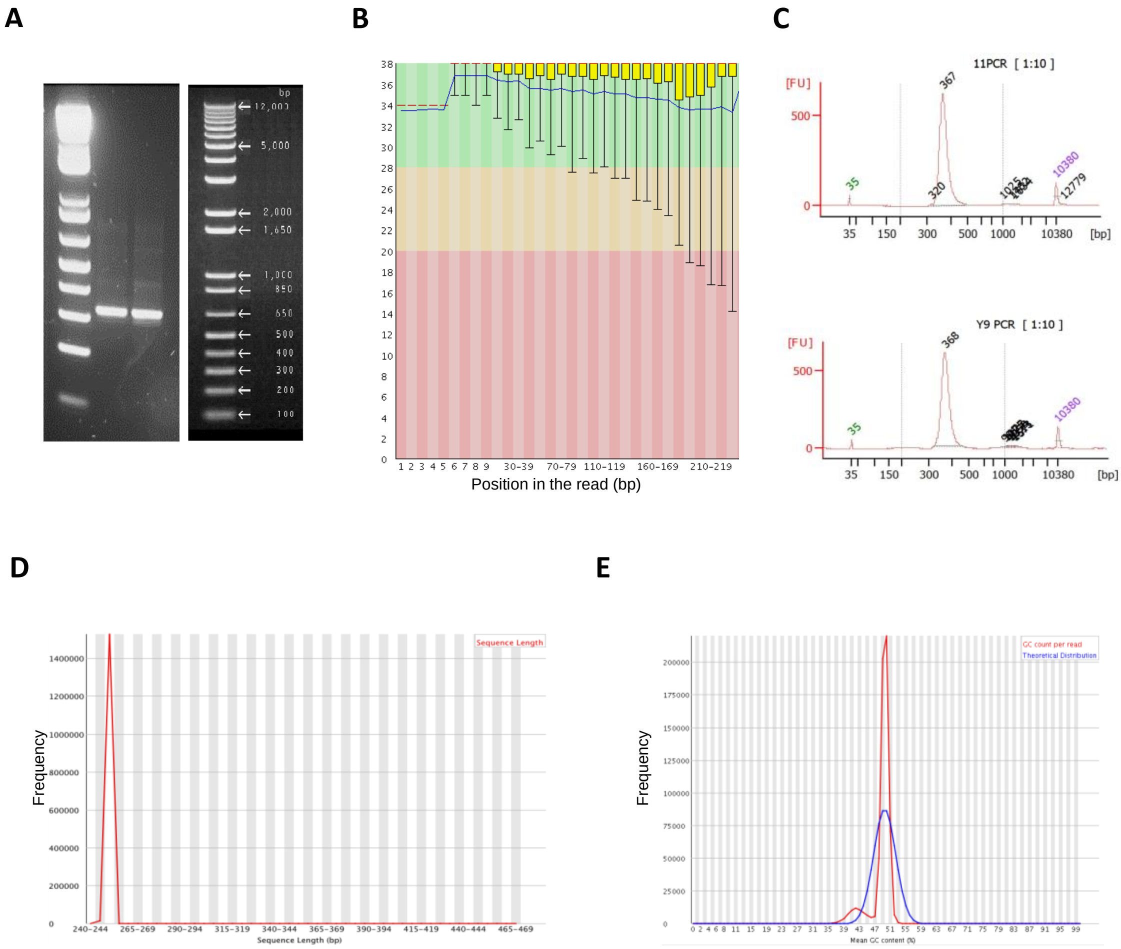


Figure 3.5. Summary of quality control statistics for barcoded Plasmid_PCR11 and 49 libraries.

(A) 1% agarose gel electrophoresis showing a 320bp barcoded plasmid PCR products. 1kb plus DNA ladder (Life technologies, ThermoFisher Scientific, Appendix B). (B) Distribution of quality values per base. The yellow boxes represent the inter-quartile range (25-75%). The upper and lower whiskers represent the 10th and 90th percentiles; the blue line represents the mean quality. The background of the graph divides the y-axis into good quality calls (green), calls of intermediate quality (orange), and calls of poor quality (red). (C) Peaks of DNA obtained during DNA quality assessment by micro chip-based capillary electrophoresis using Agilent Bioanalyzer. Results given in fluorescence units (FU). Peaks at 35bp and 10,380bp are internal controls. (D) Distribution of sequence length over all sequences. Equivalent results obtained for PlasmidPCR_49 (data not shown). (E) GC distribution over all sequences for PlasmidPCR_11. Equivalent results obtained for PlasmidPCR_49 (data not shown).

In both replicates, only 11% of the barcode variants (10,249 and 7,260, respectively) were recurrent (presented >1 replicate or copy) and thus, the majority were singletons (barcodes with a single variant). However, these singletons only contribute 6% to the total barcode population. This can mean that either a vast majority of recurrent barcodes show very low complexity and/or 6% of barcodes (those that are singletons) provide 90% of complexity to the library, although this complexity is falsly generated by sequencing errors. Interestingly, the error rate of Illumina NGS technology ranges between 1-5%.

A similar proportion of the 4 nucleotides was observed in all 14 variable positions except the 2 positions with W and S (IUPAC nomenclature) where only A/T and G/C were equally found, respectively (Figure 3.6B). These results contribute to establish a maximal nucleotide dissimilarity (based on Hamming distances[706,707]) in the barcode population thus enabling efficient barcode retrieval and clustering correction in subsequent bioinformatics analysis. Major biases in the nucleotide composition of the barcode can affect sequence complexity and yield (depending on the technology). However, the addition of a balanced and diverse PhiX DNA library allows real time control quality metrics and creates a more diverse set of clusters to normalize for the low diversity of the amplicons.

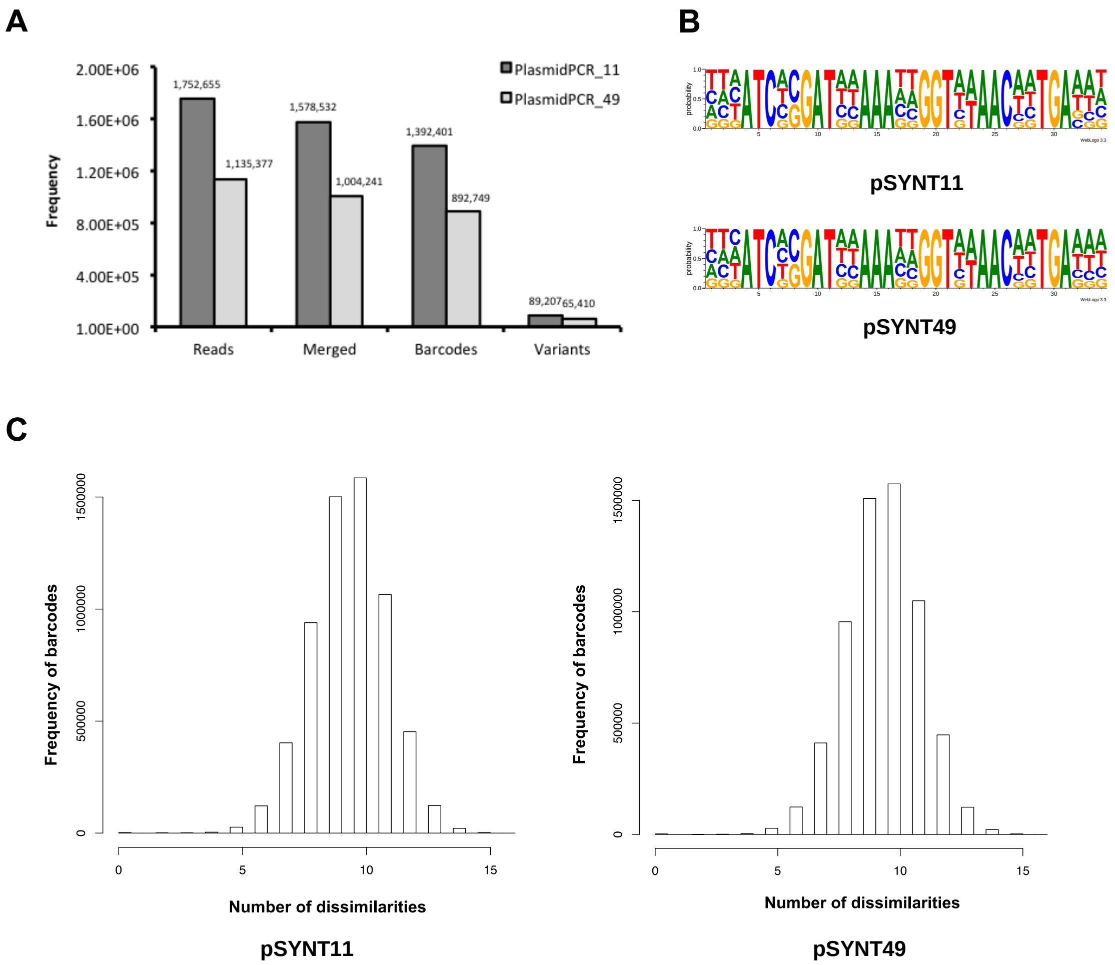


Figure 3.6. Characteristics of the pSYNT barcoded plasmid library analysed by next-generation sequencing.

(A) Number of reads obtained by high-throughput sequencing using a MiSeq 300bp paired-end strategy, successfully merged reads, extracted barcodes and unique barcode variants present in the two replicates analysed. (B) Pictogram of relative frequencies of nucleotides in sequenced barcodes Resource available at: weblogo.berkeley.edu.logo.cgi[704]. Adenine (A) is shown in green; cytosine (C) in blue; thymidine (T) in blue and Guanine (G) in yellow. Top pictogram corresponding to PlasmidPCR_11; bottom pictogram corresponding to PlasmidPCR_49. (C) Mean number of nucleotide differences between all the sequenced pSYNT11 (left) and pSYNT49 (right) barcodes.

Authors differ in the way quality criteria are applied to correct for low frequency barcode noise (different criteria detailed in Table 3.2). Arbitrary removal of low frequency variants (i.e. variants occurring <2 and up to <10 times) biases the real library complexity by underrepresenting the number of ‘true’ singletons (non-false variants occurring once/non-recurrent in the population).

Alternatively, the degree of dissimilarity could be used to discard variants close in sequence. The number of nucleotide differences between the barcode variants present in the vector library was also evaluated using a script kindly provided by Martijn H Brugman (Leids Universitair Medisch Centrum (LUMC), Germany) (Appendix B). A peak of dissimilarity around 11 nucleotides can be observed in both replicates (Figure 3.6C). This value provides the average variation expected between barcodes and could serve as a reference to establish the threshold for background noise removal. However, the hierarchy in number of counts needs to be considered not to discard predominant variants with this approach.

A potential way to minimize this bias is to integrate low frequency variant counts into the counts of their parental counterparts by clustering correction. Clustering correction aims for compensation for the appearance of false positives derived from sequencing errors. The algorithms group together sequences based on biological relationship or error threshold and assign to a mother/stem barcode. Its principle is based on the number of single-character edits necessary to change one sequence into another[708]. The Levenshtein (or editing) distance is computed between all the sequence pairs in a sequence population yielding as an output a canonical sequence (with minimal distance) and a set of several DNA sequences whose metrics are below the established threshold. Read counts of barcode variants different by less than a particular number of nucleotides (threshold) are pooled together and pictured as a network of nodes linked by edges whose distance is proportional to the number of dissimilarities between sequences. Barcode clustering consists of a computationally intense 'matching' phase in which a graph is displayed (typically resembling a star shape), followed by a cluster detection phase. Several clustering algorithms exist in the literature to correct sequencing errors from random sequences (or sequences of unknown source) when a reference library or genome is not available.

Starcode carries out an all-pair search for the number of Leveinshtein distances between sequences to construct the clustering diagram[691]. Matching is performed using lossless filtration but what makes Starcode novel is the 'poucet' algorithm search strategy, which significantly reduces time compared to other clustering algorithms. Input sequences are prefixed and these intermediates sorted and stored in alphabetical order in the edit matrix according to their prefix redundancy so that less computational effort is required to process the next search.

As discussed before, distribution of barcode frequencies shows an accumulation of low frequency barcodes up to 60% of the total populations. In other words, the majority of the population is composed of barcodes variants with a few or a single copy (singletons). Starcode clustering (using default parameters) was applied for correction on Plasmid_PCR11 and 49 and as result, the consensus showed a more even distribution of low frequency barcodes (Figure 3.7). This confirms that a fraction of barcode variants had less than 2 distinct nucleotides and less than a fifth of relative frequency compared to their corresponding 'mother' barcode (default parameters for Starcode clustering) and were likely to be originated due to sequencing errors.

After clustering correction of barcode sequences, 39,954 and 28,173 unique barcodes (or variants) were identified in clustered barcode populations, of which 3,078 overlapped. In this case, only 3,053 and 3,046 were recurrent. The profile and peak number of dissimilarities is maintained when the barcodes of the plasmid library are clustered confirming their variant proportions (data not shown).

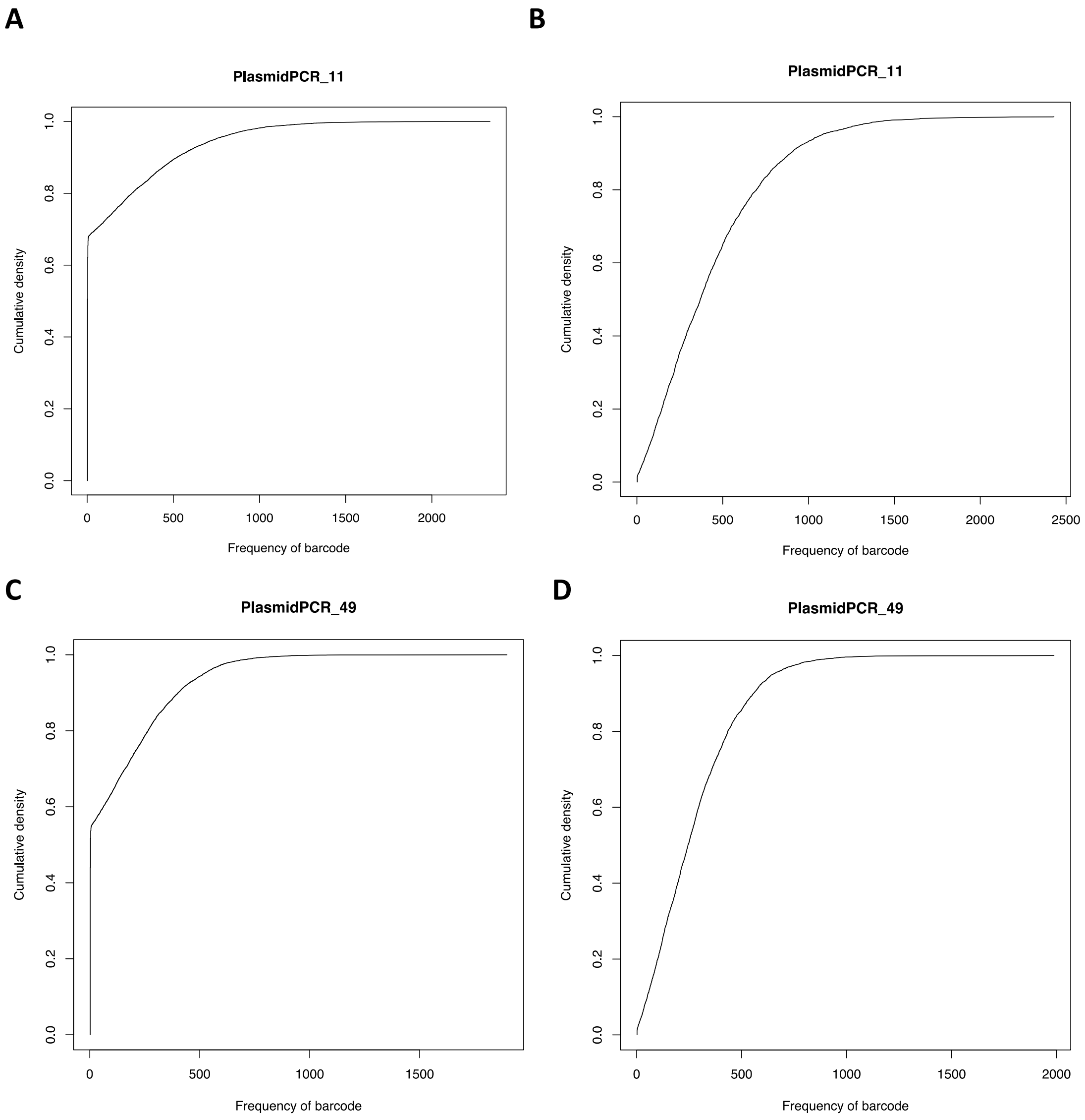


Figure 3.7. Density plot showing the cumulative frequency of barcode variants before (left) and after (right) clustering correction.

Barcode variant frequencies of plasmid libraries 11 (A) and 49 (C) expressed as cumulative percentage of the total barcode population. Starcode clustering eliminates low frequency variants (sequencing errors) in PlasmidPCR_11 (B) and Plasmid_PCR_49 (D).

Determining the library size or the number of possible barcode variants in a library is an essential step for library characterisation. The throughput or screening capacity of the lentiviral system will be dictated by the total number of distinct variants found in the library, therefore population size values under 10^3 would not signify an improvement over current non-parallel high-throughput screening procedures.

The Lincoln-Petersen estimate[709] was used in order to determine the size of the pSYNT library. The Lincoln-Petersen estimate is used in ecology to study population sizes based on the number of capture and recapture events. This model assumes the population is constant i.e. the population size remains the same between the time of the capture (mark) and the recapture implying that no individuals are born, die or migrate (which obviously does not apply in this study). Another premise of the model is that the sample is random and that all individuals have the same chances of being captured in the second sample regardless whether they were captured in the first place (independent recapture). Given the previous premises, the model predicts that the population size (N) can be calculated as follows:

$$N = \frac{M\,S}{R}$$

Where M is the number of events marked in the first 'capture', S is the number of events captured in the second sample or re-capture and R is the number of marked events (from the first sample) captured in the second sample. One of the strengths of the Lincoln-Petersen estimate is that it remains asymptotically unbiased at large sample sizes (which is the case and reason why it is used in this case over the Chapman estimator)[710]. However, despite its simplicity, population sizes calculated using the Lincoln-Petersen index tend to be overestimated, especially in samples with high heterogeneity whose spatial distribution is not uniform[711,712]. Some authors use the Schnabel index, which share the same principle but allows for multiple mark and recapture events. Applying these calculations, the population size of the pSYNT barcoded library is comprised between 485,484 and 365,700 variants (Table 3.1), which shares the same order of magnitude as the estimated number of bacterial colonies obtained in the plasmid library cloning.

Table 3.1. pSYNT library size and diversity details pre- and post-Starcode clustering correction.

	Non-clustered	Clustered (Starcode)
PlasmidPCR_11 barcode variants	89,207	39,954
PlasmidPCR_49 barcode variants	65,410	28,173
Overlapping variants	12,019	3,078
Recurrent variants PlasmidPCR_11	11.5%	7.6%
Recurrent variants PlasmidPCR_49	11.1%	10.8%
Population size (Lincoln-Petersen)	485,484	365,700

Libraries with random oligonucleotides represent a source of entropy useful to exploit high-throughput screening. Nevertheless, often the amount of diversity observed is only a small proportion of the total number of possible sequences in the sampling space. The following argument intends to estimate the real number of variants there are in a given library and also the theoretical size of a library to be 95% confident for example it contains all possible variants.

Regarding the number of physical molecules present in the sequencing reaction, typically, 600μL of a 12-20pM DNA library are sequenced under this configuration, which represents between 4.3-7.2 x 10^9 300bp molecules. This number is higher than the theoretical number of potential barcode variants in the library and also sufficiently high to be 95% sure a complete library could be fully sequenced; however, sequencing using the HiSeq technology still offers higher throughput number. Therefore, this step does not represent a bottleneck for the library size and/or complexity.

According to Patrick *et al.*,[713], given a library with L number of observed clones and providing all the variants in a library are equally represented, the mean number of occurrences of a sequence variant in the library (λ) can be defined as λ= L/V, where V is total number of possible distinct sequence variants. When λ<<L, the number of occurrences of a single sequence variant follows the Poisson distribution:

$$P(x) = \frac{e^{-\lambda}\,\lambda^{x}}{x!}$$

where P(x) denotes the probability of a library to be present x times in a library. From the previous expression, the probability of a variant not occurring in the population or occurring once can be expressed as:

$$P\ (0) = e^{-\lambda}$$

$$P\ (x \geq 1) = 1 - P\ (0) = 1 - e^{-\lambda} = 1 - e^{\ L/V}$$

Therefore, the expected or real number of different variants (C) can be described using the expression C = V P, where V represents all the possible combinations of the complete library and P is the probability of a variant to be expressed one or more times. In the case of this study, V would correspond to 14 and 2 positions with 4 and 2 possible nucleotides, respectively (4^{14} x 2^{2}), which results in 1,073,741,824 ~10^{9} possible combinations.

$$C = V\ P\ (x \geq 1) = V\ (1 - e^{\ L/V})$$

The probability of a barcode variant to occur one or more times, P (x ≥ 1), can also be understood as the fractional completeness of the library (F) or C/V. For example, a library with a completeness of 95%:

$$0.95 = (1 - e^{\ L/V})$$

$$L = -V\ \ln 0.05 = 3\ V$$

This means a 3-fold degeneracy is expected in a library to contain 95% of the clones. In our case, that would mean a library containing 3x10^{9} barcode variants should be constructed for it to contain 95% of the expected variants.

Approaching this question the other way around, we can deduce that the different barcodes sequenced in this library provide evidence that there are approximately L/3 = 400,000 clones according to Lincoln Petersen estimate divided by 3 equals 133,333 different variants in the library with 95% fractional completeness. This number actually defines the throughput of the library for further screening purposes.

Library calculation softwares are based on the assumption that all base substitutions and barcode variants are equiprobable. In reality, inherent to PCR amplification is the fact that some sequences are more easily amplified than others. As a consequence, the number of theoretically expected sequences (V) is greater than obtained sequences (sub-library size, L). If L >>V then, most variants are likely to be sampled unless the bias is very strong.

As seen in Figure 3.7, false low frequency barcode variants generated from errors in the amplification of barcoded sequence during NGS for library validation can represent an important proportion of the barcode population. Error rates (f) are calculated using the following formula:

$$Error\ rate(f) = \frac{n}{S}$$

where n is the number of mutations observed and S = (bp sequenced x d), d being the average number of doublings occurred in a reaction, which can be calculated from the expression below:

$$2^{d} = \frac{ng\ DNA\ after\ PCR}{ng\ of\ target\ DNA\ input}$$

Analytical error introduced by commercial polymerases in the sequencing process ranges from 10^{-5}-10^{-7}errors/bp[714] based on 15-20 doublings and 5kb sequenced. Adjusting the values to the 1.5M reads of 2 strands of 300bp, 35 cycles and assuming the polymerase used has an error rate of 10^{-5} (most error-prone case scenario), the total number of mutations per sequenced sample oscillates around 315,000. The probability of one of these mutations to occur in one of the 16 variable positions of the barcode is 16/300 = 2% (otherwise would have been discarded), which means 6,300 reads out of the initial 1.5M (0.42%) might contain a false barcode originated by polymerase error, a minimal proportion of the global barcode population.

In addition, errors in the barcode sequence are also introduced by the cellular RNA polymerase on transcription of plasmid molecules. These errors are not analytical but real mutations introduced in the barcode sequence that contribute to generate complexity in the library. However, since their rate is as low as those of analytical origin, the frequency and nucleotide dissimilarity of RNA-pol generated (real) barcodes would probably be underestimated by clustering algorithms.

3.4.2 Barcoded vector library validation by next-generation sequencing

Two aliquots of plasmid library (#11 and #49, randomly chosen) were amplified in vLB containing the appropriate antibiotic at 37°C overnight. Bacterial cells were pelleted and DNA was isolated to be transfected along with third generation lentiviral vector packaging plasmids as indicated in Materials and Methods (Section 2.2.16).

Ultracentrifuged lentiviral vector containing the barcoded library (vSYNT11 and 49) was prepared and titrated reporting titers comparable to a standard lentiviral preparation (Figure 3.8). No significant differences were observed between different vectors when titrated in different cell lines. These results indicate that the addition of a foreign sequence does not interfere with functional titers.

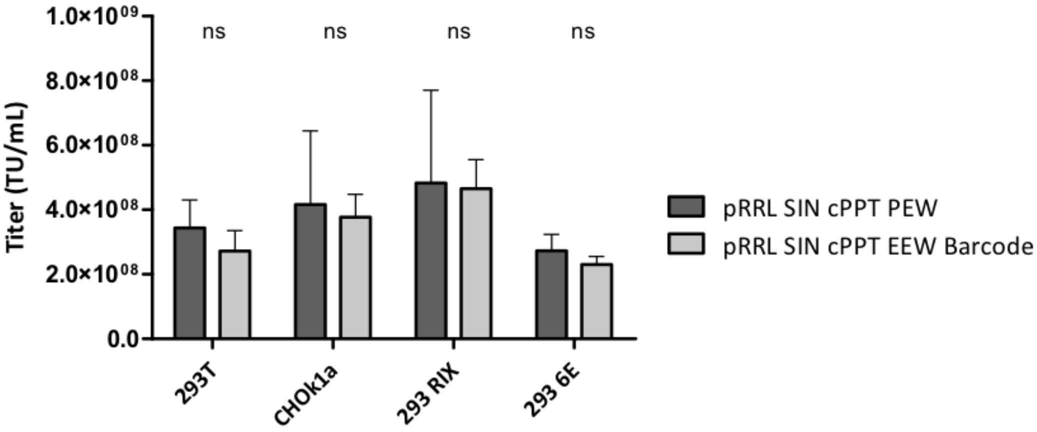


Figure 3.8. Functional lentiviral vector titration by flow cytometry on different cell lines.

All results presented (means ± SD; * $p<0.05$, ** $p<0.001$, *** $p<0.0001$, grouped per cell type, Friedman's test analysis of variance) correspond to 3 technical replicates. Values were similar for vSYNT49 (data not shown).

Viral vector genomic RNA was extracted from viral vector supernatant and DNA contaminants were removed by column purification or DNaseI. Reverse transcription and amplification of the desired barcoded sequence were performed in a single step using SuperScript® III One-Step RT-PCR System with Platinum® *Taq* DNA Polymerase and P5fwd-upstream-barcode RNAbc_150ups-fwd 5’-ACGAGTCGGATCTCCCTTTG-3’ and Barcode-PBS(*BamHI*)-rev 5’-GGATCCTAGACGGTGTTACC-3’ primers. A 220bp product was cloned into a TOPO backbone and successful retrieval of barcodes was confirmed by Sanger sequencing prior to submission of PCR product for deep sequencing (data not shown). NGS using a 100bp paired-end strategy with the HiSeq2500 sequencer yielded 12,267,014 and 11,641,759 reads for vector libraries 11 and 49, respectively. 12,235,295 (99.7%) and 11,611,553 (99.7%) R2 reads passed the criterion of having five or fewer Ns. Quality control showed good quality along all the basepairs of the 101bp reads and balanced GC content (Figure 3.9).

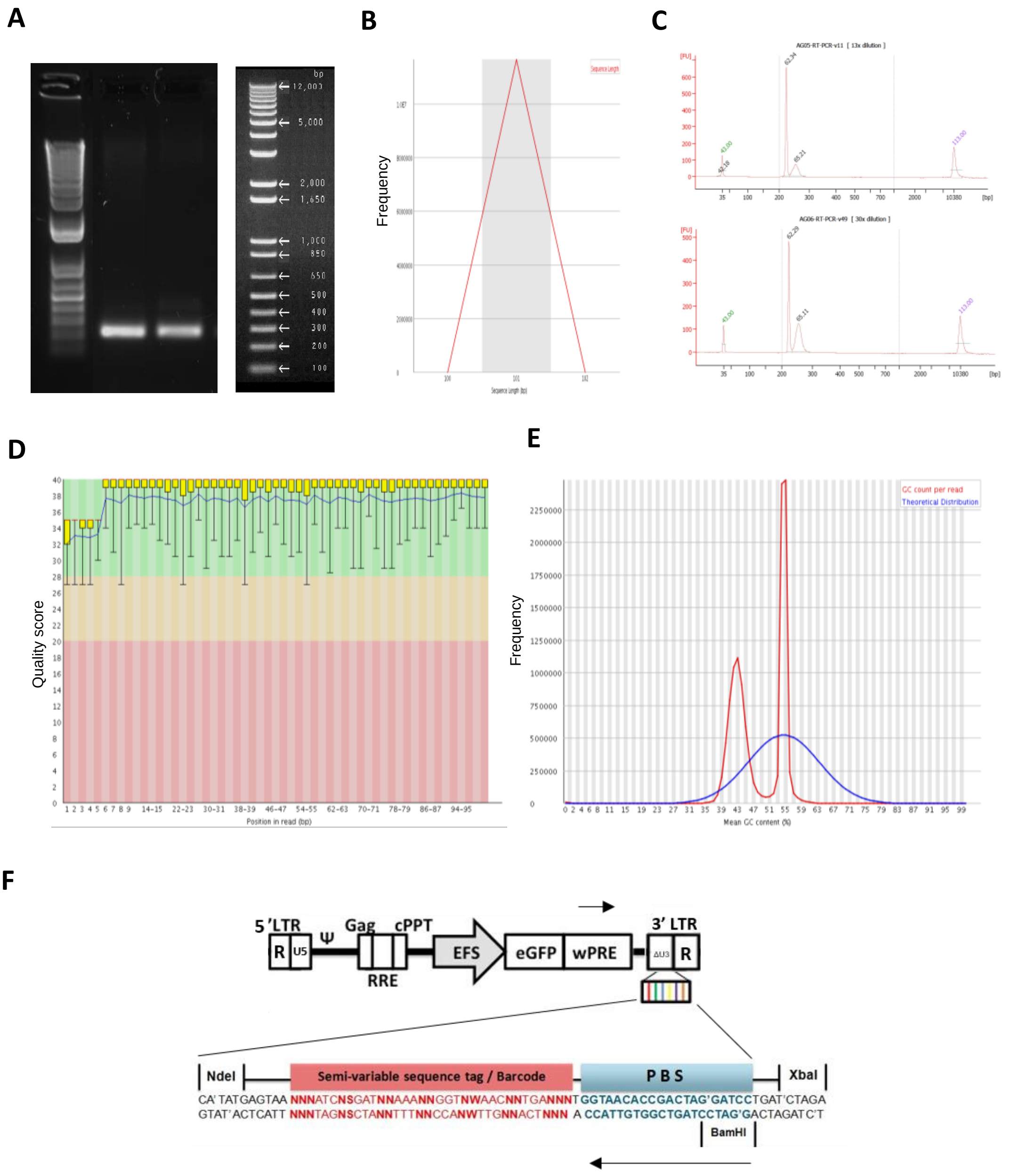


Figure 3.9. Summary of quality control statistics for barcoded vectorPCR_11 and 49 library.

(A) 1% agarose gel electrophoresis showing a 320bp barcoded plasmid PCR products. 1 kb plus DNA ladder (Life technologies, ThermoFisher Scientific, Appendix B). (B) Distribution of sequence length over all sequences. Equivalent results obtained for vectorPCR_49 (data not shown). (C) Peaks of DNA obtained during DNA quality assessment by micro chip-based capillary electrophoresis using Agilent Bioanalyzer. Results given in fluorescence units (FU). Peaks at 35bp and 10,380bp are internal controls. (D) Distribution of quality values per base. The yellow boxes represent the inter-quartile range (25-75%). The upper and lower whiskers represent the 10th and 90th percentiles; the blue line represents the mean quality. The background of the graph divides the y-axis into very good quality calls (green), calls of intermediate quality (orange), and calls of poor quality (red). (E) GC distribution over all sequences for vectorPCR_11. Equivalent results obtained for vectorPCR_49 (data not shown). (F) Location of the primers used to amplify barcoded regions from lentiviral vector RNA transcripts. LTR, HIV-1 long terminal repeat; Ψ, HIV-1 RNA packaging signal; SIN, self-inactivating (U3-deleted) HIV-1 long terminal repeat; cPPT, central polypurine tract; Gag, HIV Gag gene; RRE, Rev responsive element; eGFP, enhanced green fluorescent protein; WPRE, woodchuck posttranscriptional regulatory element; PBS, primer binding site. W is the nucleotide code for A/T; S is the nucleotide code for G/C.

For barcoded vector library 11 and 49 (vSYNT11 and 49), 5,061,108 (41.4%) and 4,688,154 (40.8%) barcodes were extracted from sequences containing the barcode regular expression (no flanking sequences) and 545,185 (10.8%) and 483,700 (10.3%) variants were counted, respectively, of which 105,261 overlapped. The number of singleton counts dropped from 90% in the plasmid library to 10-15% in the vector library; in other words, 458,841 (84.2%) and 403,001 (83.3%) variants had more than one copy in the vector library 11 and 49, respectively. This drop in the proportion of barcodes with a unique signature can be explained by the higher number of reads obtained by HiSeq sequencing. Sequencing errors are more likely to be repeated with 10 times more reads.

After clustering correction (using Starcode), the number of variants was reduced to 88,052 and 75,946, of which 3,211 overlapped between vector library 11 and 49 replicates (Figure 3.10A). This represents a 1.74% and 1.63% of the initial selected barcodes and a reduction of 80.8% and 81.1% in the number of variants, respectively. Despite the number of reads being 10 times higher, pSYNT and vSYNT retrieved Starcoded variants have the same order of magnitude and the difference between replicates remains consistent (pSYNT11 x 0.71 = pSYNT49; vSYNT11 x 0.86 = vSYNT49; calculations based on Starcode-clustered variants), confirming no major biases occurred during the lentiviral library preparation. The fact that the number of retrieved variants in the vector libraries doubled those obtained in the plasmid libraries could be attributed to the sampling difference.

The proportion of the four nucleotides in the library remains balanced (Figure 3.10B). This indicates again that no major biases were introduced during vector production. According to the dissimilarities plot, most barcodes differ from each other by 11 nucleotides (also at viral level) (Figure 3.10C), which contributes to a sufficient library complexity and eases the discrimination between true biological variants and variants originating from sequencing errors. The profile and peak number of dissimilarities is maintained when the barcodes of the vector library are clustered confirming their variant proportions (data not shown).

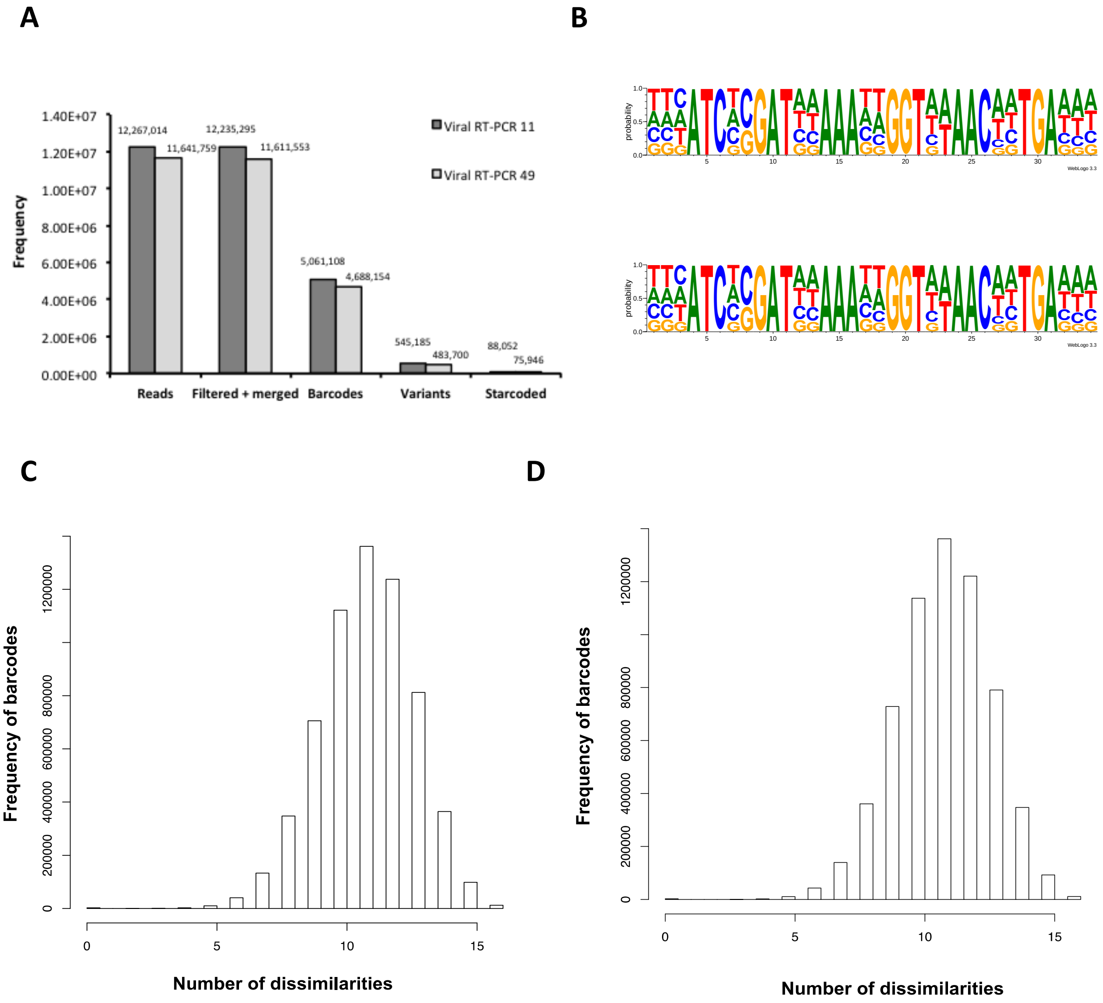


Figure 3.10. Characteristics of the vSYNT barcoded vector library analysed by next-generation sequencing.

(A) Number of reads obtained by high-throughput sequencing using a HiSeq 2500 100bp paired-end (PE) strategy, successfully merged reads, extracted barcodes and unique barcode variants present in the two replicates analysed. (B) Pictogram of relative frequencies of nucleotides in sequenced barcodes Resource available at: weblogo.berkeley.edu.logo.cgi.[704] Adenine (A) is shown in green; cytosine (C) in blue; thymidine (T) in blue and guanine (G) in yellow. Top pictogram corresponding to PlasmidPCR_11; bottom pictogram corresponding to PlasmidPCR_49. (C) Barcode comparison plot showing the number of nucleotide differences between all the sequenced barcodes. (D) Barcode comparison plot showing the number of nucleotide differences between all the sequenced vSYNT11 (left) and vSYNT49 (right) barcodes.

3.5 Delivery of barcoded lentiviral libraries

3.5.1 Transduction of the barcoded vector library into HEK293 host cell lines

Once the lentiviral barcoded library was produced and characterised, barcodes were delivered to cell lines in order to identify integration site preferences and be able to quantify their transcript expression relative abundance.

The distribution of vector particles across cells is a random process; the probability of cells receiving k number upon transduction at different MOIs is expected to follow a Poisson distribution P(k) = $e^{-m} m^k$ / k!, where m is the MOI and k is the number of integration events[715,716].

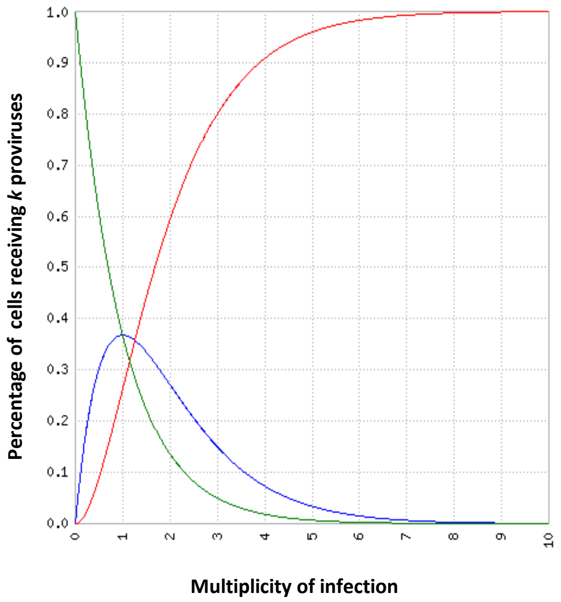


Figure 3.11. Poisson distribution describing the expected effect of MOI on the proportion of cells receiving k proviruses.

The green line represents the percentage of cells with no lentiviral proviruses (k=0), the blue and red line represent single (k=1) and multiple (k>1) integration events, respectively.

The highest proportion of cells containing only one copy of the barcode is achieved at a MOI of 1. A lower MOI could reduce the number of cells labelled with >1 barcode but would also reduce the number of uniquely labelled cells (Figure 3.11). In lineage tracking experiments, this would represent a critical issue since additional clonal populations could be found if cells have multiple tags (>1 barcode and lentiviral vector genome per cell) and is the reason why some protocols recommend using MOIs of <0.3 so that >90% of the transduced cells only contain one barcode[717–719]. In the event of measuring transgene expression by flow cytometry (not via barcoded RNA quantification), the presence of a majority of cells with a single provirus is also critical. However, in the barcode approach (expression measurement relying on barcoded RNA quantification) cells harbouring more than a single provirus do not pose a limitation for the analysis. As the sample space (the total number of possible barcode variants) is large, the chances of a cell harbouring two proviral genome copies with the same barcode variant are very low. Therefore, in this study, multiple integration of barcodes into the same cell is not a critical problem.

Suspension adapted (SA), serum-free HEK 293 cells from GSK vaccines in Rixensart (Belgium), from now on named HEK 293 SA RIX, were initially used as candidates for lentiviral vector packaging. Serum free cultures present a lower risk of adventitious viruses or prions and also for good manufacturing practices (GMP). Cell banks are available for this cell line. The barcode library was delivered to 120x10^6 HEK 293 RIX cells by transduction at a MOI low enough (MOI of 0.5) so that most cells receive only a single barcode in order to screen the genome for high transcribing sites. Cells were kept in culture for three weeks before cell sorting in order to avoid any silencing of the gene expression[720]. The number of cells was estimated so that 5% windows containing 50,000-100,000 cells. Each could be sorted depending on the percentage of GFP+ when transduced with a MOI of 1 and also the reduced viability (60-70%) of HEK 293 SA RIX cells when cultured in upright flasks (due to the lack of shaking platforms for suspension cultures).

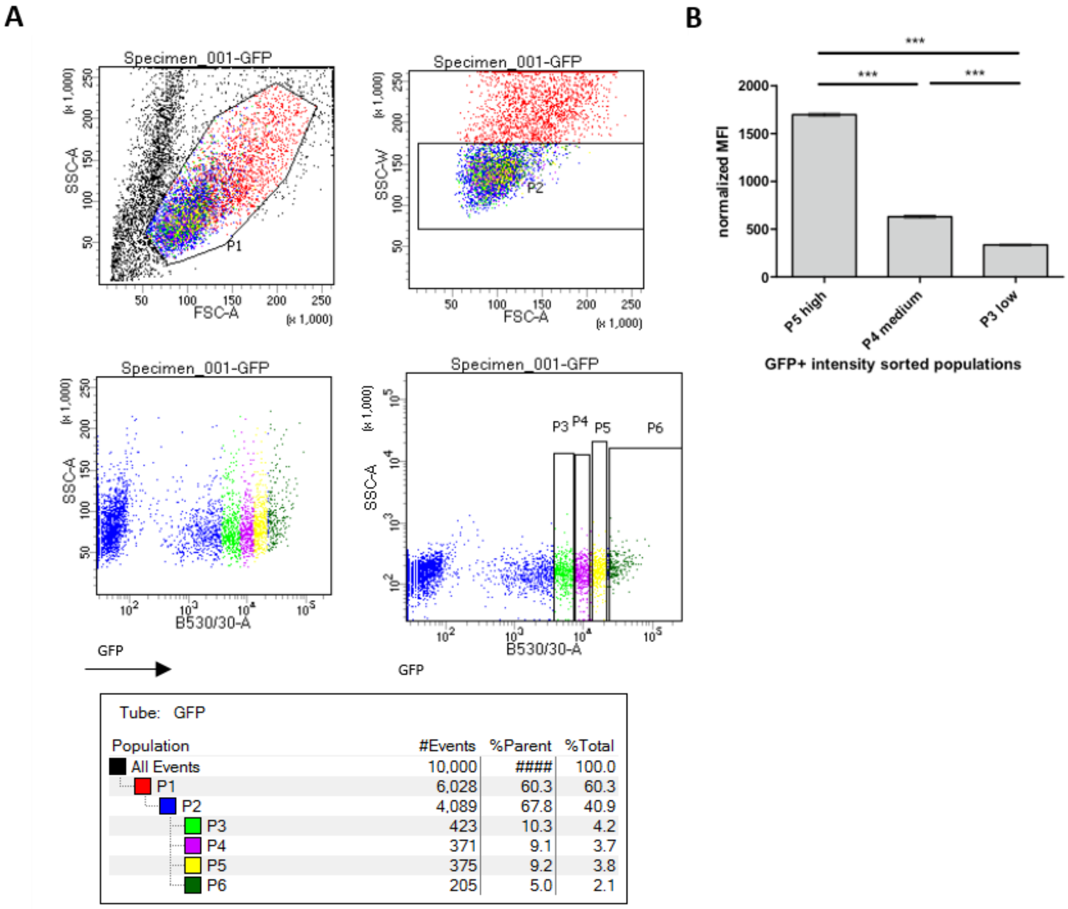
Low growth rates and relatively low titers obtained in transient transfection experiments done by others at GSK (data not shown) showed HEK 293 SA RIX cell lines are not an optimal host for vector production. Instead, HEK293 6E cell lines originally from the National Research Council of Canada were used due to their increased recombinant protein production[374,721]. HEK 293 6E express a truncated Epstein Barr virus nuclear antigen 1 (EBNA), which increase recombinant protein production in the presence of plasmids carrying Epstein Barr Virus (EBV) oriP sequences. Transient production of scFV-FC antibodies was found to be 10-fold higher than in HEK 293Ts[722]. In addition, this cell line is suspension adapted, serum-free can be combined with a family of pTT expression vectors for stable or transient expression and possess cumate and coumerycin for expression switches during production[376].

In a second experiment, 10^3, 10^4, 10^5 HEK 293 6E cells were transduced at a MOI of 1 and HEK293 6E cells were harvested (for integration site analysis and RNA-Seq) after shorter culture (7 days) to prevent fast-growing clones taking over. In terms of timing, wild type HIV particles are released as early as 18 hours post-infection in T cells, integration takes place 8.5 hours and all transcriptional species are expressed after 15 hours post-infection indicating that this selection method could theoretically be applied 24 hours post-transduction[719].

3.5.2 Cell sorting of HEK293 cell lines transduced with the barcoded vector library

Flow cytometric analysis and fluorescence-activated cell sorting (FACS) was carried out on HEK 293 SA RIX to provide separation of cellular populations based on fluorescent labelling (Figure 3.12). In this study, 4 HEK 293 SA RIX subpopulations with significantly different GFP intensities: top 5% (H), 5% mid (M), low 5% GFP expressers, as well as GFP (+ve) and GFP (-ve) were sorted, expanded and harvested for further analysis of viral integration sites.

The viability of the H sorted subpopulation dramatically dropped within 7 days after the cell sorting (Figure 3.13B). Repeated rounds of sorting changing the flow rate, along with different sorting solutions (PBS vs Hank's balanced solution) and recovery media did not solve viability problems for the highest expressers. This could be potentially attributed to the toxic effects of an accumulation of GFP722. The sorting strategy was modified to address this problem. Cells were segregated into 5 different subpopulations based on GFP intensities: top5% (H6) / high-medium 10% (HM5), mid-high 10% (MH4), mid 10% (M3 or M), low 10% (L) GFP expressers, as well as GFP+ve and GFP(-ve) (Figure 3.12). 156,900, 273,251, 267,992, 270,992 cells were recovered for HM5, MH4, M3 and L groups, respectively. All HEK 293 subpopulations were successfully expanded after cell sorting except from the top 5% (H6) group (Figure 3.13B). Several repeats of this experiments with similar outcome confirmed this phenomenon, which suggests the loss of viability is due to a cytotoxic effect caused by excessive accumulated GFP.



Tube: GFP

Population	#Events	%Parent	%Total
■ All Events	10,000	####	100.0
🟥 P1	6,028	60.3	60.3
🟦 P2	4,089	67.8	40.9
🟩 P3	423	10.3	4.2
🟪 P4	371	9.1	3.7
🟨 P5	375	9.2	3.8
🟩 P6	205	5.0	2.1

Figure 3.12. Cell sorting of HEK 293 SA RIX populations based on GFP intensity.

(A) Fluorescence-activated cell sorting (FACS) diagram corresponding to separation of HEK 293 SA RIX transduced with a barcoded lentiviral vector library. P6 corresponds to high GFP producers, P5 is mid-high GFP producers, P4 is mid GFP producers and P3 is low GFP producers. GFP (+ve) and GFP (-ve) cells, were also isolated as a control. P2 gate excludes duplets of cells and P1 corresponds to the initial population of live cells. (B) Mean fluorescence intensity values normalized by MFI of untransduced cells. All results presented (means ± SD; *** $p<0.0001$, compared between groups using one-way ANOVA and the Post hoc Tukey's test) correspond to 3 technical replicates.

Interestingly, the duplication times of the different sorted cell populations increased with GFP intensity (Figure 3.13A) possibly due to the metabolic burden on HEK 293 imposed by GFP production.

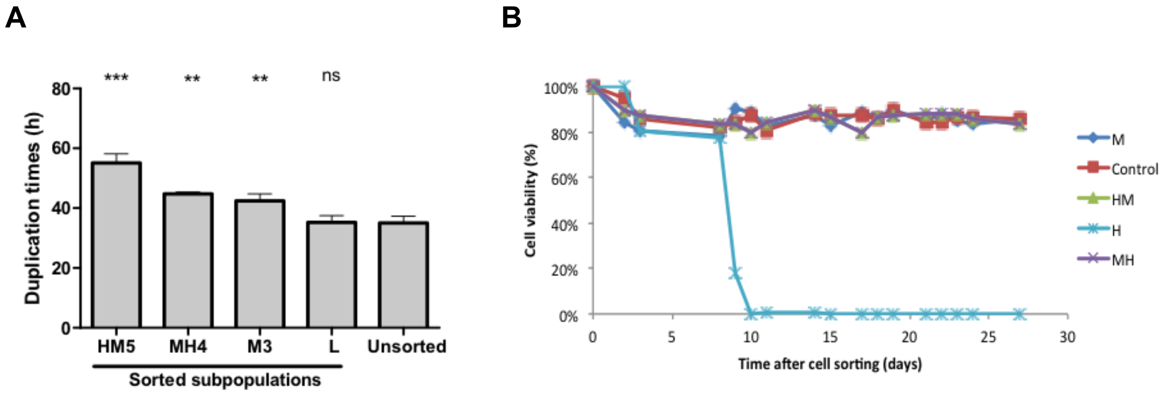


Figure 3.13. Cell line viability and duplication times of 293 SA RIX subpopulations after cell sorting.

(A) Duplication times of different HEK293 RIX sorted subpopulations. HM5, MH4, M3 and L correspond to HEK 293 SA RIX cells sorted into GFP high-medium, medium-high, medium, low. All results presented (means ± SD; * p<0.05, ** p<0.001, *** p<0.0001, ns non-significant, compared to unsorted negative control, one-way ANOVA and the Post hoc Dunnett's test) correspond to 3 technical replicates. (B) Viability percentages in different HEK293 RIX sorted subpopulations.

Duplication times are significantly lower in HEK293 SA RIX HM5 (and in HEK293 SA RIX MH4 and M3) in respect to other sorted subpopulations. No significant differences were observed between HEK293 SA RIX L and HEK293 RIX cells. This trend reinforces the hypothesis that an excessive content of GFP is deleterious for cells.

3.5.3 Lentiviral library vector copy number on transduced host cell lines

In order to normalise the expression per integrated vector copy derived from a particular integration site to a particular barcode, integration needs to be coupled with vector copy number determination by qPCR. A week after transduction genomic DNA was harvested and vector copy number was analysed by qPCR to compare the number of lentiviral genomes per cell (measured with primers and probes annealing to the WPRE sequence) to those of a housekeeping gene (endogenous beta actin) (Figure 3.14).

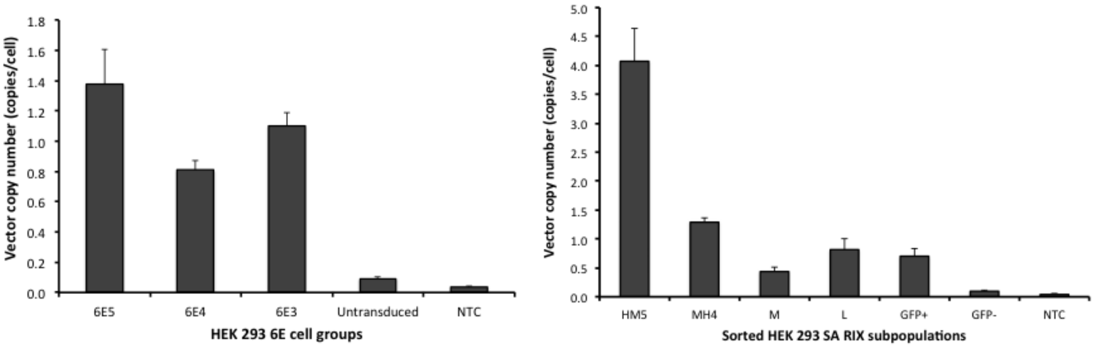


Figure 3.14. Barcode lentiviral vector copy number on different cell populations.

Samples labelled 6E3, 4, 5 stand for 10^3, 10^4, 10^5 HEK 293 6E cells, respectively transduced at an MOI~1. NTC, non-template control 1. RIX HM5, MH4, M, L, +, - correspond to HEK 293 SA RIX cells sorted into GFP high-medium, medium-high, medium, low, GFP+ and GFP- expressers. Results expressed as means ± SD.

These results indicate that cells possess a single copy of lentiviral vector integrated in their genome. Interestingly, sorting for high GFP levels may enrich for cells containing multiple proviruses. GFP is a reporter gene that has been described to quantitatively correlate its intensity with the MOI[723,724], and thus indirectly with the number of integrated vector genomes (obeying a Poisson distribution, Figure 3.11). The fact of having observed this phenomenon only in the HEK293 SA RIX HM5 population might have been caused by the concentration of the fraction of cells containing an average of 4 copies of vector, which according to the Poisson distribution is 1.5%. However, this finding does not suppose a limitation for the purpose of this project because (i) it is only observed in this particular subpopulation (ii) the probability of this 4 barcode copies in the same cell to have the same variant is negligible as it is the probability of picking the same barcode variant or one differing by only 1 nucleotide (potentially attributed to sequencing errors) based on calculations done in libraries with smaller library sizes by Bystrykh *et al*., in 2014[725].

3.6 Summary of results and concluding remarks

- A lentiviral vector was successfully engineered to harbour a DNA variable sequence tag (barcode) within a region that enables its transcription during transgene expression in target cells and allows for its retrieval during integration site analysis.

- A barcoded plasmid (4x10^5 clones) and vector library (pSYNT/vSYNT) was constructed after optimizing the amounts of backbone, stoichiometry and ligation conditions.
- Validation of both libraries by NGS revealed a sufficient size and complexity to screen 10^5 integration sites.
- Functional titers were not affected by the presence of a foreign 70bp DNA sequence within the U3 region in the 3' long terminal repeat.
- Duplication times of cell sorted subpopulations correlate with their GFP intensity.
- Vector copy number analysis by qPCR confirmed the presence of one copy in MOI-1 transductions (HEK293 6E) and all the sorted cell pools (HEK293 RIX), except for the highest producers (4 copies).

In this chapter, barcoded vector libraries were prepared and validated to efficiently deliver unique tags into thousands of genomic positions in host cell lines. One of the main challenges faced in this chapter was to achieve sufficient number of clones to guarantee a large initial sample space and minimise library diversity and size bottlenecks in subsequent stages of the project. Unlike previous reports, concatemerisation of annealed barcodes was not observed as revealed by gel electrophoresis and next-generation sequencing. Worthington *et al.*, hypothesised that an excess of 5'unphosphorylated oligonucleotides would prevent one vector end finding another due to a mass competition effect[726]. Strategies preventing mismatching did not report an increase in the transformation efficiency either.

However, in order to improve ligation efficiency, a possible improvement would be to use a cohesive end restriction enzyme with 4 overlapping nucleotides rather than the 2bp of *NdeI*[727].

Commercial oligonucleotide synthesis uses equimolar ratios of the four nucleotides to create primers with degenerate positions. However, biases in the stock mixes could affect the nucleotide representation in these positions. In order to minimize this source of bias, some companies (i.e. TriLink Biotechnologies) have done extensive research to develop a 1:1:1:1 ratio randomers. Although in this study the 4 nucleotides were quite evenly represented in the variable position, implementing an exact 25% of each-compositon would prevent potential biases from being magnified either at the bacterial subculture, viral preparation stage or during the detection step by PCR amplification.

The GC content range of the barcode design proposed in this study (including the alternated fixed position) was found to be skewed towards a low GC content. In the theoretical case of all the variable positions being occupied by Gs and Cs, the global GC content would be 61.8%, a suitable upper limit. In the most AT rich scenario, the global GC content drops to 20.6%. Although the GC content did not represent a major concern in this study, a range between 40-60% range would have been the advisable. A balanced GC range avoids the formation of secondary structures that hamper denaturation and annealing of oligonucleotides[728]. In the current barcode design, a more balanced GC content could be achieved some of the fixed nucleotides within the barcode sequence or in the surrounding nucleotides by switching from W (A or T) to S (G or C).

In line with the previous proposed optimisation, the design of the barcode could be modified to increase the current GC content of the current fixed triplet AAA (flanked by NN). Variants in which the current AAA triplet is flanked by AAs (A_7) may pose a challenge for some sequencing technologies sensitive with homopolymer-length sequencing error (i.e 454 Ion Torrent)[729].

The number of clones obtained in the library described in this study is comparable (and in most cases higher) to those reported in the literature for analysis of clonal dynamics (Table 3.2).

The position of the barcode within the lentiviral vector backbone is also relevant to the retrieval strategy and can be varied according to application. While in some publications[731,734] the barcode is located between the WPRE and the 3'LTR, CellTracker® technology locates it before the cPPT right in the middle of the lentiviral vector backbone. This location allows tracking dynamics of starter founder but does not facilitate integration site recovery. The library presented in this study not only allows cell labelling (if delivered at a low MOI) but also enables recovery of host-vector junctions.

Table 3.2. List of lentiviral and gamma-retroviral barcoded libraries recently published in the literature.

Article	Library name	Barcode pattern	Size (# of variants)	Quality control	Application
Cheung *et al.,* 2013[730]	MNDV-PGK-GFP	N_2ATC N_2 GAT N_2 AAA N_2 GGT N_2 AAC N_2	~2x10^5	Perfect match of viral vector sequences (length?) and barcode allowing up to 3 mismatches and q=20	Analysis of clonal dynamics in HSCs
Grosselin *et al.,* 2013[731]	pLentilox3.4	(N_8-(SW_5))$_5$-N_8	50	Not specified	Analysis of clonal dynamics in HSCs
Verovskaya *et al.,* 2013[732]	pMIEV and pGIPZ	GTACAAGTAAGG N_3 AC N_3GTN_3CG N_3TA N_3CA N_3TGN_3 GACGGCCAGTGAC	800 and 450	Removal of low quality sequences and sequences only occurring once	Analysis of clonal dynamics in HSCs
Hoffman *et al.,* 2007[733]	-	N_4 (barcoded primers)	7	Perfect match to the barcode and primer regions and <1 N	Identify HIV drug resistance mutations
Cornils *et al.,* 2014 [733]	LeGO	N_2ATCN_2GAT N_2AAA N_2 GGT N_2AAC N_2	>7 x10^5	Only sequences with a frequency >10 and a perfect match in the 22 variable positions were included in the analysis	Analysis of clonal dynamics in cancer
Lu *et al.,* 2012[734]	No specific name	N_6 (library ID)- N_{27}	>8 x10^4	Removal of reads with mismatches and indels up to 2bp in the 27nt random stretch. No mismatches tolerated in the 6bp ID library. Barcodes whose copy number is below a background noise threshold (algorithm) are excluded	Analysis of clonal dynamics in HSCs
Cellecta, Inc[674]	CellTracker ®	(WS)$_{15}$	5x10^7	Not specified	Multiple
Schepers *et al*., 2008[657]	pLentiLox3.4-GFP2 and pMX-GFP-bc	(N_8-(SW_5))$_5$-N_8	~5x10^3 and ~3.3x10^3	Barcodes present above background were selected based on the probability that a signal differed from an artificial background distribution	T cell lineage analysis
Brugman *et al*., 2015[735]	pTGZ	GG N_3AC N_3 GT N_3 CG N_3TA N_3 CA N_3 TG N_3 GA	485	Bioperl script filtering barcodes with desired barcode regular expression including surrounding sequences (length?). Application of clustering analysis to reduce sequencing errors (dissimilarity <2)	T cell lineage analysis
Nolan-Stveaux *et al.,* 2011[717]	CellTracker ®	(WS)$_{15}$	~27.5 x10^3	Minimum Hamming distance between barcodes in the set is 4, so up to 3 mutations in an 18-nucleotide sequence can be detected	Analysis of clonal dynamics in cancer
Porter[737] *et al.,* 2014	No specific name	Not specified	>12,000	Removal reads with >3 mismatches or >3Ns. Doping experiment to determine lower detectable limit.	Analysis of clonal dynamics
Gerrits *et al.,* 2010[659]	HC (retroviral)	N_2ATCN_2GAT N_2AAA N_2 GGT N_2AAC N_2	800 and 450	Not specified	Analysis of clonal dynamics in HSCs
This study	pSYNT	N_3 ATC NS GAT N_2 AAA N_2 GGT NW AAC N_2 TGA N_3	~4x10^5	Exact match for the barcode and a region comprising 20nt flanking the barcode. Application of clustering analysis to reduce sequencing errors (dissimilarity < 11)	Lentiviral packaging cell line development

Another useful improvement to the current design could be to clone barcode tags flanked by sequences complementary to Illumina next-generation sequencing primers into the transfer vector. This strategy, already applied by Porter *et al*., in 2014[736], would have helped to reduce the number of amplification steps necessary to prepare libraries for sequencing, thus reducing potential handling errors, PCR bias and mutations introduced by the DNA polymerase.

Grosselin *et al.,* remarks the importance of an 'arrayed' lentiviral barcoded library in order to overcome bias introduced by restriction enzymes, PCR and random ligand attachment[731]. An 'arrayed' lentiviral barcoded library involves a known library size and complexity, which helps to interpret the linkage between a particular barcode variant and the target cells. However, in order to control the exact number of variants in a library, each of them should be individually annealed and equitably pooled, which can limit the throughput of the library.

An alternative to that would be to add 'spike-in' reference controls to normalize or calibrate for the aforementioned bias, in other words, known numbers of cells (for instance 50, 500 and 1000 cells) with a known single barcode variant (with sufficient Hamming distance). This would also help evaluate loss in sequence complexity as a result of sequencing errors.

An MOI between 0.5-1 was chosen because we believed this to yield the relatively highest proportion of single transduced cells following Poisson distribution. The single integration of a lentiviral vector genome in the host cell line genome is more relevant for the FACS-LM PCR approach, since these cells will not be distinguished from those containing more than one integration event. Nevertheless, the barcode system enables the RNA-Seq approach to determine the relative abundance of integrated barcode, regardless of the number of integrated viral genomes.

FACS sorting would pool cells according to fluorescence intensity regardless of the number of integrated viral genomes the cell harbours. Ideally, site-specific expression to quantify the relative abundance of barcodes would be performed by RNA-Seq. This way, even though high fluorescent intensity pooled cells contain more than one vector copy, sequencing of barcode-containing transcripts would allow an individual assessment of their expression profile. Nevertheless, cell sorting of polyclonal populations based on the GFP intensity upon transduction with the lentiviral library was performed as an alternative to RNA-Seq due to the following reasons: (i) to back up the measurement of gene expression in case the barcode system did not work; (ii) to validate the barcode system if it works (correlate number of barcode counts with average GFP intensity in the sorted populations) and (iii) to screen for sites within high GFP expressers with high number of RNA counts, which will confirm translatability of integrated cassettes (not confirmed with the RNA-Seq approach).

In this study, high levels of GFP (associated with multiple copies of the provirus) were shown to compromise the viability of host cells. However, other factors may also contribute to cytotoxicity and a number of alternative strategies are available. Relatively low viability rates (60%) were observed in static cultures in comparison to agitated suspension cultures (>95%) (data not shown). Non-agitated cells tend to clump together and grow forming patches or aggregates, which limits oxygen supply to central cells[737,738]. These cell aggregates that occur due to environmental stresses can accelerate the rate of cell death within the sample, resulting in the release of "sticky" DNA molecules from the dying cells that can facilitate further clumping of neighbouring cells[739]. This phenomenon leads to high content of cell debris (with necrotic factors) that could induce further cell death[740] and problems during cell sorting. Adding endonuclease deoxyribonuclease I (DNase I) into the sample can minimize the presence of free-floating DNA fragments and cell clumps[741]

The addition of EDTA is also recommended because it acts as a Ca^{+2} ions chelator, preventing calcium protection of intracellular domains of adhesion molecules against proteolytic activities[742,743]. Although FCS can increase cell viability, it is important to note that these cell lines ideally would not be supplemented with FCS for safety issues (i.e. prions, adventitious bovine viruses) and manufacturing reasons such as (scalability, batch to batch variation and supply chain limitations). Serum-free media contains fewer undefined components, offers better lot-to-lot consistency and facilitates subsequent purification processes.

Additional bias can be introduced by splitting cells after transduction (or during the sorting), which skews cell populations containing different barcode variants. Experiments to evaluate the extent of this sampling bias must be designed to ensure adequate barcode representation.

Finally, there are currently no guidelines or standard pipelines for barcode analysis. The criteria ranges from various minimal read qualities combined with the exclusion of reads with lower frequencies (specific threshold for each case). Filtering protocols also discard reads not matching the expected barcode pattern. All together, around 10% of the reads are usually excluded from further analysis. However, this should depend on the error rate of the sequencing platform used. Illumina-based sequencing has a 1-5% error rate mostly caused by substitution[744,745]. In contrast, PacBio (third generation or single-molecule sequencing) reports up to 15% error rate mostly due to insertions and deletions[746]. Reference libraries provide a characterised control for these parameters and help distinguish sequencing errors from less frequent real variants. In the absence of reference libraries, astringent criteria can be applied in detriment of low frequent bona-fide barcodes. Porter *et al.*, performed a doping experiment to determine the lower detection limit for barcode representation (set to 0.0002% of the reads in their study)[736].

In conclusion, we generated, characterised and delivered a lentiviral barcoded library with enough complexity to support high throughout genomic site-specific expression. The next chapter describes the application of the barcode library to find genomic loci that support high transgene expression.

Chapter 4

RESULTS: Integration site analysis and correlation with barcode expression

4.1 Introduction

Typical strategies for the selection of high producing clones involve their individual segregation and high-throughput screening of their performance (reviewed in Section 1.22 High analytical burden of screening clones). Selection methods such as limiting dilution cloning are simple and inexpensive but often laborious and time consuming. FACS-based approaches benefit from high-throughput capacity but, on occasion, fluorescence results in toxicity to the cell[747], which limits the ability to select high producers. Secretion-based assays require handling expertise and are limited by the fragility of the cells and the costs of detection antibodies. These disadvantages are overcome by closed, automated sophisticated systems although the associated costs are considerably higher.

In this study, we propose the use of lentiviral vector integration preferences as a guide to target the insertion of therapeutic cargo genes in genomic positions with relatively high expression. Once the barcoded vector library (previously introduced in Section 3.3) is generated, most cell types (due to wide tropism of VSV-G[748]) can be transduced with the vector particles. The system relies on the evolutive ability of lentiviruses to pick stable sites with light burden on cell fitness. Highly laborious screening of individual clone titers (by qPCR, ELISA) is substituted with molecular tagging of RNA molecules and parallel screening by next-generation sequencing. However, this presents a new challenge in terms of bioinformatics. While initial concerns in Chapter 3 regarded size and complexity of barcoded libraries, in this chapter retrieval and discrimination of 'real' barcodes at a genomic and transcriptomic level will be assessed. Although multiple online tools to retrieve viral integration sites exist (QuickMap[749], Mavric[750], VISA[696], VISPA[751]), none of them facilitates the retrieval of a barcoded sequence in the vector LTR. Similarly, recovery of a reduced portion of barcoded traces from whole RNA typically fragmented in relatively short length libraries is a challenge this approach faces. In this chapter, these questions will be investigated to develop and prove an alternative approach for selection of high producing clones. Despite integration properties (efficiency, payload and copy number) would be those of the method chosen for targeted-insertion (Chapter 5), as reviewed in Chapter 3, the use of lentiviral barcoded libraries for cell line development constitutes a unique approach for packaging cell line development due to its high-resolution genomic screening and its quantitative site-specific expression analysis.

4.2 Aims

The specific aims of this chapter are:

- To show evidence of the retrieval of lentiviral barcode library-host cell line junctions containing the barcode.

- To generate a bioinformatics integration site analysis pipeline to analyse barcoded junctions at a high through put scale.

- To quantify the abundance of RNA encoding each barcode variant delivered by the lentiviral library.

- To assess the complexity of the barcoded library at different stages of the process (integrated provirus, expression levels via barcode counts).

- To compare expression derived from lentiviral integration in particular loci with basal gene expression levels of host cell lines.

- To correlate location and relative abundance using the barcode in order to select biologically relevant candidate position which can stably harbour lentiviral vector packaging components.

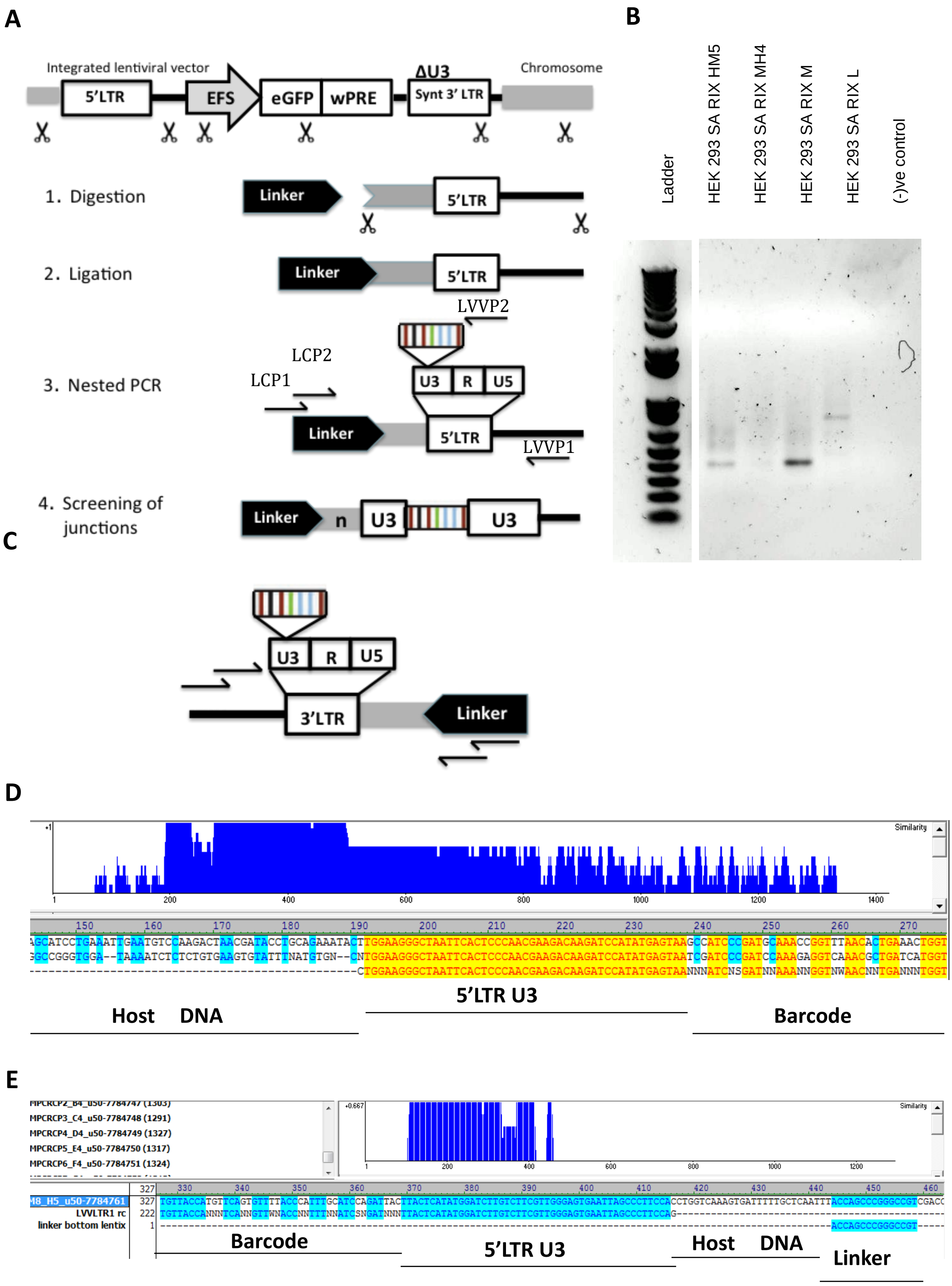
4.3 Integration site analysis and RNA expression on vSYNT-transduced HEK 293 SA RIX

4.3.1 Retrieval of lentiviral barcoded vector library – host chromosome junctions in HEK 293 SA RIX cells by LM-PCR

Typically, in cell line development, the expression of integrated transgenes is screened based on the mean fluorescence intensity of a fluorescent protein in the clone or by antibody-derived methods as described in Section 1.2.2. In this study, 4 subpopulations were isolated by cell sorting based on GFP intensity and the genomic position of integration analysed by ligation-mediated PCR (LM-PCR) to identify sites associated with high transgene expression. Alternatively, the barcoded approach suggested in this study investigates the use of barcodes present in viral vector transcripts as a quantitative method to link clonal expression to a particular integration site. Although, it is not required, the barcode approach is performed in populations sorted by FACS in order to compare their outcome and complement the former in case it is not functional.

In order to recover vector integration sites, genomic DNA from host cell lines with a low vector copy number per cell was harvested and analysed using ligation-mediated PCR. Following from Section 3.5, 1.2x10^8 HEK293 SA RIX cells were transduced with the vSYNT11 library at an MOI of 0.5 to deliver the barcode. As described in Section 3.5, cells were kept in culture for three weeks to obtain expression only from integrated provirus and sorted into 4 subpopulations displaying differential GFP intensities (high-medium 10% (HM5), mid-high 10% (MH4), 10% mid (M3 or M), low 10% GFP expressers) as well as GFP+ve and GFP(-ve) (Figure 3.12). Genomic DNA was extracted from subpopulations HM5, MH4, M3 and L and digested with *DraI*. A linker cassette with compatible ends was ligated in order to provide a known region to specifically amplify junctions by nested PCR (Figure 4.1A). In contrast to the standard LM-PCR approach, which uses 3'LTR-specific primers, this version of LM-PCR is performed on the 5' junction and allows amplification of the barcode within the junction (Figure 4.1C). This modification is designed to enable longer (and better quality) reads into host chromosome given that the deltaU3 (where the barcode is located) is closer to the end of the provirus so reverse primers can be positioned immediately downstream of the barcode. In addition, primers only binding regions proximal to the 5'LTR enable specific amplification of upstream junctions and prevent the formation of by-products (often used as control bands). A faint DNA smear of chromosome-vector junctions was observed in all HEK 293 SA RIX samples (Figure 4.1B). Amplified junctions were ligation-independently cloned into pCR4-TOPO TA vector backbone. 8 colonies per group were Sanger-sequenced and the presence of LTR-chromosome junctions was confirmed prior to NGS (Figure 4.1D and E)

Retrieved sequences were manually mapped against the hg19/GRCh37 (UCSC/NCBI) version of the human genome using BLAT. In all groups, 1-3 out of 8 colonies picked per group contained an integration site that could be mapped (data not shown), making the retrieval efficiency consistent across all groups.



```
 150       160       170       180       190       200       210       220       230       240       250       260       270
AGCATCCTGAAATTGAATGTCCAAGACTAACGATACCTGCAGAAATACTTGGAAGGGCTAATTCACTCCCAACGAAGACAAGATCCATATGAGTAAGCCATCCCGATGCAAACCGGTTTAACACTGAAACTGGT
GGCCGGGTGGA--TAAAATCTCTCTGTGAAGTGTATTTNATGTGN--CNTGGAAGGGCTAATTCACTCCCAACGAAGACAAGATCCATATGAGTAATCGATCCCGATCCAAAGAGGTCAAACGCTGATCATGGT
-----------------------------------------------CTGGAAGGGCTAATTCACTCCCAACGAAGACAAGATCCATATGAGTAANNNATCNSGATNNAAANNGGTNWAACNNTGANNNTGGT
```

```
MPCRCP2_B4_u50-7784747 (1303)
MPCRCP3_C4_u50-7784748 (1291)
MPCRCP4_D4_u50-7784749 (1327)
MPCRCP5_E4_u50-7784750 (1317)
MPCRCP6_F4_u50-7784751 (1324)

                      327  330       340       350       360       370       380       390       400       410       420       430       440       450       460
M8_H5_u50-7784761     327  TGTTACCATGTTCAGTGTTTTACCCATTTGCATCCAGATTACTTACTCATATGGATCTTGTCTTCGTTGGGAGTGAATTAGCCCTTCCACCTGGTCAAAGTGATTTTTGCTCAATTTACCAGCCCGGGCCGTCGACC
LVVLTR1 rc           222  TGTTACCANNNTCANNGTTWNACCNNTTTNNATCSNGATNNNTTACTCATATGGATCTTGTCTTCGTTGGGAGTGAATTAGCCCTTCCAG-----------------------------------------------
linker bottom lentix    1  -------------------------------------------------------------------------------------------------------------ACCAGCCCGGGCCGT-----
```

Figure 4.1. Retrieval of integration junctions between barcoded vector library and HEK 293 host chromosomes by ligation-mediated PCR.

(A) Schematics of the LM-PCR performed in this study. Primer names and sequences are detailed in Section 2.2.29, Materials and Methods. (B) 2% agarose gel electrophoresis of vector library-host cell chromosome obtained by LM-PCR. 1kb plus DNA ladder (Appendix B). (C) Schematics of the LM-PCR performed on a hypothetical 3'LTR-host DNA junction. (D) Alignment of traces obtained by Sanger sequencing showing 5'LTR barcoded U3 on the right and chromosome junction on the left. (E) Traces recovered in reverse orientation contain the barcode on the left side of the alignment and the linker on the right.

4.3.2 Next-generation sequencing of integration sites in HEK 293 SA RIX

Once LM-PCR retrieval of integration sites was confirmed by shotgun cloning and Sanger sequencing (Section 4.3.1), LM-PCR products generated from sorted HEK293 SA RIX subpopulations with vSYNT11 (RRL EEW barcoded library) were purified and submitted to Genewiz and UCL genomics for high-throughput sequencing with Illumina MiSeq using a 300bp paired-end (PE) strategy. The aim of this choice was to extend the retrieval of long junctions to 600bp. Around 2 million (M) R1 and R2 reads per sample were obtained. 50% of the reads were successfully merged using Burrows-Wheeler Aligner (BWA) *pemerge* tool in HEK293 SA RIX HM5 and MH4 samples; however, rates were lower for M and L samples (37 and 16%, respectively). Further optimization including trimming of read ends, error threshold or number of overlapping base pairs did not result in higher merging rate.

Quality control of LM-PCR SA RIX merged reads showed good (green) quality score values in most of the reads (Figure 4.2E) and along the length of the read (Figure 4.2B). Merged reads have a GC content ratio distribution of around 50% as expected (Figure 4.2D) and show a peak of frequency around 330-340bp of length although read lengths up 530bp were obtained (Figure 4.2C). Capillary electrophoresis of LM-PCR products prior to indexing is comparable with the agarose electrophoresis gel (Figure 4.1B) and did not show a high degree of polyclonality or diversity in 2HM5 and 2MH4 samples (compared to 2M and 2L) (groups' nomenclature defined in Section 3.5.2), which can be deduced from the presence of just a few strong bands (Figure 4.2A). These results confirm the good quality of most sequences and discard any potential errors originated during the library preparation.

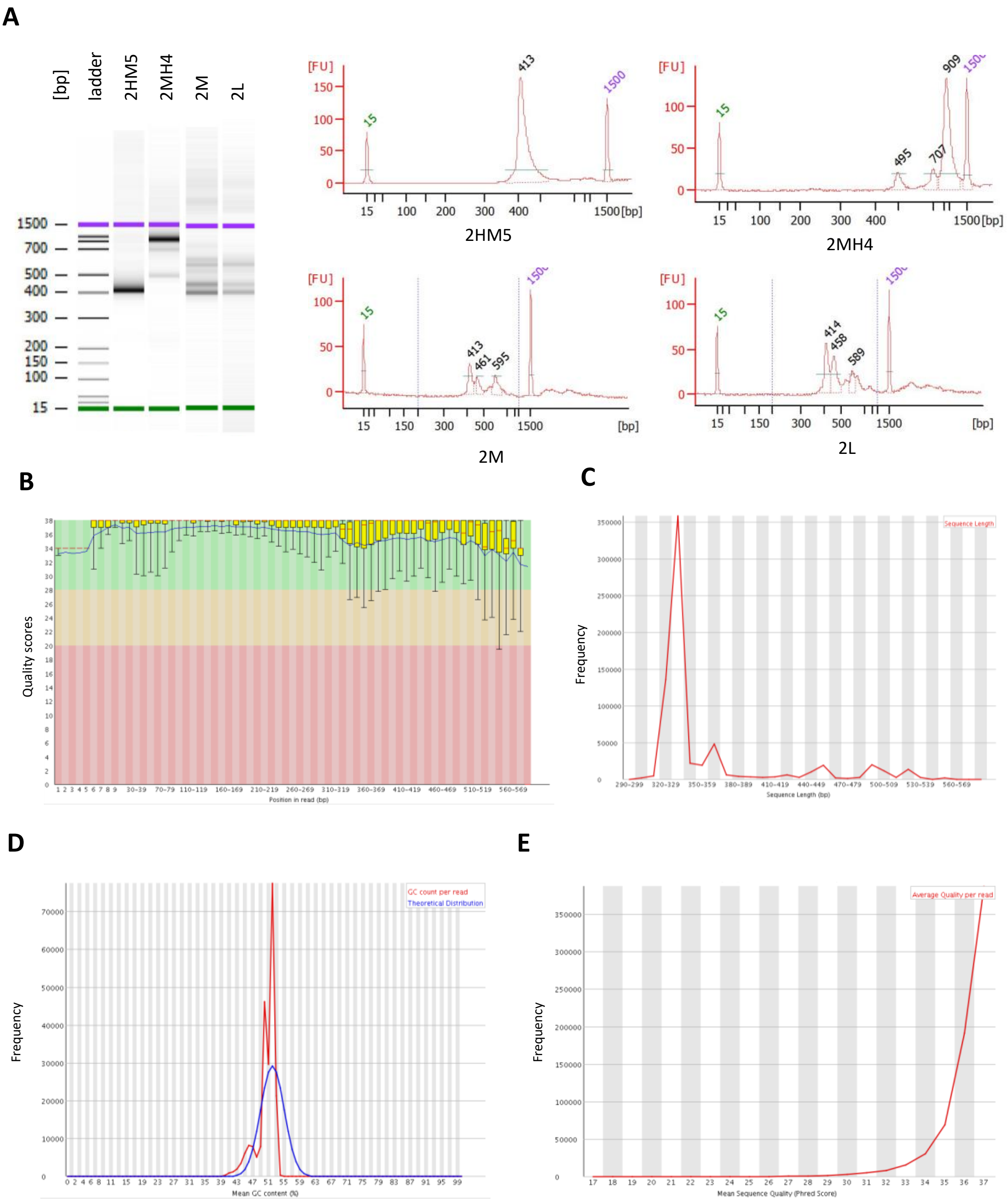


Figure 4.2. Summary of quality control statistics by FastQC for LM-PCR performed on HEK 293 SA RIX HM5 subpopulation.

(A) DNA quality assessment of vector-chromosome junctions by microchip-based capillary electrophoresis using Agilent Bioanalyzer; results given in fluorescent units (FU). Peaks at 15bp and 1,500bp are internal controls. (B) Distribution of quality scores across all 100bp of the read (Sanger/Illumina 1.9 encoding). The yellow boxes represent the inter-quartile range (25-75% of the score values per bp). The upper and lower whiskers indicate the 10th and 90th percentiles; the blue line represents the mean quality. The background of the graph divides the y-axis into very good quality calls (green), calls of intermediate quality (orange) and calls of poor quality (red). (C) Frequency of read lengths after trimming 3´ends and merging with BWA pemerge. (D) GC distribution (mean GC%) over all sequences. (E) Mean sequence quality distribution over all sequences. Similar results were obtained for LM-PCR 2MH4, LM-PCR 2M and LMPCR 2L also sequenced using MiSeq 300bp PE (data not shown).

A custom script called ‘extract_library_barcode.pl’ (AppendixB) was used to extract barcode sequences and host integration sites. The script detects the barcode and 46nt and 39nt LTR sequences flanking it and removes the linker sequence as long as all of the following criteria are satisfied:

- Alignment to linker ends at the last 5 bases (bases 32-26) on the reference linker sequence (5’-ACTATAGGGCACGCGTGGTCGACGGCCCGGGCTGGT-3’).

- The sequence identity of the linker sequence is >80%.

Additionally, host sequences <20nt in length (minimum required to unequivocally map an integration site) were filtered out. Around 160,000-200,000 extracted IS/barcodes passed the filter for HEK293 SA RIX HM5, MH4. HEK293 SA RIX M, L counts were reduced down to nearly 30,000 sequences. Integration sites were mapped against the human genome (*H. sapiens* UCSC hg19/GRCh37) using Blat. Blat was used instead of Blast because it is indexed in a different manner allowing for less usage of RAM memory and an easier mirroring than BLAST. Despite having less homology depth, Blat enables to run simultaneous queries at a higher speed with increased output options as well as different links to UCSC tools such as custom tracks and genome browsers.

A .psl file was output from Blat and was converted to bed format using a custom script (‘get_best_hit_from_psl.pl’, Appendix B) following the criteria below:

- The identity metrics set by default 1-(#mismatches/length) were replaced with the expression: 1 - (#mismatches + insertions + deletions / query + insertions + deletions). This way, insertions and deletions are included in the total number of mismatches, which contributes to accurate retrieval of sequences providing that long sequences with a high number of mismatches and gaps are not expected.

- Blat parameters were optimized, -minMatch=2 (default number of tile matches) and -minScore=20 instead of the default 30 in order to retrieve integration sites between 20 and 30 nucleotides in length (still unambiguously mappable).

-Use of –fastMap parameter; –fastMap works for alignments with high identity (>90%) and increases the speed when gaps are not expected.

- Only sequences with an exact match in the last position (the closest to the LTR) must be retrieved.

- The minimum identity threshold of host genomic sequences (-minIdentity=N) was set to 99% to enhance highly specific alignments (default is N=90%)

- Integration sites with ambiguous alignments (alignments with >1 position retrieved) were discarded (the filter command was not included in the script but performed separately).

The aforementioned criteria were established after manually mapping 100 random integration sites using Blat web tool. The results obtained were used as a reference/standard to adjust the mapping parameters and default criteria of the pipeline.

Out of the 142,749; 157,558; 21, 817 and 22,698 raw alignments with various scores output by Blat, 116,048 (81%); 127,476 (68%); 14,873 (68%) and 16,099 (71%) non-ambiguous reads containing integration sites were retrieved for 2HM5, 2MH4, 2M, 2L HEK 293 SA RIX cells, respectively (Figure 4.3). Since psl files contained repeated hits retrieved with different restriction enzymes, the different loci retrieved were counted based on the expression ‘endposition_chr’ and 310, 385, 414, 30 different IS (1,149 in total) were obtained out of the 2HM5, 2MH4, 2M, 2L HEK 293 SA RIX psl files, respectively. Providing the theoretical maximum complexity of the barcoded vector library (discussed in Section 3.4.1) and taking into account the MOI of 1 at which the >250,000 sorted were transduced (and being the NGS no limitation in screening throughput) values lower than 10^3 variants represent a substantial decrease in complexity.

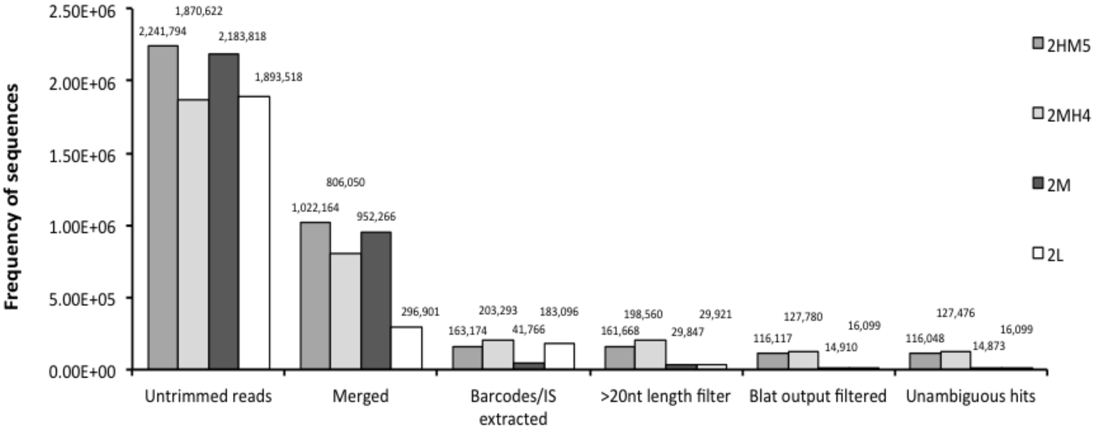


Figure 4.3. Frequency of sequences throughout the integration site analysis pipeline.

Number of reads obtained from MiSeq 300bp paired-end PE sequencing configuration, successfully merged reads with less than 5N, sequences with a complete barcode and IS, IS with a length >20 nucleotides, filtered using optimised Blat parameters and alignments with >1 genomic position retrieved. IS, integration sites.

The drop in the complexity of the library could be explained by the bias introduced at several levels: internal population dynamics during 3 weeks of cell culture (i.e. some clones overgrowing others) and the cell population drift introduced in consecutive rounds of passaging.

4.3.3 Distribution of integration sites and annotation features in HEK293 SA RIX

In order to understand the integration profile of lentiviral vectors, the 310, 385, 414, 30 different integration sites extracted from 2HM5, 2MH4, 2M, 2L HEK 293 SA RIX psl files, respectively were analysed. Unexpectedly, a major proportion (>70%) of the different integration sites recovered in all four groups were located in a few regions in the host genome within a window of a 100-200 base pairs (Table 4.1). The bias concerning reduced number of integration sites, also translated to the barcode diversity where a large part of the populations' polyclonality was depleted, limiting the throughput of the approach and the chances of finding a corresponding barcode in the RNA-Seq analysis (Figure 4.4).

Table 4.1. Clusters of integration sites retrieved from HEK 293 SA RIX cells by LM-PCR.

Subpopulation	Chromosome	Genomic position	Number of reads (%)	Gene
2HM5	**chr1**	**240504150**	38 (12.3)	FMN2
	chr12	85450800	17 (5.4)	LRRIQ1
	chr3	**152185950**	60 (19.4)	Downstream MBNL1
	chr5	**88071754**	84 (27.1)	MEF2C
	chr6	76538100	10 (3.2)	MYO6
	chr7	**4730550**	61 (19.7)	FOXK1
	chr8	71354775	17 (5.4)	Downstream NCOA2
	chrX	117405680	10 (3.2)	Upstream WDR44
2MH4	**chr1**	**240504085**	50 (13.0)	FMN2
	chr2	216556664	53 (13.8)	LINC00607
	chr3	**152185920**	107 (27.8)	Downstream MBNL1
	chr5	**88071734**	83 (21.6)	MEF2C
	chr7	**4730521**	61 (15.8)	FOXK1
2M	chr1	178838231	40 (9.7)	RALGPS2
	chr1	87493018	49 (11.8)	HS2ST
	chr12	84905305	42 (10.1)	- (in 100kb)
	chr13	73000938	34 (8.2)	- (in 100kb)
	chr18	67970581	22 (5.3)	SOCS6
	chr5	91734199	14 (3.4)	AK0568485
	chr6	72120260	33 (8.0)	LINC00472
	chr6	101103750	51 (12.3)	ASCC3
	chr8	57589988	15 (3.6)	- (in 100kb)
	chrX	24003100	61 (14.7)	KLHL15
2L	Chr7	**4730521**	11 (36.7)	FOXK1

Positions in bold were found in different subpopulations. This table shows the genomic positions retrieved with a higher proportion of the LM-PCR reads (95.7% of 2HM5, 92% of 2MH4, 87.1% 2M and 36.7% in 2L subpopulations). The remaining proportions are constituted by reads containing a more varied representation of loci.

These results could be explained due to the bias introduced during prolonged periods of culture. A small number of fast-growing clones may have formed an increasing proportion of the population. Alternatively, it may have been originated due to the accessibility to the vector-chromosome junction, dependent on the restriction site chosen in the LM-PCR step. The composition of the barcoded library retrieved by LM-PCR also showed a strong bias towards certain variants. However, the percentage of most represented variants (Figure 4.4) does not correspond to the most represented retrieved integration sites (Table 4.1), indicating that vectors with different variants integrated in the same genomic position.

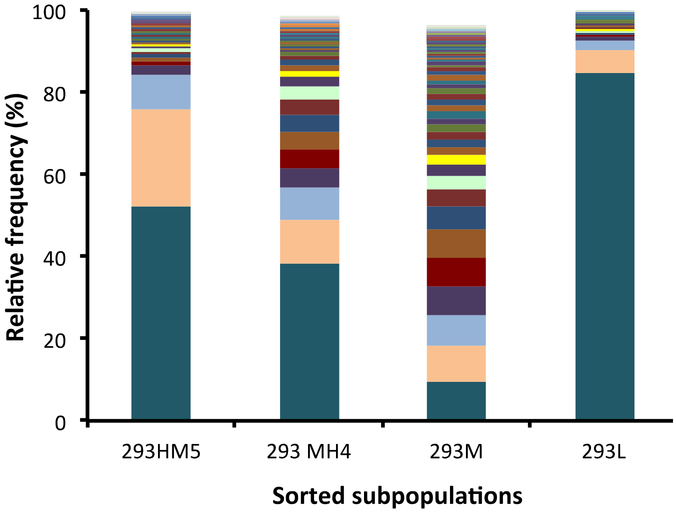


Figure 4.4. Relative abundance of barcode variants retrieved by LM-PCR on HEK 293 SA RIX cells transduced with vSYNT11.

However, since the chances of two lentiviral vector particles to integrate within a region separated by a few base pairs are very low, the clustering of multiple reads around certain positions with slightly different coordinates is likely due to the generation of chimeric PCR products. This phenomenon was first described by Saiki *et al.*, 1988 and describes an apparent recombination between different sequences with a high degree of similarity during PCR amplification[752]. The formation of chimeric PCR products is due to insufficient extension times that lead to incomplete extension of primers during the elongation phase of the PCR. These incomplete sequences cross-prime a different (but similar) molecule of template in the next cycle and complete the extension generating a strand consisting of fragments from two different parental templates. Since lentiviral junctions possess similar sequences between templates (the lentiviral LTR) and their length is variable, extension times may not be sufficient for complete amplification of certain junctions favouring the generation of chimeric sequences. As a consequence, no conclusions regarding integration preferences could be drawn from this experiment.

As shown in Table 4.1, the same integration sites were retrieved in different sorted populations. This could be explained by the variegation of expression in overgrown clones so cells containing the same IS were sorted into different subpopulations.

Alternatively to LM-PCR, high frequency unique integration sites could be retrieved by linear amplification mediated (LAM) PCR or its restriction digest-free variation non-restrictive (nr) LAM-PCR.

4.3.4 Determination of the relative abundance of barcode RNA variants by RNA-Seq on a custom sequencing library in HEK 293 SA RIX

Initially, relative abundance of barcode variants in sorted HEK 293 SA RIX HM5, MH4, M and L subpopulations was evaluated by whole transcriptome analysis. Whole RNA from the subpopulations mentioned above was submitted for sequencing using the Illumina HiSeq2500 1x (single-read) 50bp reads configuration in the Rapid Run mode. Results demonstrated quality scores >30 for more than 90% of the sequence in most reads (**Figure 4.5**C) all files along with balanced GC (**Figure 4.5**D). However, only a few tens of barcodes per sample were retrieved from fastq files containing >20M reads. Providing that RNA was extracted from 10^7 cells and that a single cell contains an average of 200,000 mRNA molecules (1-5% whole RNA)[753], a total of 10^{12} RNA molecules per RNA preparation should be expected. Under the assumption that the expression of the vector only constitutes a 0.1% of the overall protein production (to 10^9 hypothetical vector RNA molecules) and considering 20M reads per sample were obtained, the 5-log difference between the initial number of reads and the tens of barcodes eventually retrievedwhose is not difficult to explain. Lack of coverage or excessive fragmentation could explain the origin of this problem. In order to test the latter, reads containing half of the barcode pattern were extracted; 17

nucleotides following a known pattern are sufficiently specific not to retrieve unspecific sequences. Ten times more reads were retrieved in all samples suggesting 50bp fragments might be too short to integrally retrieve a sequence of 34bp (the size of the semirandom stretch of the barcode).

Alternatively, a custom library preparation was tested in order to retrieve barcoded transcripts more efficiently. In a standard library preparation for RNA-Seq (using the TruSeq RNA Sample Prep kit form Illumina), mRNA is purified, fragmented and captured with polyA magnetic beads prior to polymerisation of the first strand of cDNA. Then, the second strand is synthesised, ends are repaired, 3' ends are adenylated, adapters are ligated and indexed and P7 sequences are added by PCR prior to library validation, normalization and pooled for sequencing. The objective of the custom library preparation was to maximise barcode retrieval by (i) decreasing the RNA fragmentation time during the library preparation step and (ii) pulling down only barcoded transcripts to ease posterior bioinformatics analysis. Accordingly, fragmentation times were reduced from the standard 8 minutes to 3 minutes. Libraries significantly longer than the standard insert size (~500bp instead of ~150bp) were obtained. An additional step was included after the ligation of the adapters consisting of a selective amplification of the library by PCR using biotinylated primers specifically annealing flanking the barcode followed by an enrichment of biotinylated DNA. Custom-made libraries were sequenced using the v3 2x300bp (paired-end) MiSeq kit at UCL genomics. The ~25M reads of the flow cell were divided into 9 samples (>2M each) including HEK 293 HM5, MH4, M, L among other 5 other samples (from a different study).

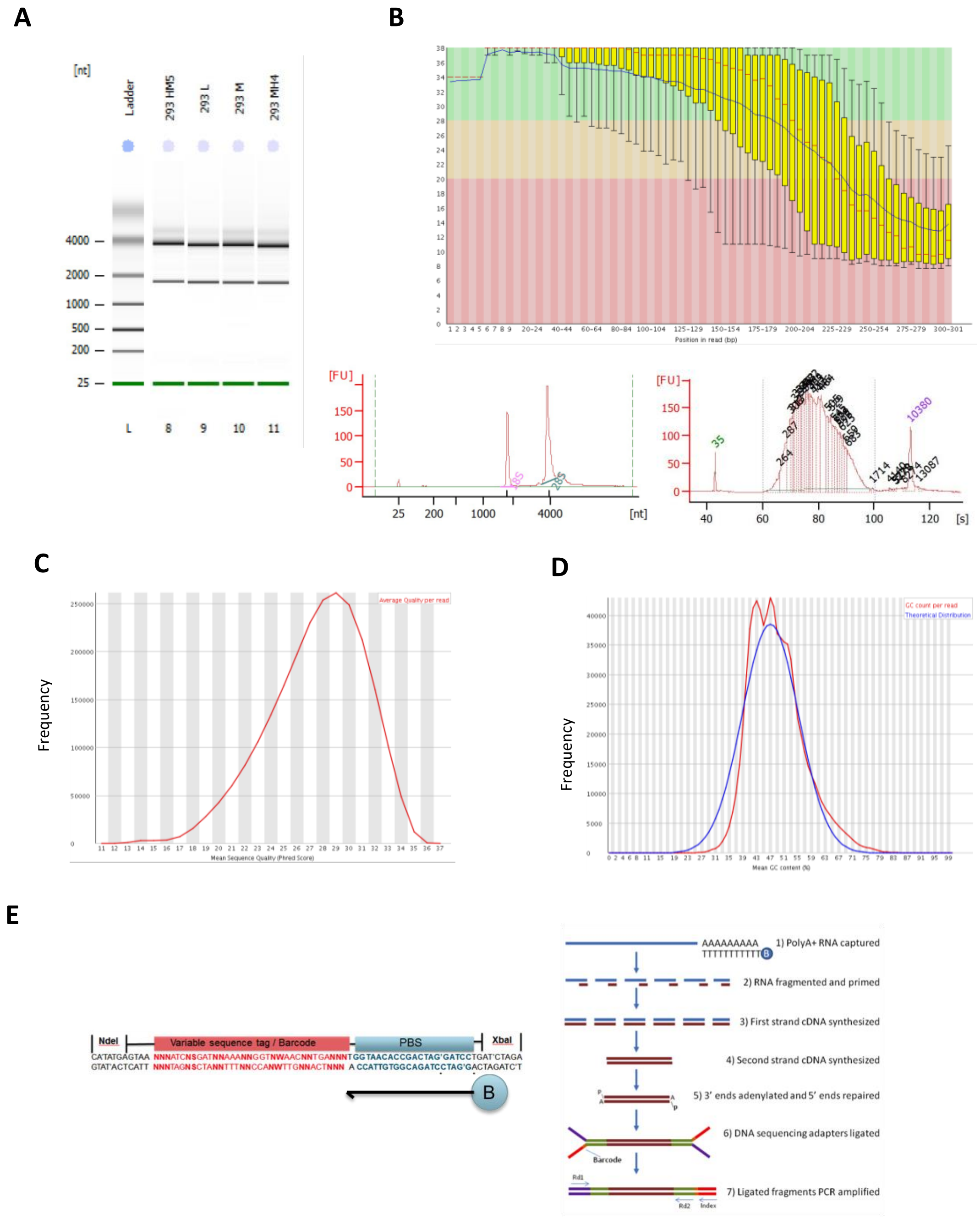


Figure 4.5. Summary of quality control statistics by FastQC from HEK SA RIX HM5 from a custom sequencing library.

(A) RNA quality assessment by microchip-based capillary electrophoresis using Agilent RNA 600 Nano; results given in fluorescent units (FU). Peaks at 25bp, 35bp and 10,380bp are internal controls. Top and bottom left diagram shows total RNA profile with peaks corresponding to 18S and 28S ribosomal RNA; bottom right diagram shows extracted barcoded RNA. (B) Distribution of quality scores across all 300bp of the read (Sanger/Illumina 1.9 encoding). The yellow boxes represent the inter-quartile range (25-75% of the score values per bp). The upper and lower whiskers indicate the 10th and 90th percentiles; the blue line represents the mean quality. The background of the graph divides the y-axis into very good quality calls (green), calls of intermediate quality (orange) and calls of poor quality (red). (C) Mean sequence quality distribution over all sequences. (D) GC distribution (mean GC%) over all sequences. Similar results were obtained for 2MH4, 2M and 2L, also sequenced using MiSeq 300bpPE (data not shown). (E) Diagram showing the use of biotinylated primers (left) as opposed to the first step of a conventional RNA-Seq library preparation workflow (right). Diagram from Corney *et al.*, [754].

Close examination of reads from both orientations revealed barcodes could successfully be extracted from R2 reads. 2-3M input R2 reads per sample were reverse-complemented with a simple script called 'rc_fastq.pl' (Appendix B) and quality trimmed to discard those with an excessive number of missing bases (N>5). >99% of sequences passed that filter in all samples. After a quality control certifying good quality scores along all base pairs of the read (Figure 4.5B) in most reads (**Figure 4.5**C) and a balanced GC content (**Figure 4.5**D), barcodes were extracted with a script called 'extract_rt-pcr_barcodes.pl', which, similarly to the script 'extract_viral_insertion_barcodes.pl', recognises 46 and 39nt LTR sequences flanking the barcode and extracts the barcode sequence.

The number of barcodes after this step dropped to 60,000-185,000 depending on the dataset (Figure 4.6A); 2L reported only 5,878 barcodes. The number of unique variants found for 2HM5, MH4, 2M and 2L were 1561; 1,578; 1,691 and 518, respectively, of which 477; 528; 605 and 107 were recurrent (Figure 4.6B). As shown in Figure 4.6C, the nucleotide composition is biased towards those variants retrieved in a major proportion. This trait is particularly strong in 2HM5, MH4, and 2L (Figure 4.6D).

Despite the reduction in the number of barcodes variants retrieved, the peak in the number of dissimilarities is maintained at 11 nucleotides as seen in plasmid and vector libraries. Interestingly, the removal of false barcode variants originated as a result of sequencing errors can be seen in Figure 4.6E where the peak around 3 nucleotides disappears upon clustering correction.

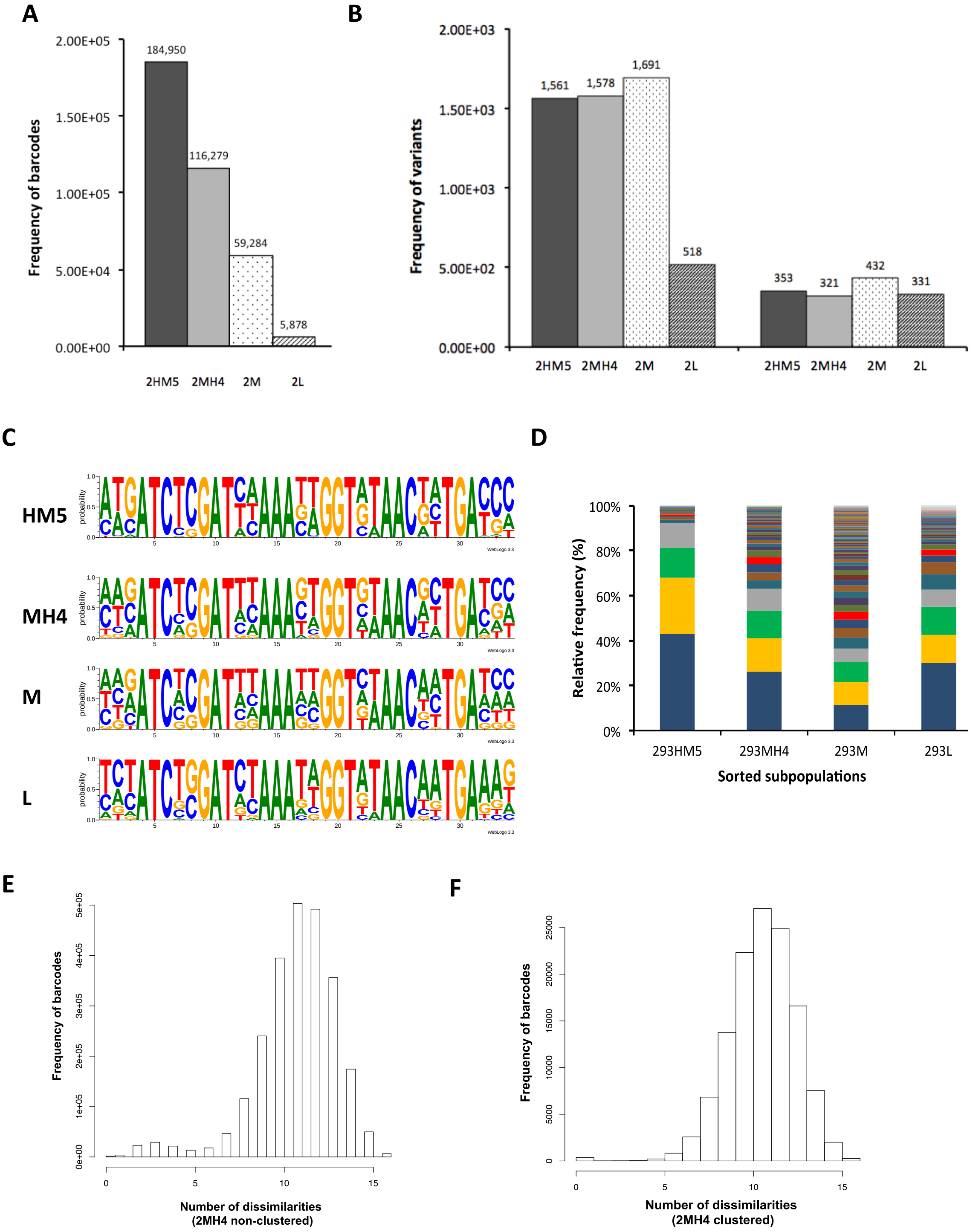


Figure 4.6. Characteristics of the barcoded counts retrieved from the custom sequencing library and analysed by next-generation sequencing.

(A) Frequency of barcodes after discarding and extracting barcodes using 'extract_rt-pcr_barcodes.pl' custom script. (B) Frequency of unique barcode variants before and after clustering using Starcode. (C) Pictogram of relative frequencies of nucleotides in sequenced barcodes Resource available at: weblogo.berkeley.edu.logo.cgi[704]. Adenine (A) is shown in green; cytosine (C) in blue; thymidine (T) in blue and Guanine (G) in yellow. Pictograms from top to bottom corresponding to HM5, MH4, M, L. (D) Stacked bar plot showing the relative abundance of each barcode within the barcode population. (E) Barcode comparison plot showing the number of nucleotide differences between clustered (right) and non-clustered (left) 2MH4 RNA-Seq barcodes.

Subsequently, clustering correction was applied to retrieved barcodes using Starcode. The thresholds for number of mismatches and 'size absorbing' ratio (explained below) were optimised for 2HM5, 2MH4, 2M and 2L. In DNA barcoding, since incomplete barcodes are not tolerated, mismatches are understood as transversions or transitions (not indels), also known as edited nucleotides or Leveinshtein distance (-d –distance- parameter in the command line instruction). The 'size absorbing' (-r) ratio is the number of times a cluster has to be larger in number of barcodes than another to be considered a single node. Two clusters with the same number of barcodes separated by a distance superior than the higher will not be clustered together unless fold difference in the number of barcode sequences is higher than the 'size absorbing' ratio[691].

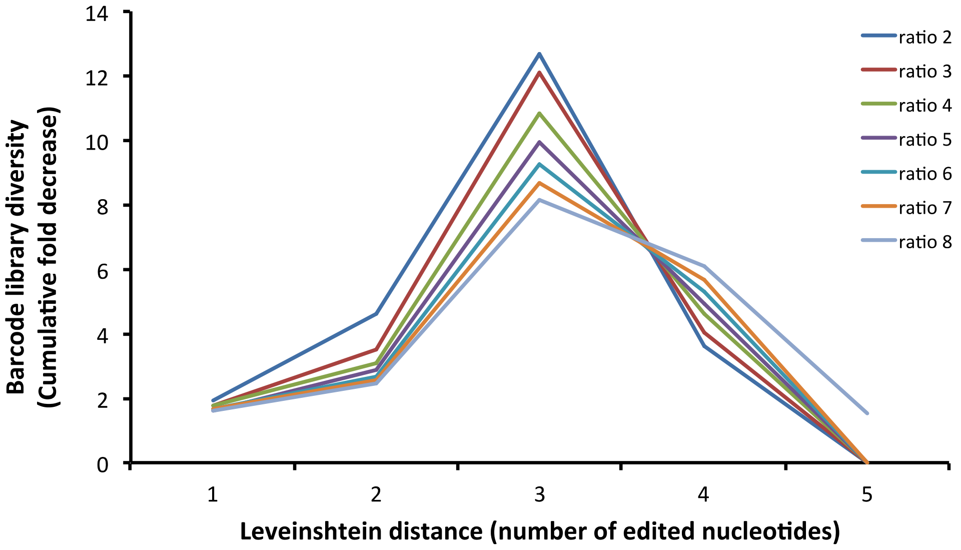


Figure 4.7. Effect of Starcode clustering parameters on frequency of barcode variants.

Increasing 'size absorbing' ratios made the clustering conditions more stringent and thus less distinct barcodes variants are obtained. (Default value is ratio=5).

Figure 4.7 shows a dramatic drop (8-12 fold) in the number of variants obtained allowing 3 mismatches. This number corresponds with the peak in the number of dissimilarities within the barcode sequence obtained cross comparing the different variants across themselves and confirms the noise-removal effect observed in Figure 4.6E around the 3 dissimilar nucleotides upon clustering correction (14 variable positions – 3 mismatches allowed = 11 peak in the frequency of dissimilarities observed between barcodes).

The number of variants was reduced around 5-fold after Starcode clustering (Figure 4.7). These results show a decrease of 3 orders of magnitude in the complexity of the barcoded library in transduced cells compared to values of barcodes and variants obtained in the plasmid and vector library (1,392,401/39,954 barcodes/clustered variants for pSYNT11 and 5,061,108/88,052 barcodes/clustered variants for vSYNT11). The same phenomenon (also with similar frequencies depending on samples) was observed in the LM-PCR experiment (Figure 4.4), where a few variants were also retrieved in a relatively high proportion. Evidence from independent sources (DNA junctions and barcode-specific RNA) consistently indicates that barcode representation is biased towards a few overrepresented variants. The main reason could be that a few clones may have overgrown the population over several rounds of passaging. Another potential explanation could be preferential amplification of certain vector-host DNA junctions or barcoded RNAs.

4.4.2 Primary sequence composition at chromosome-vector junctions

HIV integration site selection is not thought to be exclusively sequence driven. However, several authors described a short palindromic weak consensus in the chromosomal primary sequence immediately upstream of the U3 region in the 5'LTR (Table 4.2).

Table 4.2. Favoured target primary DNA sequences for HIV integration.

Publication	Weak primary consensus sequence
Carteau *et al.*, 1998[228]	5'-TNG(GTNAC)'CAN-3'
Holman and Coffin 2005[216]	5'- TDG(GTWAC)'CHA-3'
Wu *et al.*, 2005[755]	5'-TDG(GTNAC)'CHA-3'

Integration takes place in the position marked with (') on the top strand and the palindromic sequence in brackets is duplicated in the target site.

Chromosomal sequences at the integration site were analysed in HEK293 SA RIX reads. In HEK 293 SA RIX HM5, MH4 and 2M samples, the nucleotide pattern observed in Figure 4.8 (5'-TNGTAAH-3') does not correlate to that described in the literature (Table 4.2), given that the TNG motif in the genomic DNA is located only 4 nucleotides from the vector sequence (instead of 5nt, as expected for HIV). In the 293 2M sample, the nucleotidic pattern is more random and does not follow the weak consensus sequence whereas. In 293L the nucleotide composition is strongly biased possibly due to a clone that outgrew the population in the rounds of passaging during the month the cells were kept in culture. In addition to that, the presence of a thymidine in the 5' end of the integration site is not expected since it has been reported to cause steric hindering with the phosphate backbone of the newly made bond[172].

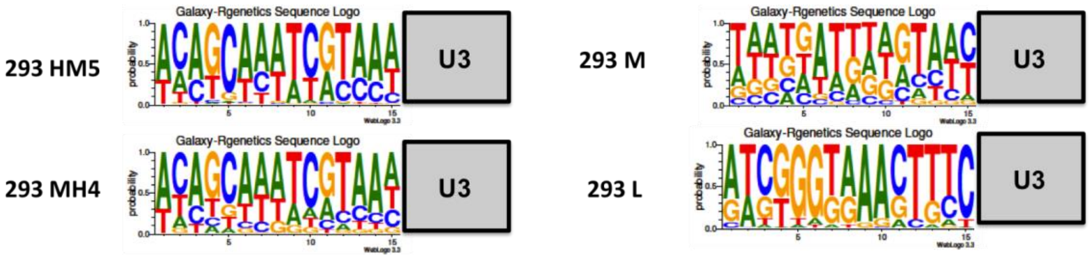


Figure 4.8. Pictogram showing relative frequencies of nucleotides on the 5'vector LTR-host chromosome junction in HEK 293 SA RIX transduced with vSYNT11.

(A) Resource available at: weblogo.berkeley.edu.logo.cgi[704]. Adenine (A) is shown in green; cytosine (C) in blue; thymidine (T) in blue and Guanine (G) in yellow. (A) Primary sequences retrieved from HEK 293 SA RIX sorted subpopulations.

4.3.5 Correlation of integration site analysis and RNA counts on HEK 293 SA RIX data

Barcode variants from the different groups (2HM5, 2MH4, M, L) were ranked based on their abundance of RNA counts after clustering correction. The top 10 most expressed variants were interrogated to the LM-PCR dataset containing genomic positions associated with a particular barcode (Table 4.1).

Although barcodes were clustered using Starcode, none of the top 10 barcodes changed in sequence or position (based on number of counts) compared to their position prior to clustering. These results indicate that pre-clustered variants were distant in sequence similarity (at least >2 editing or Leveinshtein distance, the value of the parameter used in Starcode).

From the correlation of both datasets, several integration sites were found to be associated with the same barcode variant. However, not all the genomic positions recovered by LM-PCR had the same number of reads. Some cases did not allow unique association of a barcode to a single integration site whereas in others, a particular integration site was substantially more represented than the rest of genomic positions. The threshold for establishing endposition_chr abundance obeys the following formula:

$$\frac{\text{\#reads of the most represented genomic position}}{\text{\#reads of the 2nd mostrepresented genomic position}} > 10$$

A 10-fold threshold difference in the ratio between the first and second highest number of LM-PCR counts was established as a filter to consider the association a clear signal rather than background noise. The criterion to set to the '10-fold ratio' threshold was based on the profile of signal/noise observed in different hits (linear increase for background noise versus logarithmic increase for *real* integration sites).

Table 4.3. Correlation of RNA counts with vector integration sites using the barcode.

Sub-population	Top #	#RNA counts	Barcode variant sequence	#LM-PCR reads	Signal*	Genomic position	Chr	Gene
2HM5	1	80,324	ATG-TC-CA-TA-AT-TA-CCC	145	<10-fold	**		
	2	46,673	CAC-TC-TC-GT-GT-GC-TCA	314	<10-fold	**		
	3	25,523	ATG-TC-TT-CT-CA-AT-CGC	none	NA			
	4	21,184	AAG-CG-AT-CT-GT-GT-TGT	29,717	11.3	88,071,769	5	MEF2C
	5	2,641	TAA-CG-AT-GC-CA-AC-AAC	144	<10-fold	**		
	6	2,178	ACA-AG-AG-CA-CA-AC-TTC	69	<10x			
	7	1,184	TCA-GC-GT-TC-CA-AT-AAA	70,947	11.5	4,730,541	7	FOXK1
	8	1,045	AGA-TC-TT-TC-GA-CT-GCC	none	NA			
	9	756	CGG-CC-TG-TT-AT-GA-CTA	5	<10-fold	**		
	10	554	TAG-CC-TT-AC-CT-AA-TCG	none	NA			
2MH4	1	29,948	CAC-TC-TC-GT-GT-GC-TCA	146	<10-fold	**		
	2	17,670	ATG-TC-TT-CT-CA-AT-CGC	none	NA			
	3	14,530	ATG-TC-CA-TA-AT-TA-CCC	16	<10-fold	**		
	4	12,351	AAG-CG-AT-CT-GT-GT-TGT	11,863	10.1	88,071,769	5	MEF2C
	5	4,457	CGA-AC-GT-GT-TT-TC-TAT	none	NA			
	6	4,115	TAA-CG-AT-GC-CA-AC-AAC	267	<10-fold	**		
	7	3,992	CCC-GC-TC-GG-GA-CC-TAT	7,394	13.2	152,185,965	3	
	8	3,664	ACA-AG-AG-CA-CA-AC-TTC	110	<10-fold	**		
	9	3,529	GTC-AG-TT-AT-TA-CC-GTT	none	NA			
	10	2,188	CAA-AC-CT-TA-AT-AT-AAC	none	NA			
2M	1	6,535	ATG-TC-CA-TA-AT-TA-CCC	none	NA			
	2	6,046	TAG-CC-TT-AC-CT-AA-TCG	none	NA			
	3	5,082	ACA-AG-AG-CA-CA-AC-TTC	none	NA			
	4	3,582	CAC-TC-TC-GT-GT-GC-TCA	none	NA			
	5	2,816	TCA-GC-GT-TC-CA-AT-AAA	6	<10-fold	**		
	6	2,581	AGA-TC-TT-TC-GA-CT-GCC	none	NA			
	7	2,331	CCC-GC-TC-GG-GA-CC-TAT	none	NA			
	8	2,265	CGA-AC-GT-GT-TT-TC-TAT	none	NA			
	9	1,978	ACC-AC-AC-AT-AA-AA-AAA	none	NA			
	10	1,866	CGG-CC-TG-TT-AT-GA-CTA	none	NA			
2L	1	1,905	TCT-TG-CT-TA-AT-AA-AAG	2	<10-fold	**		
	2	748	CAT-GC-CT-TT-GT-AA-GAA	none	NA			
	3	663	TGC-TG-AC-TT-AT-TG-AGT	none	NA			
	4	426	CAC-TG-AT-TA-CA-AA-ATC	none	NA			
	5	350	TTT-TC-GT-GC-TT-AT-TCG	none	NA			
	6	334	ACA-TC-CC-GT-AT-CA-AAG	none	NA			
	7	310	CCC-CC-GT-AG-GT-GA-AAT	9,718	1618.2	108,685,768	6	LACE1
	8	155	AAG-TC-TG-CG-TT-CT-GAT	none	NA			
	9	150	TCT-CG-TC-CA-TT-TC-AAT	16	<10-fold	**		
	10	115	TAC-CC-CC-AC-GT-AA-AAT	none	NA			

Barcodes present in >1 dataset are highlighted with the same colour. (*) Signal refers to the ratio between LM-PCR reads recovered between the first and the second genomic position with the same barcode. Column 1, sorted subpopulation of HEK 293 6E based on GFP intensity; column 2, position in the ranking of top expressed barcode variants; column 3, number of barcode counts per barcode variant retrieved by RT-PCR and NGS; column 4, variable nucleotides extracted from the whole barcode in this particular variant; column 5, number of reads retrieved by LM-PCR; column 6, Signal (ambiguity measurement) – ratio between the number of LM-PCR reads between the top two genomic positions associated with the same barcode variant; column 7, chromosomal position; column 8, chromosome; column 9, RefSeq gene associated with the genomic position. (**) Integration sites with a signal below the threshold (10-fold) were not mapped.

Association of RNA and LM-PCR barcodes derived from datasets 2HM5, 2MH4 and 2L under the aforementioned premises resulted in 4 candidate genomic positions associated with different genes:

- *FOXK1* (also known as myocyte nuclear factor, MNF) stands for forkhead box protein K1 and is RNA polymerase II regulator that binds to the upstream enhancer region (CCAC box) of myoglobin gene.

- *MEF2C* is a protein-coding gene whose position was retrieved from both 2HM5 and 2MH4 datasets. MEF2C (myocyte specific-enhancer factor 2C) is a transcription factor, which binds to regulatory regions of muscle-specific genes. As well as FOXK1, it is involved in cardiac morphogenesis and myogenesis. It also plays a role in neuronal development and it is necessary for megakaryocytes, platelets, and B-cell lymphopoiesis.

The third position (chr3:152,185,965) is not associated with a gene but it is located <3kb downstream of the human musclebind like splicing regulator 1 (*MBNL1*) gene, a zinc-finger protein that participates in alternative splicing of myogenic pre-mRNAs.

- *LACE1*, whose signal/noise ratio was 100 times higher than the rest of positions, stands for lactation-elevated protein 1 and is a protein-coding gene with a possible ATPase function.

However, screening of 300-500 genomic positions/barcodes does not represent a significant improvement compared to current selection methods for high expressing clones. Due to the high degree of library complexity (throughput) lost over several weeks of culture, a new strategy was explored where integration sites were screened for expression 10 days post-transduction. This strategy maximises the maintenance of entropy/complexity by transduction of the barcoded library in detriment of stability of gene expression. From a biological point of view, this strategy would present a problem if gamma-retroviral vectors were used to screen the genome for high expressing sites due to the gene silencing associated with CpG islands, more predominant in the TSS, where gamma-retroviral vector are more likely to integrate. However, since lentiviral vectors are being used in this study and these vectors tend to integrate along the transcription unit, this risk becomes minor.

4.4 Integration site analysis and RNA expression on vSYNT-transduced HEK 293 6E

4.3.4 Next-generation sequencing of integration sites in HEK 293 6E

In addition to the modifications introduced in the last section, our Cell and Gene Therapy CMC group at GlaxoSmithKline switched from HEK 293 SA RIX to HEK 293 6E cell lines. This cell line is also suspension adapted, grows in serum-free media and according to the Biotechnology Research Institute of the National Research Council of Canada (NRC-BRI) allows for high yield production of viral vector and recombinant proteins[374,721]. The HEK 293 6E cell line expresses a truncated form of the Epstein Barr Virus (EBV) nuclear antigen (EBNA) 1 which enables episomal persistence and amplification of plasmids possessing the EBV oriP sequence with yields up to 10-fold higher for antibody production[722]. This feature together with a faster growth rate make this host cell line suitable for large-scale transfection and biomanufacturing[376].

Following the procedure described in Section 3.5.1, either 10^3, 10^4 or 10^5 HEK293 6E cells were transduced with vSYNT11 at an MOI of 1 and cultured for a week before RNA and DNA harvest. Such numbers of cells respond to the expected library complexity and throughput that the barcode method is expected to reach given the results obtained in Table 3.1. pSYNT library size and diversity details pre- and post-Starcode clustering correction. The introduction of a known number of transducing units delivered to a known number of cells constitutes an internal control for recovery of integration sites. Genomic DNA from HEK 293 6E cells transduced with vSYNT11 (at a MOI of 1) was extracted and integration sites captured by ligation-mediated PCR (LM-PCR) as described in Section 4.3 for HEK 293 SA RIX host cell lines. LM-PCR was performed using 4 restriction enzymes

(*DraI, NlaIII BsuRI* and *HpyCH4v*) to screen a higher proportion of sequences with different characteristics.

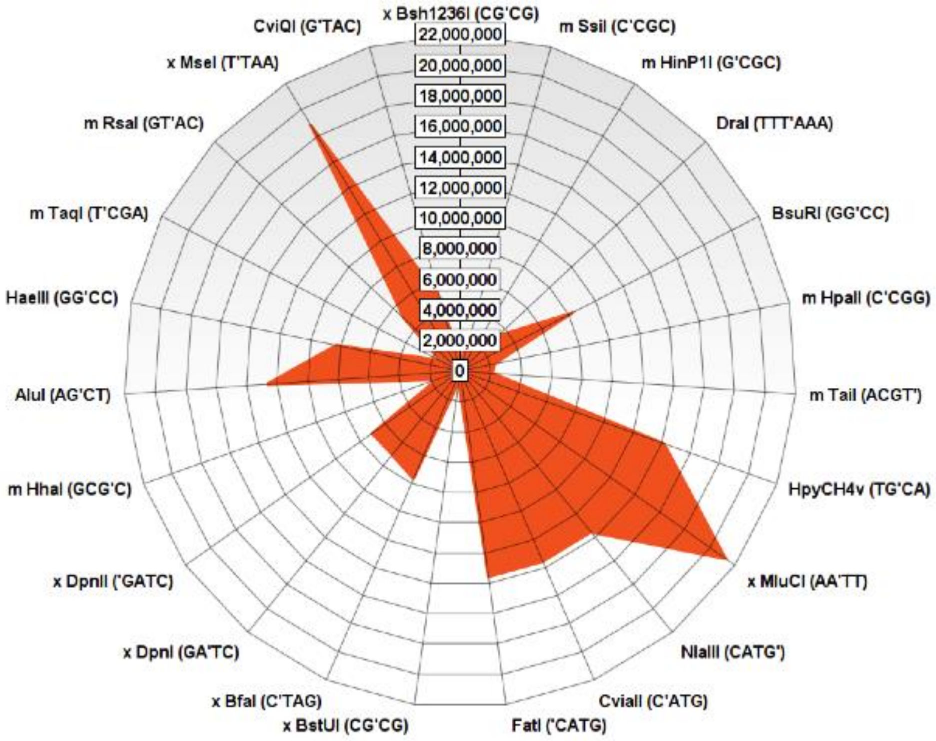


Figure 4.9. Radar diagram showing the frequency of different 4-base cutters in the human genome (*H. sapiens* GRCh37/hg19).

Enzyme sites labelled with an 'm' were discarded for being blocked by methylation or for being present in the LTR/ (labelled with 'x'). All enzymes shown are palindromic (except *SsiI*) and thus are recognised in both strands of DNA.

The choice of restriction enzymes was also optimised to maximise the access to genomic fragments with different GC content and cutting fashions and discarded those cutting within the LTR or linker sequence or blocked by methylation. Regarding the choice of restriction enzyme, *NlaIII*, *BsuRI* and *HpyCH4v* are the 4-base cutters not present in the LTR and/or linker sequence with more representation in the human genome (Figure 4.9). Their recognition sequence is also different (CATG', GG'CC and TG'CA, respectively), with different properties (blunt/sticky, GC/AT rich). Nonetheless, some studies showed no significant differences in integration preferences detected between restriction enzyme compared to an *in vitro* integration control[249] or cloning of integration sites using different restriction enzymes[228,234,249].

The presence of fixed nucleotides in the barcode stretch minimizes the number of restrictions sites randomly generated in barcodes potentially cleaved to 10 combinations (using *DraI, NlaIII, BsuRI* and *HpyCH4v*) (Table 4.4).

Table 4.4. LM-PCR restriction sites found in barcode sequence

Restriction enzyme	Barcode sequence
HpyCH4v (5'-TG'CA-3')	**TAa**tgcATCNSGATNNAAANNGGTNWAACNNTGANNN**TGG**
	TAANNNATCNSGAtgcaAANNGGTNWAACNNTGANNN**TGG**
DraI (5'-TTT'AAA-3')	**TAA**NNNATCNSGAtttaaaNNGGTNWAACNNTGANNN**TGG**
NlaIII (5'-CATG'-3')	**TAA**NNNATCNSGATNNAAANNGGTNWAACcatgANNN**TGG**
	TAANNNATCNSGATNNAAANNGGTNWAACNNTGANca**tgG**
Original barcode	**TAA**NNNATCNSGATNNAAANNGGTNWAACNNTGANNN**TGG**

Sequences flanking the barcode and restriction sites are highlighted in bold and lowercase, respectively.

Retrieval of integration sites by TOPO TA shotgun cloning prior to next-generation sequencing was performed as described for HEK 293 SA RIX LM-PCR with *DraI* with similar recovery rates (data not shown). From the number of bands that can be seen in the electrophoresis gel (Figure 4.10A) the polyclonality of the samples seems to be higher than that of HEK 293 SA RIX samples (Figure 4.2A). Even though a single junction would represent a minimal band in a DNA smear on the gel, the presence of several visible bands anticipates the recovery of more (and more diverse) IS.

LM_PCR products from 10^3, 10^4, 10^5 cells transduced at a MOI of 1 were recovered by next-generation sequencing following the same strategy as in SA RIX integration sites (MiSeq 300bp PE). Around 1.7M reads R1 (left) and R2 (right) reads were obtained for each of the 3 samples (10^3, 10^4, 10^5 cells) and 4 restriction enzymes (12 total). The same pipeline applied previously on 293SA RIX 2HM5, MH4, M, L fastq files was used to process from HEK 293 6E integration sites. R1 (left) and R2 (right) integration sites/barcodes retrieved from the same sample with different enzymes were collated together and trimmed to optimise merging and subsequently merged using Burrows-Wheeler Aligner (BWA) pemerge tool. Around 3.4M out of 6.4M reads (53%) were successfully merged across samples.

Quality control of the merged sequences for the ‘10^4 cells’ dataset revealed a balanced global GC content (Figure 4.10C) as well as good quality calls along the length of the read (Figure 4.10D). The frequency of reads increases exponentially beyond a mean sequence quality score of 30 (Figure 4.10 B) and the frequency in read length of the merged reads decreases between 150 and 290bp (Figure 4.10E). Compared to SA RIX read lengths the presence of numerous band sizes observed in this HEK 293 6E LM-PCR anticipates a higher diversity of junction lengths (Figure 4.10A). These results were comparable for the other two datasets (10^3 cells and 10^5 cells).

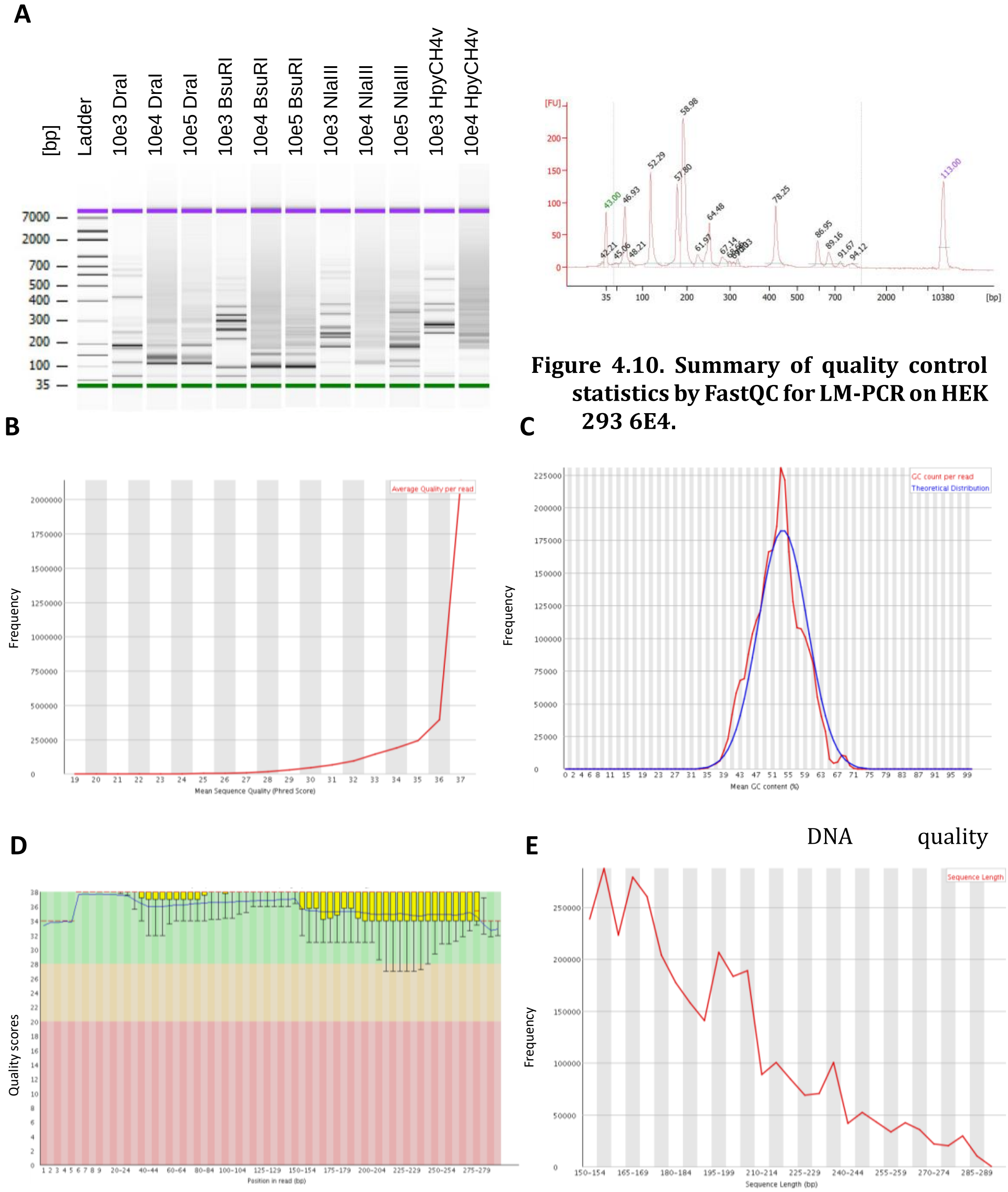


Figure 4.10. Summary of quality control statistics by FastQC for LM-PCR on HEK 293 6E4. DNA quality assessment of vector-chromosome junctions for HEK 293 6E 10e3 LM-PCR products digested with *DraI* by microchip-based capillary electrophoresis using Agilent Bioanalyzer; results given in fluorescent units (FU). Peaks at 35bp and 10,380bp are internal controls. (B) Mean sequence quality distribution over all sequences. (C) GC distribution (mean GC%) over all sequences. Similar results were obtained for LM-PCR6E3 and LM-PCR6E5, also sequenced using MiSeq 300bp PE (data not shown). (D) Distribution of quality scores across all 100bp of the read (Sanger/Illumina 1.9 encoding). The yellow boxes represent the inter-quartile range (25-75% of the score values per bp). The upper and lower whiskers indicate the 10th and 90th percentiles; the blue line represents the mean quality. The background of the graph divides the y-axis into very good quality calls (green), calls of intermediate quality (orange) and calls of poor quality (red). (E) Frequency of read lengths after trimming 3´ends and merging with BWA pemerge. Results shown for LM-PCR6E3 and LM-PCR6E5 are comparable to the results shown.

After quality control, a custom script previously used in Section 4.3.2 was used to extract integration sites and barcode sequences. Around 2.4M (71%) IS/barcodes were successfully retrieved across the 3 samples. IS were mapped using Blat against the human genome (same version and parameters as described in Section 4.3.2). Out of the 21,522,222; 32,644,849; and 30,650,729 raw alignments with various scores output by Blat, 807,411 (82%); 1,089,554 (87%) and 836,938 (86%) non-ambiguous integration sites were retrieved for 10^3, 10^4, 10^5 HEK 293 SA RIX cells, respectively (Figure 4.11.A).

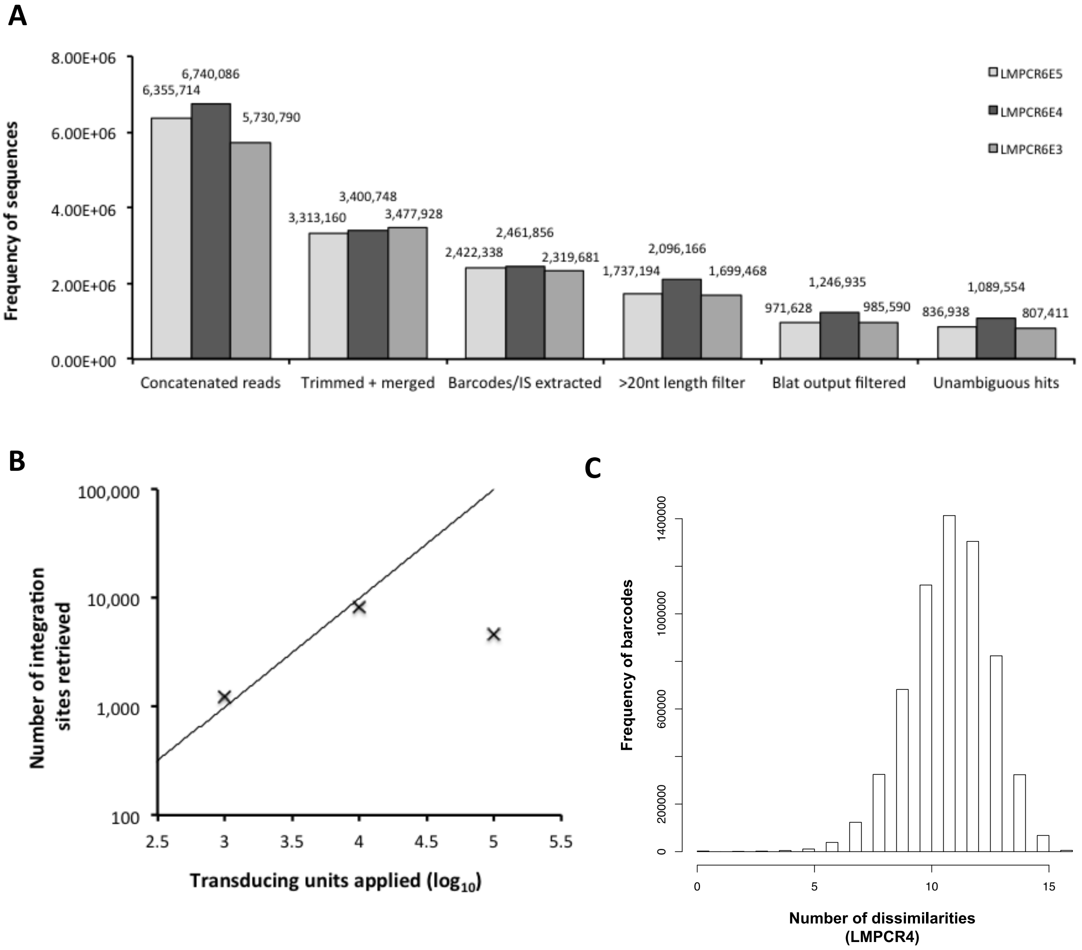


Figure 4.11. Barcode retrieval from LM-PCR reads using integration site processing pipeline

(A) Frequency of sequences obtained throughout the integration site-processing pipeline. (B) Correlation between the number of transducing units HEK 293 6E were transduced with vSYNT11 at an MOI of 1 and the number of IS retrieved after the bioinformatics pipeline. (C) Barcode comparison plot showing the number of nucleotide differences between all the sequenced LMPCR4 barcodes. LMPCR3 and LMPCR show comparable profiles (data not shown).

Since psl files contained repeated hits retrieved with different restriction enzymes, the different retrieved loci were counted based on the expression endposition_chr and 8,261 different IS were obtained out of the 1,089,554 unambiguous sequences of the 10^4 psl file. These ~8,000 integration sites correspond to the 10,000 transducing units (TU) that were applied during transduction (Figure 4.11B). The +1,700 missing IS could be partially justified by short junctions discarded in the PCR product purification prior to NGS submission. Another source of missing barcodes could be unbarcoded lentiviral vectors, which transduced cells but did not deliver a tag into their genome. Similarly 1,245 were retrieved in a sample transduced with 1,000 TU (10^3 sample). However, the '10^5' sample did not follow the expected number of transducing units (4,604 seen vs ~100,000 expected). A potential explanation could be that transduction was done in a 24-well plate format instead of 96-well plate as for 10^3 and 10^4, so the distribution of the viral vector over the cells was not uniform and consequently less TU were retrieved. The number of nucleotides dissimilarities in the barcodes of the viral and plasmid libraries is maintained at 11 nucleotides in LM-PCR junctions (Figure 4.11C).

4.4.2 Primary sequence composition at chromosome-vector junctions

In HEK293 6E, all samples display a similar nucleotide composition compatible sequence with the reference pattern (summarised with the expression TDG(G)TAAC) although the proportion of the nucleotides is better balanced than those retrieved in HEK 293 SA RIX (Figure 4.12B). This result reinforces the observation that HEK 293 SA RIX junctions have less diversity and polyclonality than HEK 293 6E.

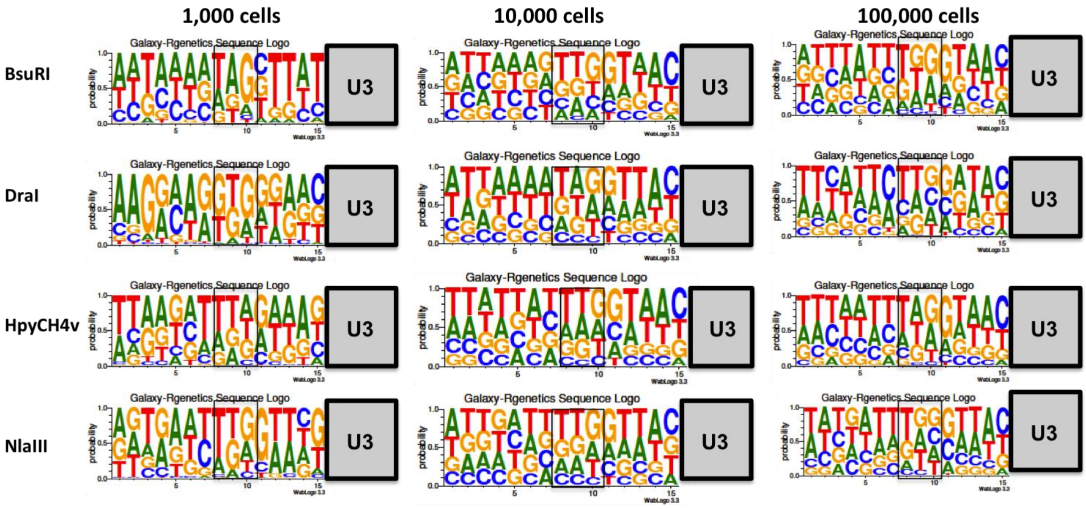


Figure 4.12. Pictogram showing relative frequencies of nucleotides on the 5'vector LTR-host chromosome junction.

Resource available at: weblogo.berkeley.edu.logo.cgi[704]. Adenine (A) is shown in green; cytosine (C) in blue; thymidine (T) in blue and Guanine (G) in yellow. (A) Primary sequences retrieved from HEK 293 SA RIX sorted subpopulations. Primary sequences retrieved from HEK 293 by LM-PCR with different enzymes. The 5’-TNG-3’ trinucleotide of the 5’-TNG’(GTNAC)CAN-3’ pattern of the is indicated in a square.

4.4.3 Genomic distribution of lentiviral vector integration sites

Integration sites were retrieved from HEK 293 6E 10^4 and 10^3 sample and plotted on chromosomes arranged in a karyotype-like fashion using the Galaxy karyotype plotting tool (Figure 4.13). In parallel, the same number of random genomic positions (8,261) were generated using VISA (Vector Integration Sites Analysis server, default filtering parameters)[696] to use as a control and were processed in the same manner as HEK 293 NGS files.

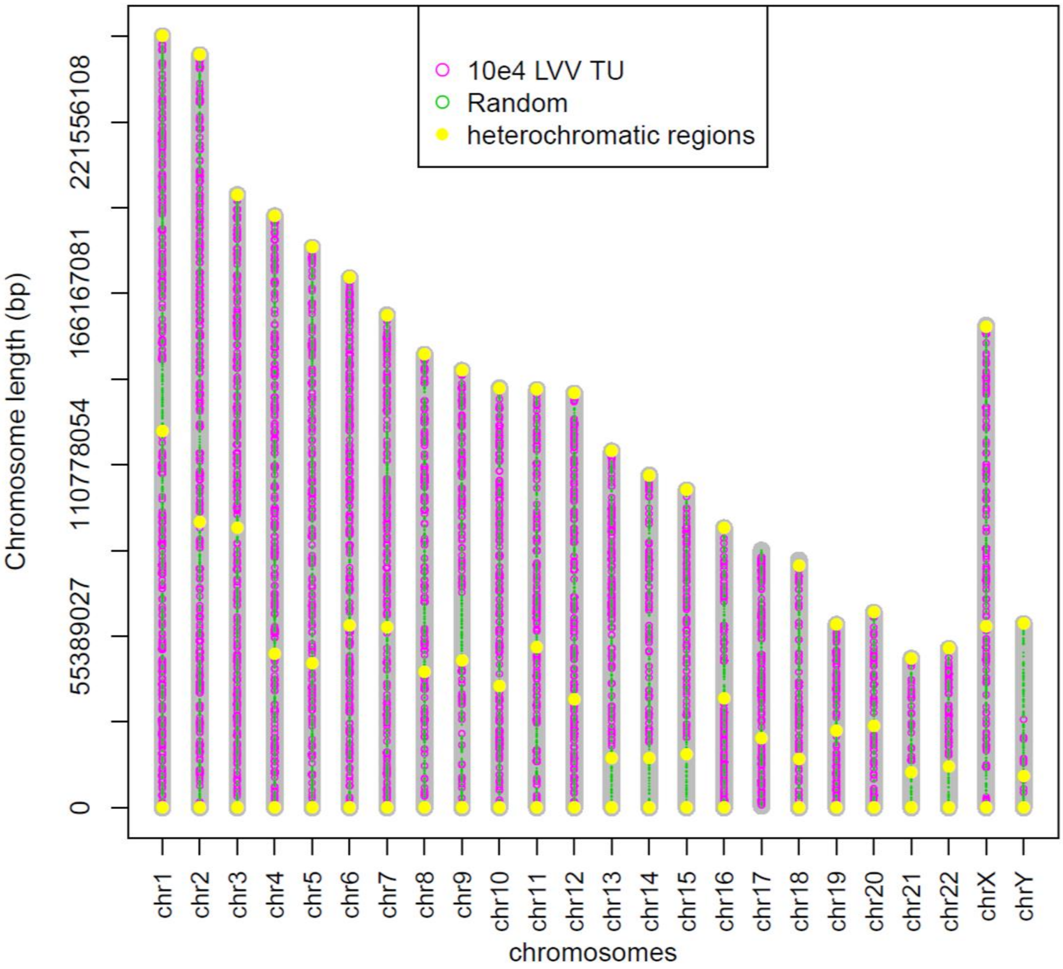
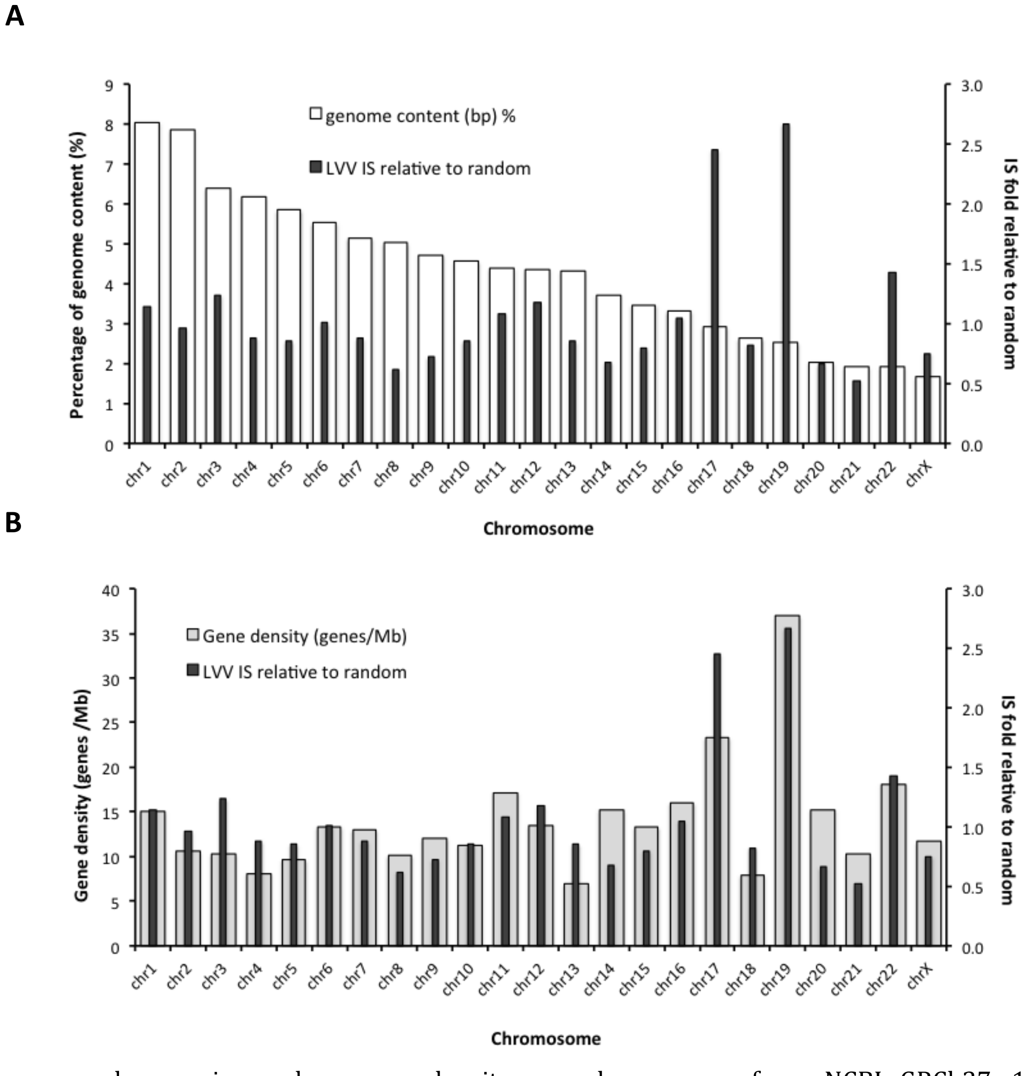


Figure 4.13. Distribution of vSYNT integration sites in the human genome.

Karyotype view of integration sites in the human genome. 10^4 HEK 293 6E cells were transduced with 10^4 LVV TU resulting in 8,261 and an equivalent amount of randomly generated sites represented in magenta and green, respectively. Random integration sites were generated using the online tool "Vector Integration Site Analysis" from Trobridge Lab at University of Washington, College of Pharmacy (https://visa.pharmacy.wsu.edu/bioinformatics)[696]. A similar profile is obtained when plotting 10^3 HEK 293 6E cells were transduced with 10^3 LVV TU (data not shown). Heterochromatic regions (displayed in yellow) were downloaded from the UCSC Table browser 'Gap' database creating a filter "centromere telomere".

As can be seen in Figure 4.13, lentiviral vector integrations are less frequent in centromeric regions, rich in constitutive heterochromatin[756], compared to a random integration profile. Figure 4.14 demonstrates that the frequency of lentiviral integration sites relative to random integration does not correlate with the chromosome size (Figure 4.14A) but with gene density. These results are consistent with the earliest models[204–206] proposing chromatin conformation playing a key role on integration site selection, disfavouring integration in heterochromatic regions (Figure 4.13) and displaying positive biasing towards RefSeq genes (Figure 4.14B). As more recently described by Wang *et al.*, in 2009 and Biffi *et al.*, in 2011 lentiviral vectors tend to integrate into gene dense chromosomes such as chromosome 17, 19 (with >20genes/Mb), 3 and 22 (less gene dense) in comparison to a random integration pattern (Figure 4.14B)[225,757].

Figure 4.14. Integration sites relative to random displayed on genome content (A) or gene density (B) per chromosome.

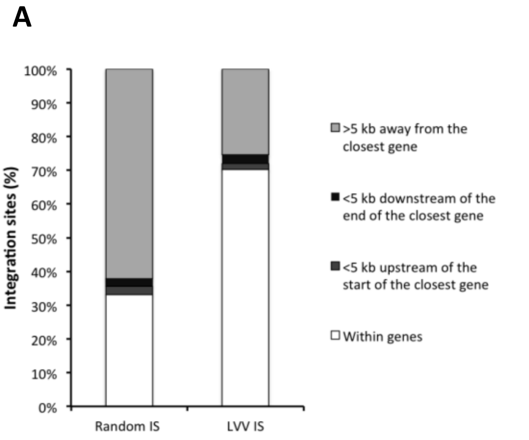


Chromosome base pairs and genome density per chromosome from NCBI GRCh37.p13 (https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.25). Random integration sites were generated using the online tool "Vector Integration Site Analysis" from Trobridge laboratory at University of Washington, College of Pharmacy (https://visa.pharmacy.wsu.edu/ bioinformatics)[696].

Bedtools was used to associate integration sites coordinated to different annotation features. Similarly to the rates described in the literature[249,758,759], 71% ($p<0.0001$) of the integrations (n=5,891) were found within genes (Figure 4.15A), a percentage significantly higher than the frequency of randomly generated IS integrated within genes (35%, generated using VISA[696]). The latter is consistent with the proportion of transcription units in the human genome (33%)[229]. LVV IS hit 3,032 different genes, 17% of the 18,041 genes. Within the gene, integrations were predominantly located within the transcription unit (TU) (Figure 4.15D) evenly distributed along the length of the gene (Figure 4.15C) ($p=0.318>0.05$, null hypothesis). Such phenomenon is due to the tethering effect LEDGF/p75 protein. Depletion of this protein has been reported to reduce the preference towards transcription units[243].

When looking at integration within repetitive elements, LVV IS were found to be underrepresented compared to random ($p<0.05$) (Figure 4.15B) agreeing with Stevens and Griffith and Moiani *et al.*,[251,760]. Frequencies lower than random were observed in LINE, LTR and DNA elements and especially in Satellite DNA, predominantly found in centromeres and telomeres, theoretically disfavoured due to their heterochromatic conformation. On the contrary, IS located in short interspersed nuclear elements (SINEs) were more frequent when compared to random, as expected given their location close to RNA polymerase II promoters[256]. The same procedure was applied to compare the genomic positions retrieved from the HEK 293 6E 10^3 sample, which reported comparable results.

Custom tracks containing annotation for repetitive elements were obtained from UCSC Table browser, which utilises data from ‘Repbase update library of repeats’ from the GIRI (Genetic Information Research Institute)[695].



B

Repetitive element	Random IS (%)	LVVvSYNT11 IS (%)
SINE	14.32	16.21 (p=0.32)
LINE	19.40	18.12 (p=0.61)
LTR elements	10.43	4.19 (p=0.0002)
DNA transposons	2.56	0.71 (p=0.001)
Satellites	0.41	0.04 (p<0.0001)
Others	4.67	4.8 (p=0.43)
Total	**51.79**	**44.07 (p<0.05)**

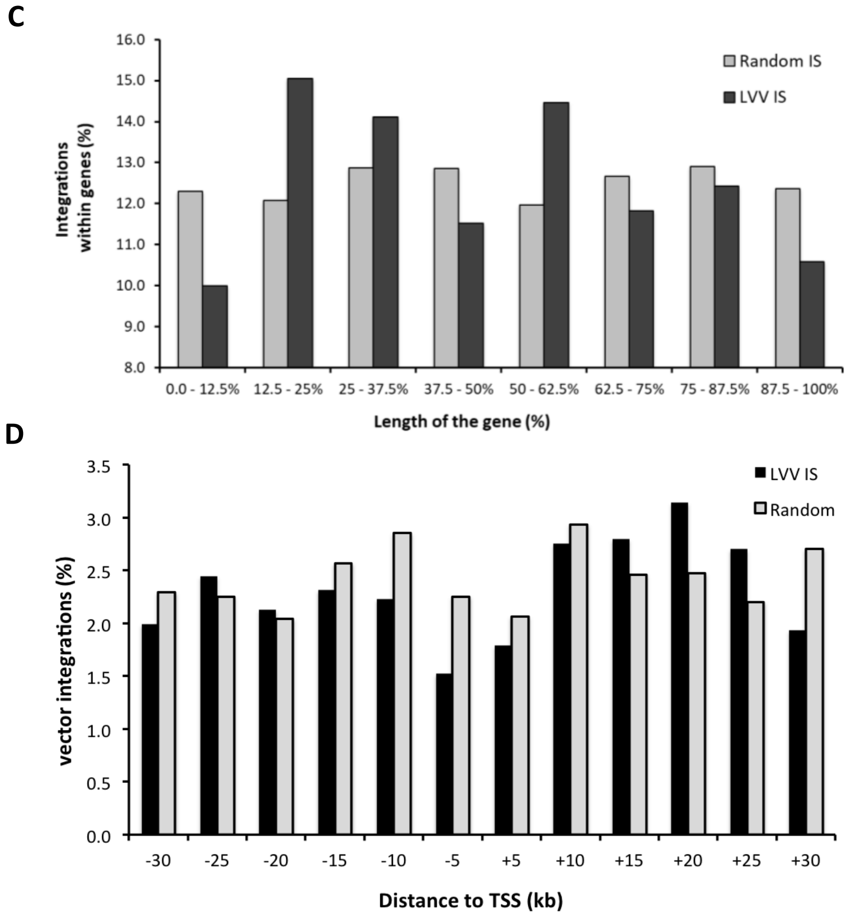


Figure 4.15. Distribution of vSYNT vector library IS compared to genomic annotation features.

(A) Distribution of vector integration sites relative to RefSeq genes compared to random integration profile. (B) Distribution of vector IS relative to repetitive elements compared to a random integration profile (values in brackets show percentage, n=8,261). (C) Distribution of vector integrations along the length of the gene; length of the gene body was segmented into 8 fragments. (D) Distribution of vector integrations relative to the transcription start site (TSS) of RefSeq genes relative to random. Chi-squared tests were performed to determine whether the probabilities are due to chance using the same number of random generated integrated sites. P-values are determined from the Chi-square statistics; Yate's correction was applied in the case of a 2x2 contingency table and 1 degree of freedom.

4.4.4 Study of the barcode complexity and duplication across samples

LMPCR6E integration sites from 10e3, 10e4 and 10e5 cells were cross-compared and <5% of them were found to be repeated in different samples (Figure 4.16A). Given that the chances of a viral vector genome independently integrating in the exact genomic position multiple times are extremely low, this phenomenon could be explained by the existence of cross-contamination between samples. LMPCR6E integration sites with multiple copies were discarded from the candidate list to avoid duplicities in the retrieval or correlation of barcode variants.

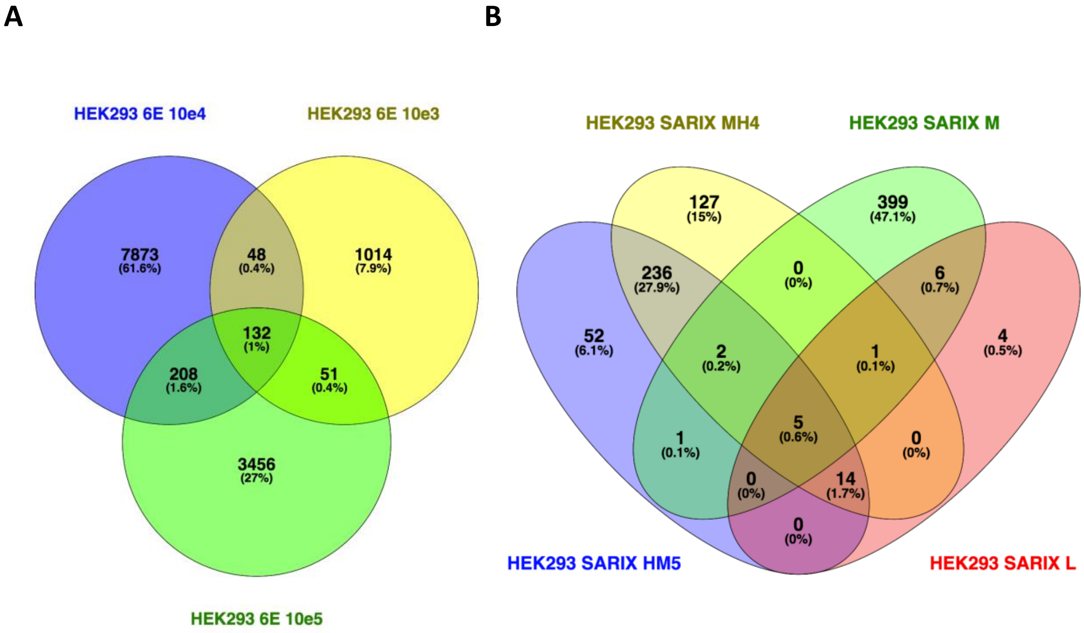


Figure 4.16. Venn diagrams representing the number of overlapping integration sites found between samples.

Percentage of total sequences per sample indicated in brackets.

In the case of HEK293 SA RIX cells, sorted into 4 subset populations 30 days post-transduction, MH4 and HM5 present the highest number of duplicities, which can be attributed to the fact that cells with the same origin were sorted into different subpopulation due to having similar fluorescence intensities. Percentages of integration site duplicities in the rest of groups are negligible (Figure 4.16B).

4.5.2 Determination of the relative abundance of barcode in vSYNT11-transduced HEK 293 6E cells

In parallel to the DNA harvest of transduced (MOI of 1) 10^3, 10^4 and 10^5 samples and in order to perform integration site analysis, RNA was also extracted to quantify the expression of each barcode variant in that same sample. DNAseI was applied to 1µg of total RNA and the mixture was column-purified to avoid any amplification from DNA and conditions were optimised in order to synthesize cDNA and amplify a specific region of the RNA in a single reaction. RNA was reverse-transcribed and amplified with primers binding a region flanking the barcode. A band of the expected size (220bp) was observed in prior to next-generation sequencing of cDNA (Figure 4.17A). Nomenclature adopted for these samples is RTPCR 3, 4 and 5 for specific PCR products amplified from reverse transcribed RNA.

PCR reactions were purified and sent for sequencing at Genewiz using HiSeq 2500 100bp PE sequencing strategy (custom-made libraries were not available). Close examination of reads from both orientations revealed barcodes that could successfully be extracted from R2 reads. Consequently, >10M input R2 reads per sample were reverse complemented with a simple script called 'rc_fastq.pl' (Appendix B) and quality trimmed to discard those with an excessive number of missing bases (N>5).

Quality control of 100bp R2 reads showed base calls of good quality along the length of the read (Figure 4.17B) in most reads (Figure 4.17C) and the GC content was well balanced (Figure 4.17D).

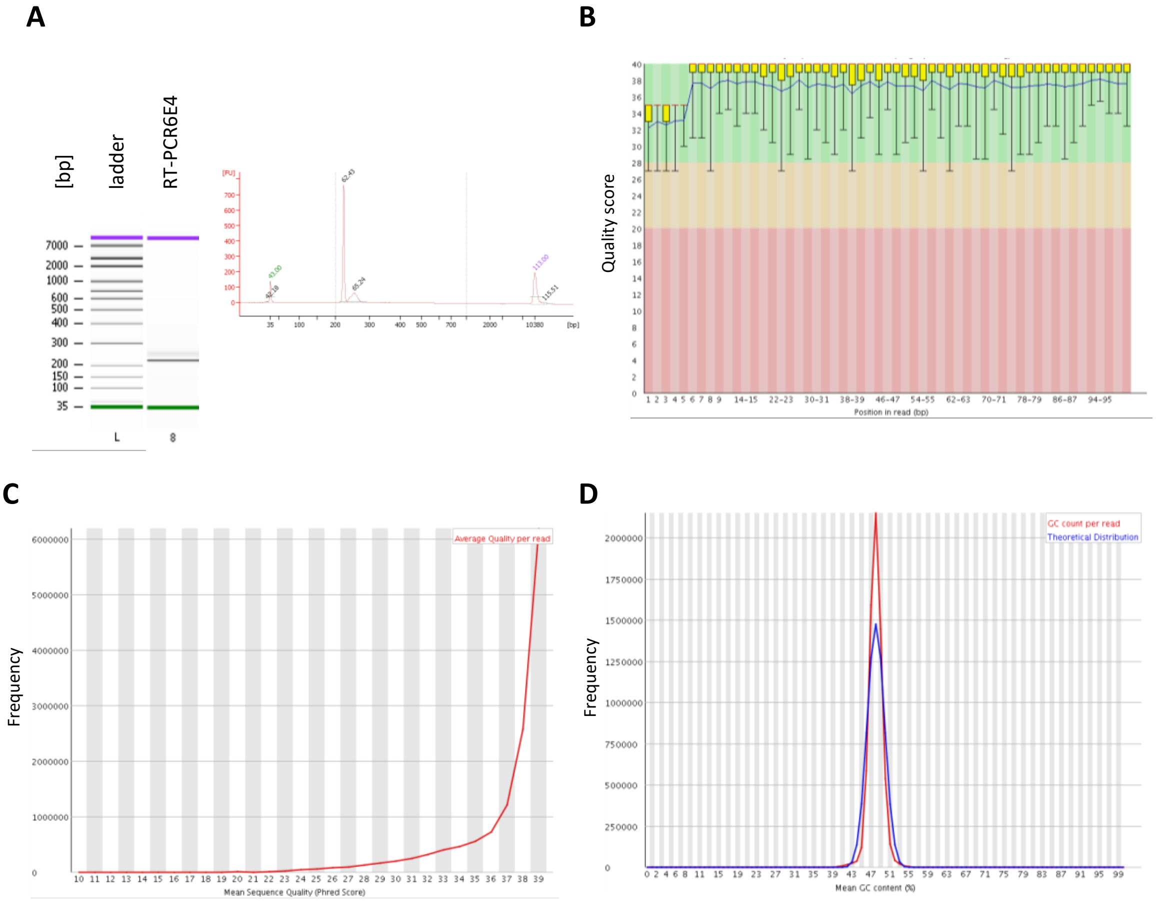


Figure 4.17. Summary of quality control statistics by FastQC performed on RT-PCR6E4.

(A) DNA quality assessment by micro chip-based capillary electrophoresis using Agilent Bioanalyzer showing a 220bp band corresponding to the barcoded band reverse transcribed from cellular RNA. On the right, peaks of DNA are given in fluorescent units (FU). Peaks at 35bp and 10,380bp are internal controls. (B) Distribution of quality scores across all 100bp of the read (Sanger/Illumina 1.9 encoding). The yellow boxes represent the inter-quartile range (25-75% of the score values per bp). The upper and lower whiskers indicate the 10th and 90th percentiles; the blue line represents the mean quality. The background of the graph divides the y-axis into very good quality calls (green), calls of intermediate quality (orange) and calls of poor quality (red). (C) Mean sequence quality distribution over all sequences. (D) GC distribution (mean GC%) over all sequences. (E) Similar results were obtained for RT-PCR6E3 and RT-PCR6E3, also sequenced using HiSeq2500 100bpPE (data not shown).

After quality control, barcodes were extracted with a script called 'extract_rt-pcr_barcodes.pl' (Appendix B) similar to the script 'extract_viral_insertion _barcodes.pl', which recognises 46 and 39nt LTR sequences flanking the barcode and extracts the barcode sequence.

Out of >12M initial reads per sample, over 99% of the sequences passed the filter and yielded 4.8M, 4.8M and 4.9M successfully extracted barcodes (for RTPCR3, 4, 5, respectively), which converged into 300,943; 385,927 and 408,338 variants (reduced to 1,398; 9,561 and 5,457 unique clustered variants, respectively). These values are in agreement with the expected number of variants considering the number of transducing units applied to the different samples (except in the RTPCR5). In addition, these results also correlate with the LM-PCR results shown in Section 4.4.1, which reinforces the robustness of this screening method at DNA and RNA level.

As shown in Figure 4.18, the complexity of the retrieved barcode remained high, with no variants representing more than 5% of the whole barcode population. These results contrast with those obtained from transduced HEK 293 SA RIX (MOI of 1) where a few variants were retrieved in a relatively high proportion suggesting that a few clones might have overgrown the population. This might be attributed to the extended times of culture the transduced HEK 293 SA RIX were kept for, which contrasts with HEK 293 6E cells, harvested for NGS a week post-transduction. The level of complexity and throughput of the barcoded library retrieved at a transcriptional level (10^4 barcodes variants) is compatible in order of magnitude with that of barcodes retrieved by integration site recovery techniques (10^4 unambiguous IS). An alternative method to assess clonal expansion would be to use a vector library consisting of different fluorescent markers in order to track the expansion of individual clones although the throughput of this approach is limited.

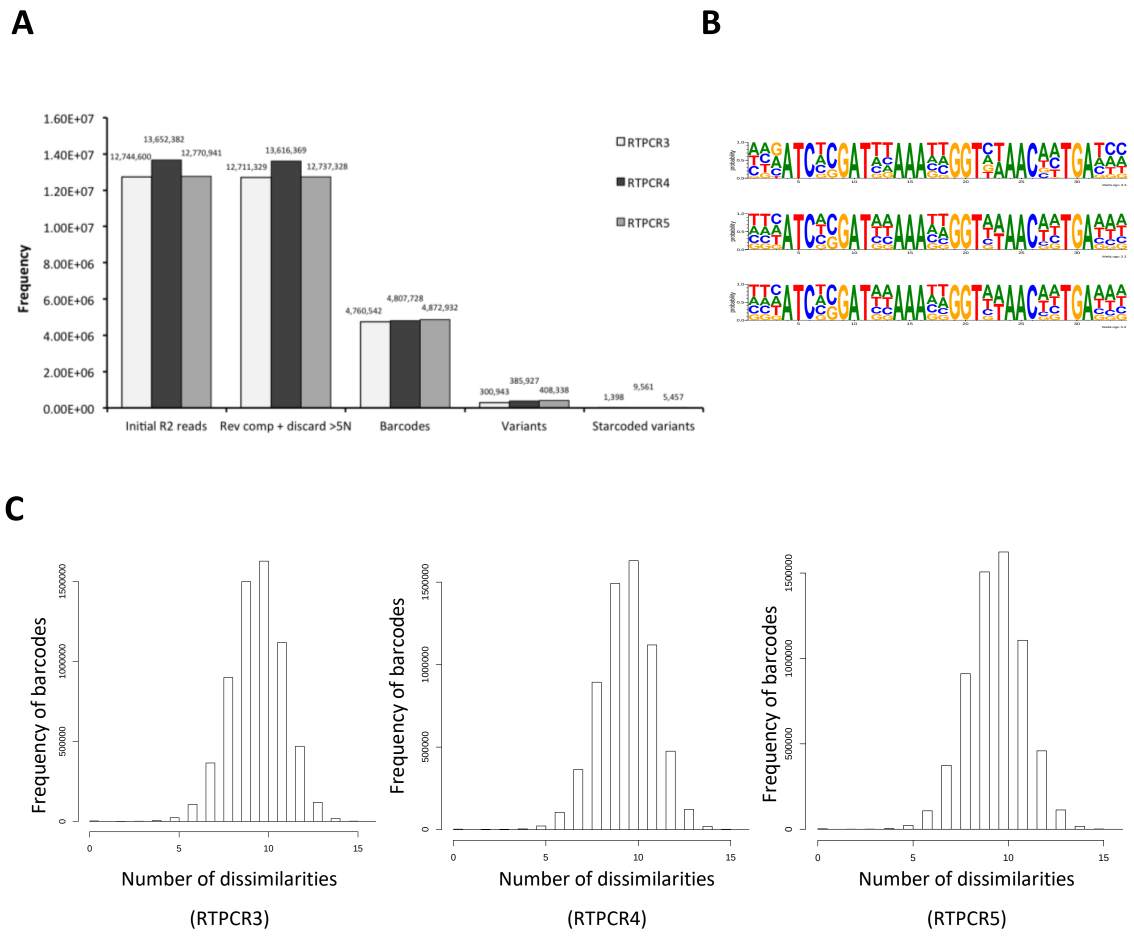


Figure 4.18. Characteristics of the barcode counts retrieved from reverse transcribed cellular RNA analysed by next-generation sequencing.

(A) Number of R2 reads obtained by high-throughput sequencing using a HiSeq 2500 100bp paired-end (PE) strategy, successfully merged reads with less than 5N, extracted barcodes, unique barcode variants and clustered variants (using Starcode) present in the three samples analysed. (B) Pictogram of relative frequencies of nucleotides in sequenced barcodes Resource available at: weblogo.berkeley.edu.logo.cgi[704]. Adenine (A) is shown in green; cytosine (C) in blue; thymidine (T) in blue and Guanine (G) in yellow. Pictograms from top to bottom corresponding to RTPCR6E5, RTPCR6E4, RTPCR6E3; (C) Barcode comparison plot showing the number of nucleotide differences between all the sequenced barcodes of clustered RTPCR samples.

Clustering correction was applied to extracted barcodes RTPCR3 and RTPCR4 different Leveinshtein distances (-d1 to -d5) as in Figure 4.7; RTPCR5 was discarded because its integration site counterpart file did not contain the expected number of IS. In both cases, top expressers remained in the same positions of the ranking and unclustered as separate nodes (or centroids), indicating they were significantly different below a threshold of –d=3, corresponding with the results seen in Figure 4.18C.

All four nucleotides in Starcoded barcodes extracted from RNA counts were equally represented in variable positions of the barcodes in all samples (Figure 4.18B). In conjunction with these results, the cumulative density demonstrates also a balanced profile. The profile and peak number of dissimilarities (at 11 nucleotides) (Figure 4.18C) is maintained when the barcodes from RTPCR3, RTPCR4 and RTPCR5 samples are clustered confirming their variant proportions (data not shown).

4.5.2 Correlation of retrieved integration sites - relative abundance via barcode

In order to know the number of barcodes in the RNA-Seq analysis that could be mapped to LM-PCR reads, only one barcode variant per integration site should have been retrieved. However, from the 8,261 IS retrieved in the 6E4 dataset, the multiple barcode variants obtained per integration site make the total number of theoretical IS-variant combination rises up to 142,256. Of these, 85,157 would find correspondence in the RNA-Seq dataset. This association is not meaningful, as the element used to correlate both datasets (the barcode variants) is duplicated. Therefore, there are many barcode variants per integration site and also some of the variants are common in between different integration sites.

Following the same procedure used with HEK 293 SA RIX (Section 4.3.5), a list of top 15 candidate positions with relatively high barcode expression and unambiguous correspondence in the host cell genome was generated for genome editing candidate selection in HEK 293 6E 104, 103 datasets (Table 4.5).

Table 4.5. Correlation of lentiviral integration sites with their derived RNA counts via barcode.

Sample	Top #	#RNA counts	Barcode variant sequence	Number of containing this variant	Signal? **	#LM-PCR reads	Genomic position	Chr	Gene
6E4	1	24,032	TTC-TC-CA-AT-TA-CG-ACT	254	<10-fold		***		
	2	**20,017**	**AGC-TC-AA-AT-TT-AC-CAC**	**66**	**160**	**802**	**168,010,733**	**3**	**EGFEM1P**
	3	19,928	GGT-AC-CA-AT-TA-CG-GTA	199	<10-fold		***		
	4	19,907	TCC-CC-CA-GT-GA-AC-TGT	96	255	1,529	80,591,158	17	WDR45B
	5	18,778	CTT-TG-AC-GC-CA-AG-ATA	58	<10-fold		***		
	6	17,028	AGG-CC-GT-AA-TA-TA-AAA	19	<10-fold		***		
	7	16,624	CAA-TC-AG-AA-AA-AT-CAT	2	<10-fold		***		
	8	16,001	TAA-TC-TT-AT-TT-TA-AAG	14	<10-fold		***		
	9	**15,034**	**TGC-CG-CG-TG-TT-TA-TAC**	**185**	**14**	**361**	**107,887,987**	**11**	**CUL5**
	10	14,242	TTC-TG-TA-TT-GT-GT-GCA	106	<10-fold		***		
	11	13,944	AAT-AC-CA-AA-TT-AT-GTT	193	<10-fold		***		
	12	**12,605**	**GAC-AC-CA-CT-TT-TA-TAT**	**214**	**11**	**4,491**	**15,409,560**	**21**	**none**
	13	12,556	ATA-AC-GC-TT-TA-TA-AAA	257	24	573	37,722,505	14	MIPOL1
	14	12,330	TGA-GC-GA-TT-TA-AT-TTG	74	15	511	151,738,902	4	LRBA
	15	12,311	TTA-TC-TA-GA-AA-GA-CTG	298	<10-fold		***		
6E3	1	72,976	ACA-TG-TA-CT-AA-GT-CTA	none	NA				
	2	59,439	CTT-GG-AG-AT-AT-TG-TTT	11	<10-fold		***		
	3	47,243	GCA-CC-AA-TA-AT-AA-TAT	none	NA				
	4	45,842	ATC-TC-AT-GT-CT-TC-TAC	5	<10-fold		***		
	5	44,350	TGT-TC-AT-GT-AT-TT-GGT	none	NA				
	6	42,378	TCT-TG-TT-AT-TT-AT-GGT	none	NA				
	7	38,362	AAT-TG-AA-TT-AT-AA-ACA	none	NA				
	8	37,101	TTT-AC-TT-AA-GA-TT-GTT	25	<10-fold		***		
	9	35,877	ATG-GC-TT-CA-CT-GC-TTA	none	NA				
	10	35,610	ATT-AC-TG-GA-GA-TG-ATG	none	NA				
	11	34,181	AGC-TC-AA-AT-TT-AC-CAC	none	NA				
	12	32,200	GAA-CG-AA-GC-AA-CC-ATT	none	NA				
	13	31,784	ATT-TG-GA-TT-TT-TC-AAG	33	<10-fold		***		
	14	31,683	TCC-CC-CA-AA-GA-AT-GAC	none	NA				
	15	28,180	CTA-CG-TT-CA-AA-TA-GCA	none	NA				

Column 5, number of IS containing this barcode variant, number of genomic positions correlated to a single barcode variant; column 6, signal (**), in the case of >1 IS per barcode variant, ratio between the 2 first genomic positions with more reads retrieved by LM-PCR and NGS. Candidate positions selected for genome editing of lentiviral transfer vector are highlighted in bold. Highlighted barcode sequences are common in both samples. (***) Integration sites with a signal below the threshold (10-fold) were not mapped.

Again, as shown in Table 4.5 column 5, there is more than one genomic locus retrieved per barcode variant. In this particular example, 66 different loci respond to the variant AGC-TC-AA-AT-TT-AC-CAC. However, the position with ID 268 (chr3:168,010,733) displays a relatively high signal considering the number of times it was retrieved by LM-PCR. The term signal refers to a relatively abundant number of reads with the same end position compared to other positions that share the same barcode variant. As in HEK 293 SA RIX, if the ratio between the first and the second genomic position with more LM-PCR reads is greater than an order of magnitude of difference, it was considered for correlation with the barcoded RNA counts. Otherwise, positions were considered background noise originated due to cross-priming during PCR and were subsequently discarded.

Alternatively to a cross-comparison of datasets, the top most expressed barcodes obtained by RNA-Seq were searched for a barcode counterpart in the LM-PCR dataset. This approach is suboptimal as a result duplicity in barcode assignation encountered in the LM-PCR. This limits the criteria to chose candidates for targeted integration and in order to make a decision, only integration sites with a high number of reads retrieved were considered for RNA-Seq barcode correlation. The threshold (or strength of signal) was defined as the ratio number of LM-PCR reads obtained between the IS-most_abundant_variant and the IS-second_most_abundant_variant. Therefore, no integration sites were unambiguously retrieved as candidates with a higher transcription rate as a result of the integration of the lentiviral vector.

No genomic positions were unambiguously retrieved from the 6E 10^3 dataset. RNA counts per variant are higher in 6E 10^3 dataset because fewer positions were amplified and thus the throughput is divided/shared between fewer reverse transcribed mRNA molecules. Only one barcode combination was found in both datasets minimising the effect of any potential contamination. This contrasts with the results obtained in the SA RIX analysis Table 4.3 where half of the barcodes were also found in other datasets.

Three candidate positions with relatively high expression (barcoded RNA counts) were chosen from the HEK 293 6E 104 dataset for site-specific integration of a lentiviral transfer vector using the CRISPR-Cas9 system.

Loci 1 - Chr3:168,010,733	(EGFEM1P pseudogene)
Loci 2 - Chr11:107,887,987	(in the first CUL5 intron)
Loci 3 - Chr21:15,409,560	(intergenic position)

Such positions were chosen from the HEK 293 6E 10^4 dataset because more univocal/unambiguous candidates were in the first 15 positions of RNA-Seq variant counts compared to the HEK 293 6E 10^3 sample. In the case where positions were found in both datasets, candidates would still have been chosen from 6E 10^4 dataset, since it is a 10-fold larger pool of IS. The HEK293 6E LM-PCR 10^5 dataset did not show the expected number of integration sites providing the transducing units that were applied to the cells and thus was discarded.

EGFEM1P was chosen as a first candidate because due to the abundance of RT-PCR reads. *EGFEM1P* stands for EGF-like and EMI domain containing 1 and it is annotated as a pseudogene particularly expressed in the pituitary. Pseudogenes are vestigial, non-essential fragments of genes that have lost their ability to code for protein as a result of multiple mutations. However, pseudogenes can undergo transcription of non-coding RNA driven by a nearby promoter[761]. Their function is not totally understood although they might play a regulatory role similarly to other non-coding RNA[762].

The second candidate position (chr14:107,887,987) is located in cullin5 (*CUL5*). *CUL5* inhibits cellular proliferation, potentially through its involvement in the SOCS/ BC-box/ eloBC/ cul5/ RING E3 ligase complex, which functions as part of the ubiquitin system for protein degradation[763]. Interestingly, CUL5 protein is also reported in the literature to interact with viral trans-activating regulatory protein *tat* and viral accessory protein *vif* (although they are not present in a 3rd generation lentiviral vector). It might also play a role in the reelin signaling cascade[764]. In terms of tissue distribution, studies have shown that cullin5 is highly expressed in heart and skeletal tissue, and is specifically expressed in vascular endothelium and renal collecting tubules. The renal origin of the HEK293 cells could possibly explain observed expression levels[765].

The third position (chr21:15,409,560) was chosen for two reasons. Firstly, the LM-PCR signal is relatively high compared to other candidates; this integration site was clearly distinguishable from the background noise of genomic positions with the same barcode variant. Secondly, because this position does not correspond to a RefSeq gene; the ankyrin repeat domain 20 family member A11 (*ANKRD20A11P*), a non-coding pseudogene RNA and the human phospholipase I (*LIPI*) are located more than 50kb upstream and downstream, respectively. Therefore, our intention was to assess titers resulting from targeting a genomic position that does not disrupt gene expression at all.

Although *WDR45* has a higher signal ratio than any of the other candidates, this genomic position was not chosen for genome editing because its function is not as critical in the biology of the cell as CUL5 and was found to be not as accurately annotated. *WDR45* encodes a member of the WD40 repeat protein family and participates in cell progression, regulation and apoptosis (NCBI Gene ID: 11152).

The following two positions also present in Table 4.5 were not considered because presented a lower number of corresponding RNA variants. Time an resources were the main limitations to keep the number of candidate positions to test for genome editing to 3. However, the biological function of these loci is described below for interest.

MIPOL1 stands for mirror-image polydactyly 1 and it is a protein-coding gene whose truncation is associated with the above mentioned syndrome (also known as Laurin-Sandrow Syndrome)[766].

LRBA is LPS (lipopolysaccharide) responsive beige-like anchor protein and plays a role in secretion of vesicles containing immune effector molecules.

4.5 Summary of results and concluding remarks

- An integration site analysis technique (ligation-mediated PCR) was optimized for detection of vector-chromosome junctions. Ligation-mediated PCR performed at the 5' end of the junction allows for longer reads and avoids NGS amplification of the internal band (commonly used as a reaction control).
- A bioinformatics pipeline was built and optimised to isolate integration sites and their associated barcodes.
- Integration site preferences of barcoded lentiviral vectors confirmed their predisposition for transcription units, no TSS and allowed for weak consensus sequence
- The amount and complexity of the barcodes retrieved by integration site analysis (LM-PCR) and expression analysis (RNA-Seq) were equivalent and sufficient to simultaneously screen more than 10^4 sites.
- Three candidate loci unequivocally reporting a high number of barcoded RNA counts were chosen for subsequent targeted integration of a lentiviral transfer vector using CRISPR-Cas9 genome editing technology.

In this chapter, a variation of the LM-PCR was performed to successfully retrieve lentiviral barcoded junctions. The variation comprises amplification of 5' junctions (instead of the 3' junction) with a vector primer annealing immediately downstream of the 5'LTR (and thus not amplifying the 3'LTR barcode). This impedes amplification of an internal control band typically used in conventional LM-PCR as a control of the technique. In addition, amplification of the barcode from the 5'LTR U3 allows for the recovery of longer and better quality reads. Following this procedure, integration site analysis was initially performed on HEK 293 SA RIX cells that had been transduced at a low MOI and sorted into different populations based on GFP intensity. An integration site analysis pipeline was built in order to recover integration sites with high identity rates to the human genome and perfect matching with known flanking sequences (LTR and linker if present).

Out of the 10^4-10^5 positions recovered, only 10^2-10^3 distinct sites were unique. Such a reduction in the complexity of the IS-barcode system may be explained by (i) the effects of clonal dynamics after prolonged periods of culture (with the aim of preventing silencing of gene expression) and (ii) the generation of chimeric PCR products. As a result of these phenomenons, (i) fewer clones (and their respective barcodes) overgrew the population and (ii) different loci associated to the same barcode, which precluded any concluding remarks. While this measure aimed for selection of stably expressing sites, population dynamics and several rounds of passaging are factors that contribute to diminish initial entropy introduced by the barcode vector library. Nevertheless, lentiviral vectors naturally select for stably expressing sites and amplification and the screening of genome-edited clones for several weeks of growth in a later stage can anticipate and prevent eventual silencing. Additionally, balanced and low relative abundance of variants in all barcode formats/supports (oligonucleotide, plasmid but mainly vector library level prior to transduction) was demonstrated to contribute to bias minimisation. Finally, genomic positions that were extracted from HEK 293 SA RIX, were mostly transcription factors related with myogenic function. Nevertheless, their annotation did not reveal any link with the biological precedence or origin of HEK 293 cells unlike candidates selected from HEK 293 6E.

A second attempt using HEK 293 6E (which have a better yield in vector production) was then performed avoiding prolonged periods of culture. The choice of restriction sites was also optimised to avoid any bias and maximise genome accessibility. In this experiment, 1,245 and 8,261 unique integration sites were recovered by next-generation sequencing upon transduction of 1,000 and 10,000 cells with an MOI of 1. The number of IS retrieved in this validation experiment supposed a considerable increase compared to the size of previous genome-wide massive parallel DNA sequencing experiments such as Schroder *et al.*, in 2002 (524 sites)[249]; Cattoglio *et al.* (849 sites)[767] and Mitchell *et al.*, in 2004 (407-528 sites)[231] but below Wang *et al.*, in 2007 (40,596 sites)[759]. Lentiviral integration was found predominantly in gene rich areas and confirmed not obeying the integration into larger chromosomes (C-paradox). This agrees with the currently accepted (though incomplete) model for retroviral site selection, in which chromatin openness (if not excessively) of actively transcribed genes located in regions proximal to the nuclear envelope promote accessibility to the viral/vector particles, that have evolved to target PolII transcribed genes that enhance transcription upon integration. Integration was also found throughout the length of the transcription unit and not only close to the TSS. A weak consensus in the primary sequence (TDG(G)TAAC) was found in the IS retrieved from HEK 293 6E cells. Taken together, barcoded lentiviral vector libraries completely follow the integration features and preferences described in the literature.

However, a few technical complications related to the nature of this method have been identified and may skew the results and anticipate potential solutions:

- The first amplification steps of PCR are characterised by stochastic fluctuations in priming, which can lead to a disproportion in frequency of (barcode) template and the amplified product, commonly known as PCR jackpotting. A potential future improvement could be the implementation of Unique Molecular Identifiers, a second set of DNA tags (with a large sample space and sufficient editing distance) that provide information about PCR dynamics during the first two rounds of amplification[768]. This way a barcode variant that contains the same tag are likely to be formed as a result of PCR mutation or a sequencing error and can be easily removed from the pool of real barcodes.

- Secondly, unlike linear amplification of barcode molecules in plasmids, the size of the genome of mammalian host cell lines together with the difficulty of extracting barcodes require several rounds of amplification, which introduces differences in PCR bias. Isolation or enrichment of barcoded sequences could alternatively be achieved by restriction digestion and size exclusion. Other methods based on capture would involve DNA or RNA "hooks"[769–771].

- A different source of bias could be introduced by PCR purification methods. Most commonly used methods for purification of DNA fragments (silica columns or

AMPureXP beads) enable recovery of fragments greater than 100bp, excluding short junctions. That could explain the difference between the number of transducing units applied to 6E cells and the amount of sites recovered. Short junctions could have flowed through the purification columns reducing the number of IS retrieved.

In parallel to the determination of vector integration sites, retrieval of barcodes by RNA-Seq can be achieved by biotinylated primer selection/pooling and whole transcriptome analysis (WTA). However, the retrieval efficiency ($4x10^6$ barcodes for WTA vs 10^5 biotinylated primer) of both methods could not be compared due to the variation on the experimental design. Read lengths and fragmentation times are very important factors for the latter strategy. In any case, the throughput values achieved with any of these two strategies can complement those obtained by LM-PCR, indicating that the lack of throughput is not a limitation in the rationale of this project. In addition, 1,398 and 9,561 different barcode variants were retrieved for 10^3 and 10^4 samples, respectively, which confirms the values obtained by LM-PCR (1,245 and 8,261, respectively).

However, in order to study the impact of transcriptional activity of HEK 293 on integration preferences, whole transcriptome (or microarray) analysis could have been performed on untransduced host cells. While in general terms, integration is favoured in actively transcribed genes, some studies show that the highest expressed genes report low levels of integration supporting what was suggested by Weidhaas *et al.*, in 2000[254]. Schroder revealed that the correlation between integration and transcriptional activity already existing in non-transduced cells becomes stronger in transduced cells[249]. This analysis would have contributed to confirm lentiviral integration active sites and also to understand the extent to which barcode counts observed in genes at a local level correlate with global transcriptional activity assigned to a gene. This could help provide insight to establish a background noise threshold or a normalisation for basal level of activity associated to a gene. However, another way to discriminate between background noise and real low frequency barcodes could be to evaluate the consistency of variant representation over time. Population dynamics should be evaluated to examine the model they follow (stochastic, stable, etc.). Once identified their trend in a particular cell line (exposed to stimuli different from hematopoietic stem cells, where this experiments are typically performed) their general contribution could be assessed and that could allow exclusion of outliers or artefacts generated at any stage of the process.

An important aspect to consider is the fact that barcoded RNA transcripts will be mainly driven by the internal lentiviral vector promoter. Once integrated the 5'LTR contains the SIN U3 and thus the enhancer/enhancer promoter activity is supressed. The influence of the chromatin environment and neighbouring genes to the barcode expression is questionable and limited given the position of the barcode in the 3'LTR.

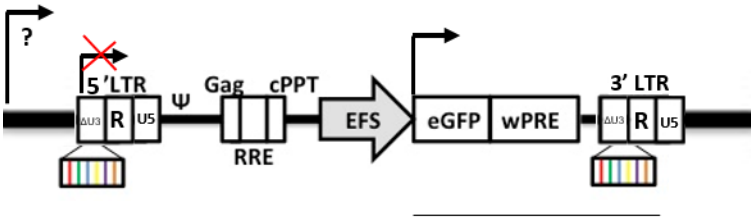


Figure 4.19. Schematics illustrating the role of different hypothetical promoter regions on the expression of barcode

The viral transcript containing the barcode is represented by a black line below the integrated vector. Promoters are represented as angled arrows. LTR, HIV-1 long terminal repeat; Ψ, HIV-1 RNA packaging signal; SIN, self-inactivating (U3-deleted) HIV-1 long terminal repeat; cPPT, central polypurine tract; Gag, HIV Gag gene; RRE, Rev responsive element; eGFP, enhanced green fluorescent protein; WPRE, woodchuck posttranscriptional regulatory element;

In the event of evaluating the promoter that drives the expression of the whole transfer vector, this should be the same as the internal promoter; otherwise the difference between promoters (in this study EFS internal and RSV driving the transfer vector genome) can lead to inconclusive results.

In this study, the barcode system is postulated as an alternative selection method with a screening capacity of 10^4 integration sites at a bp precision and such simplicity represents a major improvement compared to high-throughput systems, which need costly antibodies or sophisticated automated closed systems. Although the libraries presented in this project could theoretically handle a few million cells with their >10^5 barcode variants and the majority of cells will receive unique barcodes, with larger transductions, the same barcode could be delivered to more than one cell more frequently. In addition, successful IS retrieval has not been validated with 10^5 cells. Other potential concerns like the loss of complexity in the number of dissimilarities were discarded as 11bp harbour enough variants to meet any throughput need. Elimination of overlapping integration sites attributable to contamination between samples represents <4% and is not a limitation either.

The deletion of 400bp containing termination enhancer motifs in self-inactivating (SIN) retroviral vectors enhanced the leakiness of 3'LTR transcriptional termination (up to MLV levels), which results in read-through of vector transcripts into host genomic content[772,773]. This could be utilised to capture genomic position and expression derived from each integration event directly from a single RNA molecule avoiding any restriction or PCR bias. RNA molecules containing vector-host junctions could be selectively primed using a biotinylated primer that allows amplification of sequences downstream of the 3'LTR R region utilising a polyA signal instead of the conventional oligodT priming to synthesise the first strand of cDNA upstream of the polyA tail.

No replicates were run in next-generation sequencing experiments. Due to the number of conditions to be tested in a single run and the sequence capacity of a flow cell, the presence of multiple replicates per condition would have compromised the sequencing depth. Introduction of technical replicates is critical in next-generation sequencing of rare single nucleotide variants[774]. In the case of small fragments of DNA with multiple variations, depth within a sample becomes a key component, especially if expected sample complexity raises up to 10^5 distinct variants. Biological replicates imply separate experiments and sequencing runs, which suppose a limitation due to their elevated costs. The similarity of the results achieved between the HEK 293 6E 10^3, 10^4 and 10^5 conditions (and also pSYNT and vSYNT 11 and 49 libraries in Chapter 3) indicates significant differences would not be expected among technical replicates.

In the next chapter, specific integration of a reporter gene into the positions discovered in this chapter will be tested in order to prove the hypothesis proposed in this study.

Chapter 5

RESULTS: CRISPR-Cas9 knock-in of donor constructs in identified loci

5.1 Introduction

The traditional strategy for recombinant protein production typically involves the delivery of a gene of interest into the host cell lines followed by stable integration and selection and screening of multiple clones[464]. The lack of control of random integration of transfected plasmids often leads to phenotypic heterogeneity, also termed position effect variation[775]. This work presents an alternative, non-random way to drive integration and reduce screening and selection timelines. Therefore, a controlled means of DNA insertion is required.

CRISPR-Cas9 genome editing technology is quite recent and has not been fully explored for biopharmaceutical applications. In contrast, Lee *et al.*, employed genome editing in CHO cells for protein production[589]. Site–specific integration of expression cassettes for protein production had only been seen in antibody production. In packaging cell line development, Sanber *et al.*, used the integration preferences of MLV-derived vectors to stably drive the expression from highly transcribing sites, although integration was not effectively site-specific. In their study, they tagged such sites with MLV and posteriorly introduced *gag-pol* genes via RMCE[381]. A similar strategy was used by Carrondo *et al*., to explore the impact in stoichiometry of Gibbon ape leukaemia virus *env* gene in combination with *gag-pol* on titers in the context of a packaging cell line[431]. To our knowledge, this is the first study in which CRISPR-Cas9 technology has been used to rationally integrate a lentiviral vector component with the objective of generating a packaging cell line. In addition, none of these studies rationally targeted integration to a specific locus based on their quantified expression.

Several authors showed the expression of the transfer vector is the limiting factor for lentiviral vector production[401,414,463]. For this reason, the transfer vector constitutes the majority of the plasmid DNA in transient transfections. In order to simplify the project and demonstrate proof of concept in its limited timeline, a lentiviral transfer vector was introduced in the three different positions discovered in Chapter 4 instead of separately integrating *vsv-g*, *gag-pol* and *rev*. The remaining viral genes were then complemented in *trans* by transient transfection in order to assess titers. This constitutes the inverse approach to what is typically done (to keep the transfer vector flexible/modular) but allows effective screening and titration of functional titers by GFP fluorescence quantification. The fundamentals for this decision lie on the identification of the expression of the viral vector genome as one of the main limitations for a vector production in packaging cell lines[420,430,776]. In addition, transient expression of the remaining viral genes eliminates the consequences of *gag-pol* expression when not balanced with that of the envelope protein[431,776].

The three different positions discovered in Chapter 4 that were tested in Chapter 5. Chr3:168,010,733 plus strand (EGFEM1P pseudogene), Chr11:107,887,987 plus strand (located in the first intron of CUL5) and Chr21:15,409,560 minus strand (located in an intergenic position) show no disruption of the gene product. In this chapter, targeting efficiencies and efficacies of the CRISPR-Cas9 strategy were assessed and packaging cell lines containing transfer vectors integrated in high transcribing positions were produced and evaluated.

5.2 Aims

The specific aims of this chapter were:

- To demonstrate site-specific integration of a lentiviral transfer vector cassette >5kb fragment of DNA into a control position described in the literature and high transcribing positions discovered in Chapter 4 using genome editing techniques (CRISPR-Cas9).

- To assess the efficiency and rates of recombination-mediated gene addition of the CRISPR-Cas9 genome editing technique.

- To evaluate the titer of virus from a producer cell line containing a lentiviral backbone expressing GFP.

5.3 Preparation of the CRISPR-Cas9 plasmids

5.3.1 Cloning of donor construct (pRRL 2HA SIN cPPT EFS eGFP WPRE Zeo BFP)

In order to explore the potential of CRISPR-Cas9 as a genome-editing tool, we tested the integration of a reporter system into a genomic position previously reported in the literature. Insertion (knock-in) of a 1kb of foreign DNA into the EMX1 locus (Chr2: 73160998 - 73160999) has been previously reported by Cong *et al.*, using 800bp homology regions[777]. Cong *et al.*, (2013) used CRISPR-Cas9 technology to site-specifically integrate a 1kb fragment into the EMX1 locus. Of relevance to this work, the size of our donor fragment is >5kb, which requires the optimization of delivery parameters (length of homology arms, number of cells per transfection, amount/ratio of plasmid DNA, transfection method, selection).

Two components are necessary in order to specifically modify a sequence in a particular position of the genome via CRISPR-Cas9: (i) a guiding RNA containing the insertion or deletion at the loci of interest and (ii) a protein which catalyses the excision of a DNA strand (nuclease for double-strand break –DSB- or nickase for single strand cuts). If site-specific integration of a stretch of DNA is desired, then a donor construct with homology arms is required to be added by homologous recombination.

A donor plasmid containing homology arms flanking a lentiviral transfer vector containing a reporter gene was constructed (Figure 5.1). The lentiviral transfer vector comprised a 3rd generation lentiviral RRL backbone with the EFS promoter driving the expression of eGFP. The woodchuck hepatitis virus post-transcription regulatory element (WPRE) is located downstream of the eGFP reporter gene. The donor plasmid also contains an antibiotic resistance gene downstream of the transfer vector and within the homology arms, whose function is to act as a selectable marker in the presence of zeocin. The zeocin cassette consists of the resistance gene under the control of the SV40early promoter and upstream of the SV40 polyA signal from pcDNA4 TO (Figure 5.1B). The use of zeocin resistance as a selectable marker has reported higher and more stable GFP intensity in cell pools compared to other antibiotics[778].

In order to assemble the donor plasmid, the zeocin cassette was cloned downstream the 3'LTR of the original lentiviral transfer vector (Figure 5.1A). The 1,145bp zeocin cassette was amplified from pcDNA 3.1 Zeo (+) (Invitrogen, V860-20, Appendix A) with primers (NheI-SV40P-Zeo-fwd 5'-AGGAT*GCTAGC* gaatgtgtgtcagttagggtg-3' and Zeo-MCS-NheI-rev 5'-ATCGC*GCTAGC*ACTAGTAC GCGTGGTCACCctagaggtcgacggtatacag-3'; *NheI* sites indicated in *italics;* overlapping base pairs in lowercase) including *NheI* sites and the PCR product was digested with *NheI*, column-purified prior to ligation of compatible ends to the *AvrII* site present in the lentiviral backbone (pRRL SIN cPPT EFS eGFP WPRE) giving rise to pRRL SIN cPPT EFS eGFP WPRE Zeo (Figure 5.1B).

Next, the lentiviral backbone containing the zeocin resistance downstream the 3´LTR (pRRL SIN cPPT EFS eGFP WPRE Zeo) was surrounded by several restriction sites in order to provide multiple options when incorporating the fragment into further constructs containing homology regions with potential conflicting sites. To achieve this, the lentiviral backbone + zeocin resistance cassette was amplified with primers containing *MluI*, *BstBI*, *AscI* and *PacI*, *MreI*, *AsiSI*, *NheI* restriction sites on the 5' and 3' side, respectively, and topo-cloned into a pCR4 TOPO TA backbone (Figure 5.1B).

MluI, *BstBI*, *AscI* T3promoter fwd primer sequence (MlBAT3_A1)
5'-AGCTA*ACGCGT*ATATA*TTCGAA*CGAAT*GGCGCGCC*aattaaccctcactaaaggg-3

T7promoter rev primer sequence with *PacI*, *MreI*, *AsiSI*, *NheI* (PMrAsINT7_A2)
5'-TGATT*TTAATTAA*ATTAT*GCGATCGC*ATTG*CGCCGGCG*AAG*GCTAGC*taatacgac
tcactatagg-3'

Restriction sites indicated in *italics;* overlapping base pairs in lowercase.

Subsequently, 4 plasmids (pMS-RQ HA-MCS-HA2-BFP DCAF6, pMA-RQ HA-MCS-HA2-BFP CUL5, pMK-RQ HA-MCS-HA2-BFP 'inter' and pMS-RQ HA-MCS-HA2 EMX1) containing a 3,398bp fragment consisting of a multicloning site with the aforementioned restriction sites flanked by 800bp homology arms were ordered from GeneArt (Figure 5.1C). Homology sequences for the 3 candidate loci are:

EGFEM1P (Chr3 Left:168009927–168010727; Right 168010731-168011531)

CUL5 (Chr11 Left:107887184-107887984; Right 107887985 - 107888785)

intergenic (Chr21 Left 15409567 -15410366; Right 15408767-15409566)

A blue fluorescent marker (strongly enhanced blue fluorescent protein, seBFP, under the control of the SV40 promoter) was placed outside the homology arms (downstream the 3' arm) to monitor integration not occurring via homologous recombination. In the event of a double recombination, the third generation transfer vector harboured in the donor construct together with the zeocin marker would be delivered to the host cell genome and cells would be exclusively GFP positive; otherwise, cells would be double positive for green and blue florescence. The choice of the BFP fluorochrome was based on excitation/emission wavelengths, sufficiently distinct from that of GFP (present in the transfer vector) and RFP (present in the Cas9/sgRNA vector for EGFEM1P, CUL5 and 'intergenic').

Assembly of pMK-RQ HA-MCS-HA2-BFP ‘inter’ could not be achieved by GeneArt after more than 20 cloning attempts and was thus dropped from our list of candidate positions due to limitations in our timelines. Both, TOPO TA vector backbone pCR4 containing pRRL SIN cPPT EFS eGFP WPRE Zeo flanked by *NheI* sites (insert) and GeneArt synthesised plasmids (backbone) containing a multicloning site flanked by the candidates’ respective homology arms (pMS-RQ HA-MCS-HA2-BFP EGFEM1P, pMA-RQ HA-MCS-HA2-BFP CUL5, and pMS-RQ HA-MCS-HA2 EMX1) were digested with *NheI* and *BstBI* and ligated together resulting in the final donor constructs (Figure 5.1C). In all cases, DNA isolated from bacterial clones was tested by restriction digest analysis and confirmed by Sanger sequencing (data not shown).

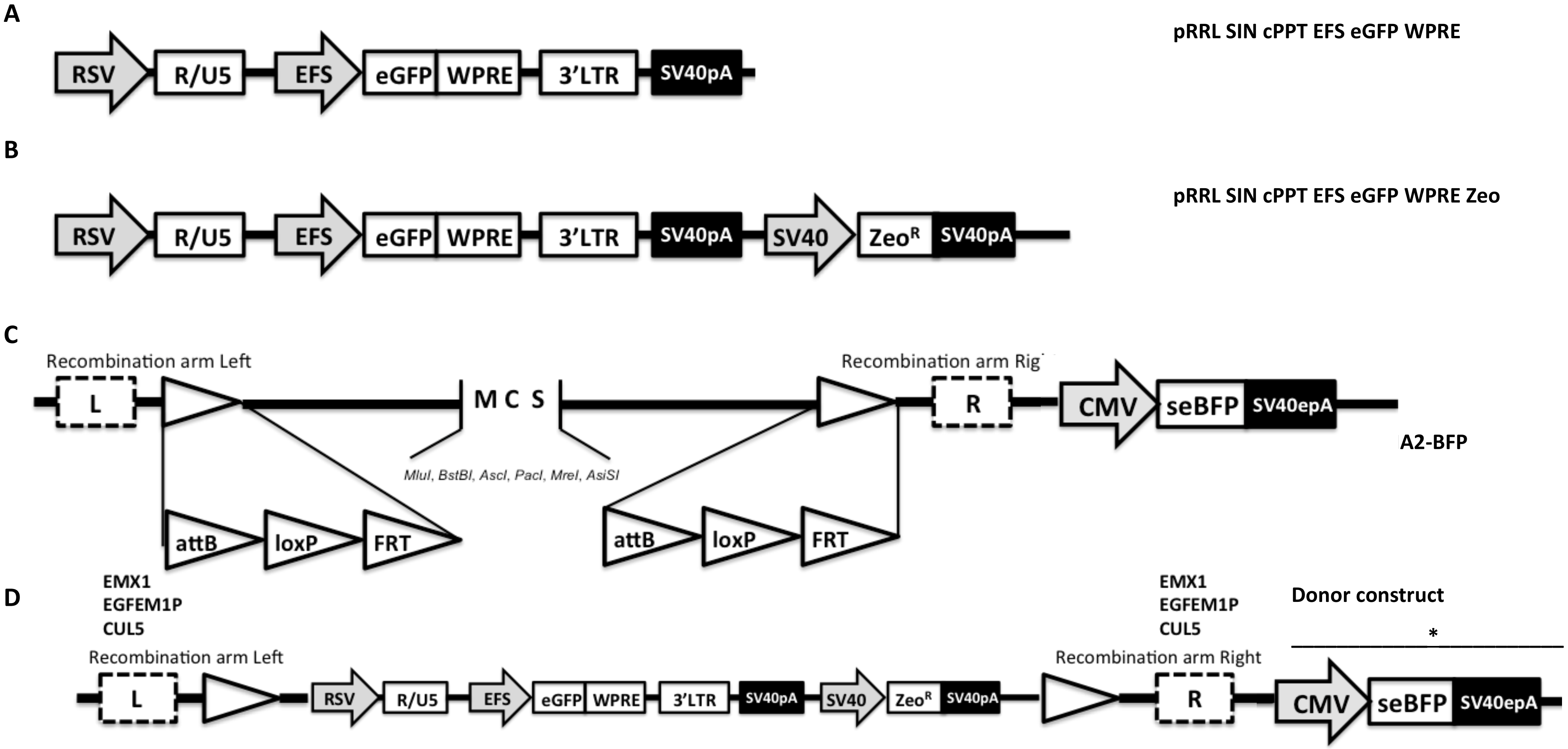


Figure 5.1. Schematics of donor construct cloning procedure.

(A) Original third generation pRRL SIN cPPT EFS eGFP WPRE transfer vector plasmid. (B) pRRL SIN cPPT EFS eGFP WPRE with zeocin resistance gene downstream the lentiviral transfer vector under the control the SV40 promoter and followed by a SV40polyA signal. (C) Backbone ordered from GeneArt containing a multicloning site (MCS) flanked by recombinase recognition sites (attB, loxP and FRT) and homology arms (right, R; left, L) upstream of a blue fluorescent protein gene under the control of the CMV promoter and upstream the SV40 early polyA signal. Distances are not to scale (D) Donor construct resulting from the cloning of pRRL SIN cPPT EFS eGFP WPRE Zeo into the pMS/K/A-RQ HA-MCS-HA2-BFP backbone containing homology arms and the BFP cassette. RSV Rous sarcoma virus; LTR, long terminal repeat; eGFP, enhanced green fluorescent protein; WPRE, woodchuck posttranscriptional regulatory element; CMV, cytomegalovirus; SV40pA, simian virus 40 polyA signal; seBFP, strongly enhanced blue fluorescent protein. (*) CMV-seBFP-SV40pA not present in the EMX1 donor.

5.3.2 Description of Cas9 and sgRNA plasmids

Two separate plasmids containing the Cas9 nuclease under the control of the CMV promoter and the U6 promoter driving the expression of 20bp EMX-1 sgRNA sequence (5'-GAGTCCGAGCAGAAGAAGAA-3') prior to the *S. pyogenes* protospacer adjacent motif (PAM, NGG in the genomic sequence in *S. pyogenes* CRISPR system) described by Cong *et al*., were obtained from Sigma-Aldrich (Figure 5.2A and B). The U6 promoter is a RNApol III promoter that allows ubiquitous expression of the sgRNA in human cells and specific initiation and termination of transcription[779]. However, it requires a guanine immediately before the first sgRNA nucleotide. In the case of the three candidate genomic positions, the sgRNA sequences containing the PAM motif were chosen using the Zhang laboratory (http://crispr.mit.edu/) online tool according to the following criteria: minimum risk of off-target effect, minimum distance from a PAM motif to the candidate position selected in Chapter 4.

A single vector including the Cas9 gene and the sgRNA was used instead of two separate plasmids for the knock-in into the 3 candidate positions (Figure 5.2C). The sgRNA sequences were 5'-TTAATGCTTATTTATTTTGT-3', 5'-TACCTGGGGGTGGTGGTGTA-3' and 5'-TACCTTCTTCCCTACAGGTC-3' for EMX1, EGFEM1P, CUL5 and 'intergenic', respectively. . Transfection using a 2-plasmid system (donor + sgRNA/Cas9) maximises the chances of successful cellular delivery. Additionally, a RFP gene is co-expressed from the same mRNA as the Cas9 protein via a 2A peptide linkage enabling tracking of transfection efficiency in cell populations via flow cytometry. A T7 promoter sequence is located immediately upstream the Cas9 cDNA sequence, allowing *in vitro* Cas9-RFP mRNA synthesis.

The Cas9 protein implemented in these vectors contains both HNH and RuvC activities enabling the creation of double strand breaks. Cas9 is linked to EVROGEN™ TagRFP fluorescent proteins. TagRFP is a monomeric red (orange) fluorescent protein generated from the wild-type RFP from sea anemone *Entacmaea quadricolor*[780]. It possesses bright fluorescence with excitation/emission maxima at 555 and 584 nm, respectively. In both GFP and RFP vectors, the 2A-FP encoding sequence is flanked by two *Hpa*I restriction sites, which allows removal or replacement of the 2A-FP element. The *XbaI* site can be used to linearise the vector for production of Cas9-FP mRNA via *in vitro* transcription using T7 RNA polymerase.

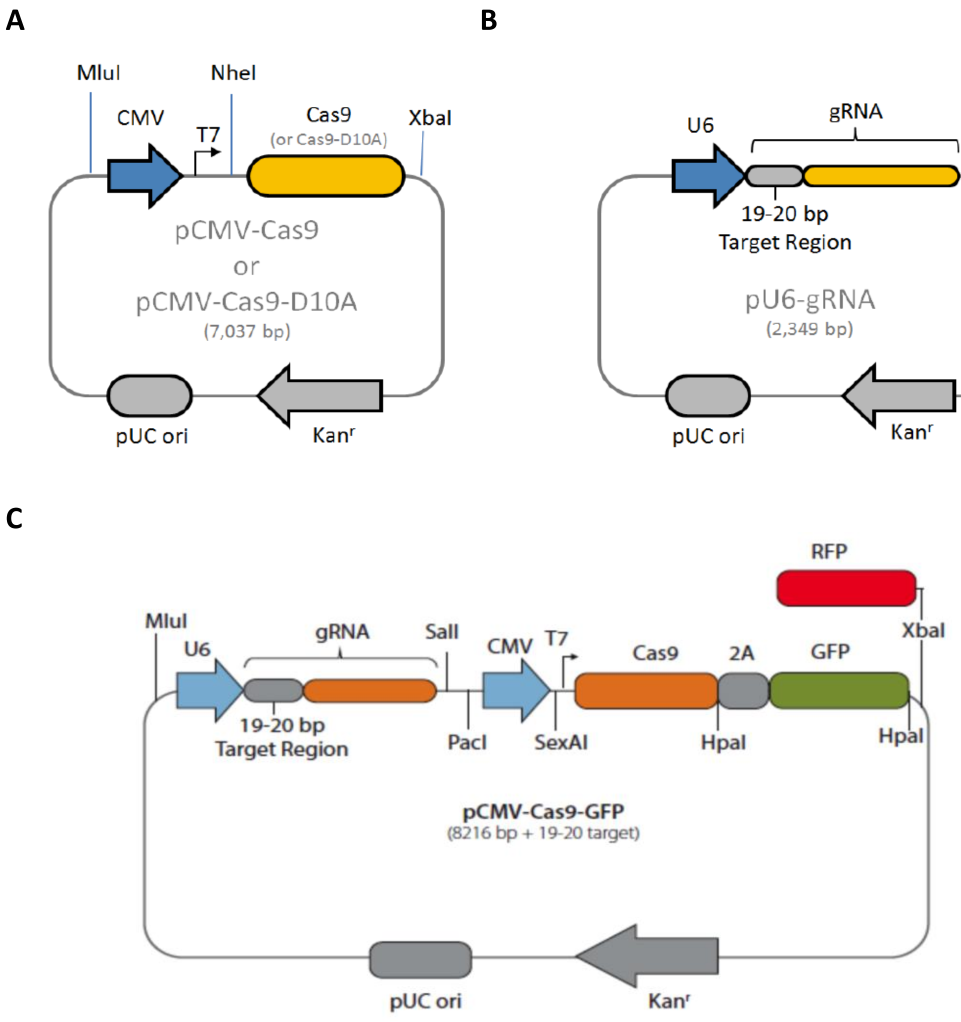


Figure 5.2. Schematics of CRISPR-Cas9 plasmids used for knock-in of a donor construct on HEK 293 6E cell line genomic positions.

(A) pCMV-Cas9 plasmid map. **(B)** pU6-sgRNA plasmid. **(C)** pCMV-Cas9-U6-sgRNA-RFP (sgRNA and Cas9 combined in the same plasmid). pCMV-Cas9 plasmid map and pU6-sgRNA plasmids were used for targeting the RRL SIN cPPT 2HA EEW Zeo transfer vector into the EMX1 locus. pCMV-Cas9-U6-sgRNA-RFP was used for targeting the RRL SIN cPPT 2HA EEW Zeo BFP transfer vector into the EGFEM1P, CUL5 and 'intergenic' locus.

5.4 Delivery of CRISPR-Cas9 and donor plasmids and screening

5.4.1 Validation of targeted integration into the genome of HEK293 6E

In order to deliver the CRISPR-Cas9 plasmids, 2µg of each plasmid (Sigma U6-sgRNA, CMV-Cas9 and EMX1 donor, Figure 5.2A, B and Figure 5.1D, respectively) were nucleofected into 2x10^6 cells per condition using the Amaxa Nucleofector 2b and program S-018. Such conditions were previously optimised to maximise transfection efficiency (data not shown). In the case of the two genomic loci discovered in Chapter 4 (EGFEM1P Chr3: 168,010,733- 168,010,734 and CUL5 Chr11:107,887,987-107,887,988), 2µg of each plasmid of each of the two plasmids (Figure 5.2C and Figure 5.1D, respectively) were co-transfected into HEK-293 6E adapted to adherent culture.

Two days post-transfection, zeocin selection (500µg/mL) was applied to the cells for 14 days. 24 EMX1, 34 EGFEM1P and 24 CUL5 colonies were randomly selected, isolated and amplified separately in adherent conditions for 15 days prior to re-adaptation to suspension conditions; EMX1 colonies 18, 19 and 21 did not re-adapt to suspension conditions; all EGFEM1P and CUL5 colonies successfully re-adapted to suspension cultures. The CRISPR-Cas9 editing strategy and timelines are schematically represented in Figure 5.3A.

Precise integration of donor constructs was assessed by PCR amplification of donor construct-host genome junctions. Genomic DNA was extracted from all cell clones and PCR analysis was performed to verify the integration of the lentiviral vector. 5 different sets of primers binding genomic regions outside homology arms and 2 sets of primers binding the internal lentiviral *gag* and zeocin cassette (for 5' and 3' ends, respectively) were tested for each candidate site in order to optimize the PCR conditions in Figure 5.3B. Eventually, EMX1 right/left, EGFEM1P junction and CUL5 right/left junction were screened with primer sets detailed in Materials and Methods 2.2.43.

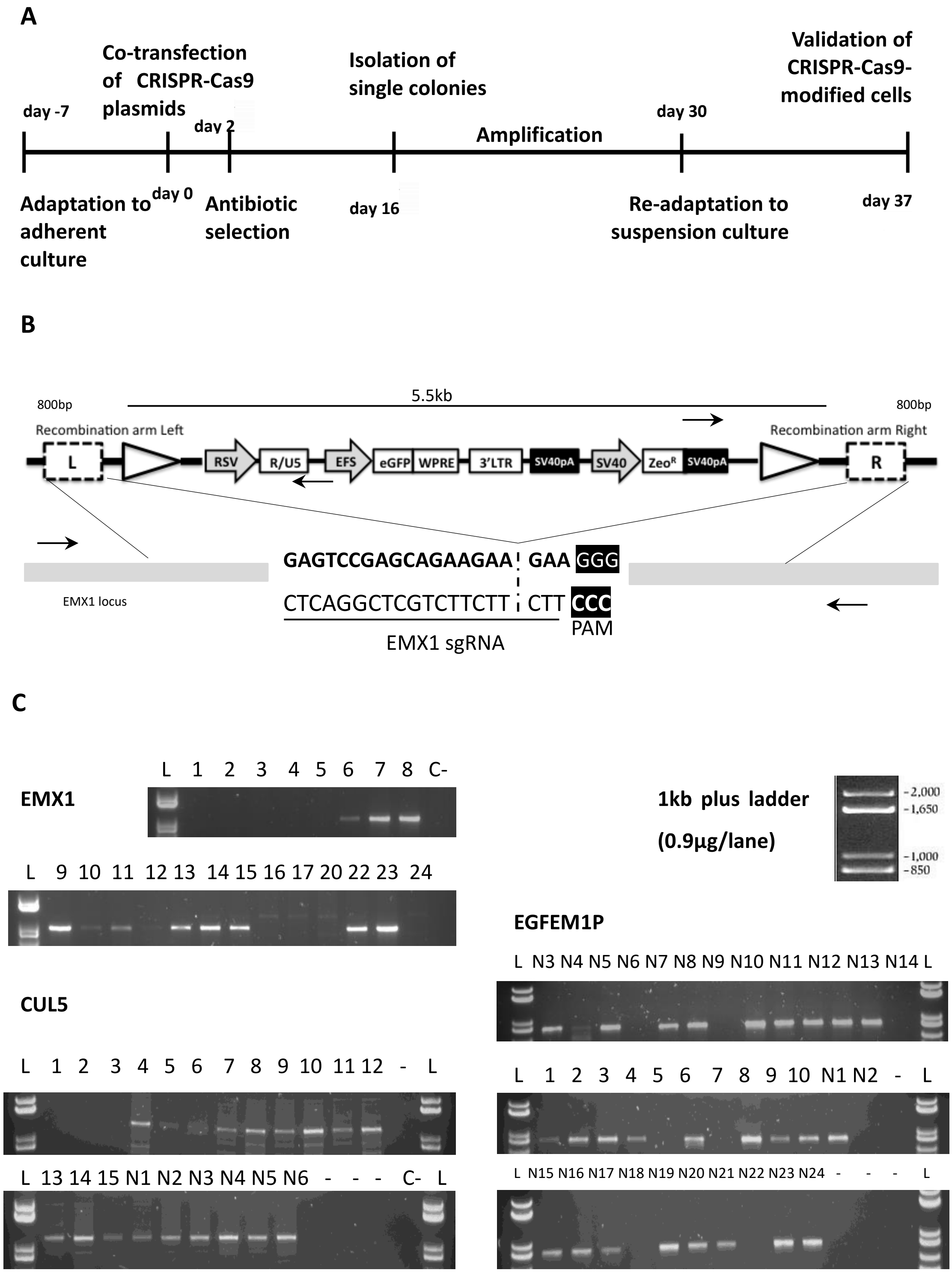

A
Co-transfection of CRISPR-Cas9 plasmids
Isolation of single colonies
Validation of CRISPR-Cas9-modified cells
day -7
day 2
Amplification
day 30
day 0
Adaptation to adherent culture
Antibiotic selection
day 16
Re-adaptation to suspension culture
day 37
B
800bp
5.5kb
800bp
Recombination arm Left
Recombination arm Right
L
RSV
R/U5
EFS
eGFP
WPRE
3'LTR
SV40pA
SV40
Zeo^R
SV40pA
R
GAGTCCGAGCAGAAGAA GAA GGG
CTCAGGCTCGTCTTCTT CTT CCC
EMX1 locus
EMX1 sgRNA
PAM
C
EMX1
L 1 2 3 4 5 6 7 8 C-
L 9 10 11 12 13 14 15 16 17 20 22 23 24
1kb plus ladder
(0.9µg/lane)
-2,000
-1,650
-1,000
-850
EGFEM1P
L N3 N4 N5 N6 N7 N8 N9 N10 N11 N12 N13 N14 L
CUL5
L 1 2 3 4 5 6 7 8 9 10 11 12 - L
L 1 2 3 4 5 6 7 8 9 10 N1 N2 - L
L 13 14 15 N1 N2 N3 N4 N5 N6 - - - C- L
L N15 N16 N17 N18 N19 N20 N21 N22 N23 N24 - - - L

Figure 5.3. CRISPR-Cas9 'knock-in' strategy and validation by junction PCR.

(A) Overview of CRISPR-Cas9 genome editing strategy timelines. (B) Schematic illustrating integration of a donor fragment of DNA into the desired locus; primers for screening the junctions for correct HR are indicated with arrows. (C) Predicted band sizes of CRISPR-Cas9 junctions are comprised between 1.1kb, 850bp and 1.3kb for EMX1, EGFEM1P and CUL5, respectively. L, ladder; C-, negative control.

11 of 21 EMX clones (52%) showed a band for specific integration of a donor construct. For EGFEM1P and CUL5 clones, targeting efficiencies (with selection) raised up to 76% (26 and 16 positive clones out of 34 and 21, respectively) (Figure 5.3C). The similarity in targeting efficiency shows consistency between different targeted genomic locations. The difference in the targeting efficiency between EGFEM1P/CUL5 and EMX1 could be explained by the fact that sgRNA and the Cas9 were co-delivered into HEK293 6E cells instead of using a 3 plasmids (donor + sgRNA + Cas9) system, which contributed to simplify the system and enhance transfection efficiency.

PCR products resulting from the amplification of CRISPR junctions were ligation-independently cloned into a pCR4 TOPO-TA vector backbone. Sanger sequencing analysis with the M13reverse primer (5'-CAGGAAACAGCTATGAC-3') verified expected sequences demonstrated precise integration on both genome-donor 5' and 3' boundaries, proving HDR integration of donor plasmid DNA into the desired position (Chr2: 73160998 - 73160999) in the host cell line genome (verified using Blat). The PAM motif (GGG following the *S.pyogenes* pattern 5'-NGG-3') can be seen immediately downstream to the three last nucleotides of the EMX1 sgRNA (GAA) (Figure 5.3B and Figure 5.4). This confirms that the DSB took place at the expected position (between the 17th and 18th position of the EMX1 sgRNA, right upstream the remaining GAA sequence) and also the correct directionality of the integration. DNA was harvested 37 days post-transfection so that potential amplification from residual transfected donor plasmid was not possible.

Interestingly, Sanger sequencing of the DNA fragments displaying the bands (with expected DNA sizes) shown in Figure 1.3C revealed amplification of sequences located in Chr5:127,818,094 for EGFEM1P and Chr12:103,973,128 for CUL5 instead of their expected loci (Chr3:168,010,733 plus strand and Chr11:107,887,987 plus strand, respectively) and no trace of donor construct. No sequence homology was found within 20kb surrounding these loci, which indicates this result could be due to a PCR artefact. Analysis of sgRNA specificity shows that no off-target effects are expected in chromosome 5 and 12 with less than 4 mismatches in the sgRNA (data not shown). In addition, the predicted potential positions in those chromosomes do not correspond with the amplified sequences. Therefore, no evidence of targeted integration was obtained from the amplification of EGFEM1P and CUL5 junctions.

This result contrasts with the successful integration shown for the EMX1 position obtained following the same procedure. A reasonable complementary approach to confirm these results could be to attempt amplification of donor-genome junctions on the off-sites predicted by the sgRNA design tool (http://crispr.mit.edu/).

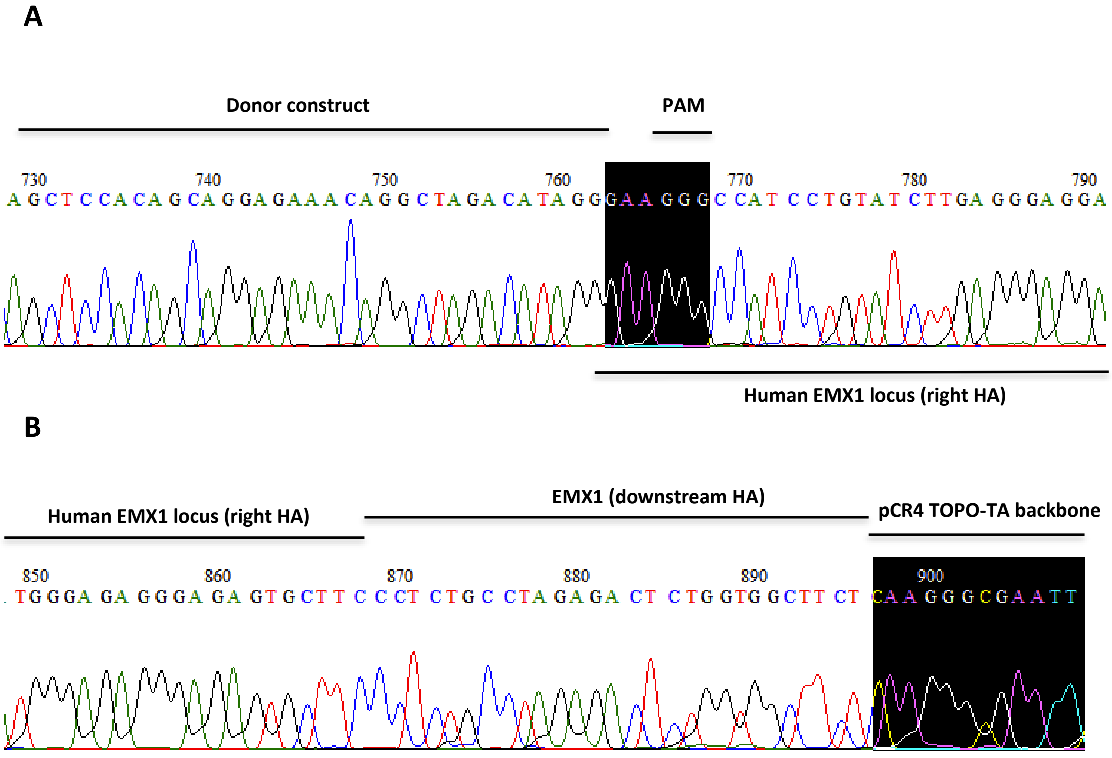


Figure 5.4. Sanger sequencing results from integrated donor plasmid-host cell line chromosome 3' junction amplified by PCR.

(A) Trace containing the junction between the donor construct and the beginning of the homology arm. The highlighted area shows the last 3bp (GAA) of the EMX1 sgRNA followed by the *S.pyogenes* PAM pattern (NGG). (B) Comparable junctions were obtained for the remaining GFP positive clones screened.

5.4.2 Assessment of GFP expression in CRISPR-Cas9-modified HEK 293 6E cell clones

GFP expression from the integrated transfer vector was assessed by FACS 15 days after isolation of single colonies grown under zeocin selection. All clones were GFP+ and 16 of the 21 EMX1 clones (76%) presented a defined single GFP peak suggesting that the composition of cells in that clone is uniform. Interestingly, the viability of clones 8, 9, 14 and 20, some of them reporting high levels of GFP (Figure 5.5), dramatically dropped after 3-4 days of isolation (data not shown). This result agrees with that of Chapter 3 (Figure 3.13) suggesting that high levels of GFP expression are detrimental for cell viability. EGFEM1P and CUL5 GFP intensity was also assessed by flow cytometry. 100% of the EGFEM1P clones were GFP+ (presented ratios >1 compared to non-fluorescent cells) while in CUL5 this percentage was lower (76%) (Figure 5.5B, C). A single peak of GFP intensity indicating homogeneity in the composition of the clone was obtained in 30 (88%) and 16 (76%) of them, respectively. Clones showing 2 peaks of GFP intensity were excluded from the screening in order to discard heterogeneous signal/integration.

Interestingly, while EMX1 stable clones with the highest ratios were close to 30 times more fluorescent than control cells, the ratios of highest expressers for EGFEM1P and CUL5 resulted to be 2-3 times lower than EMX1 values. The presence of GFP in samples in which integration was not detected can be attributed to off-target integration.

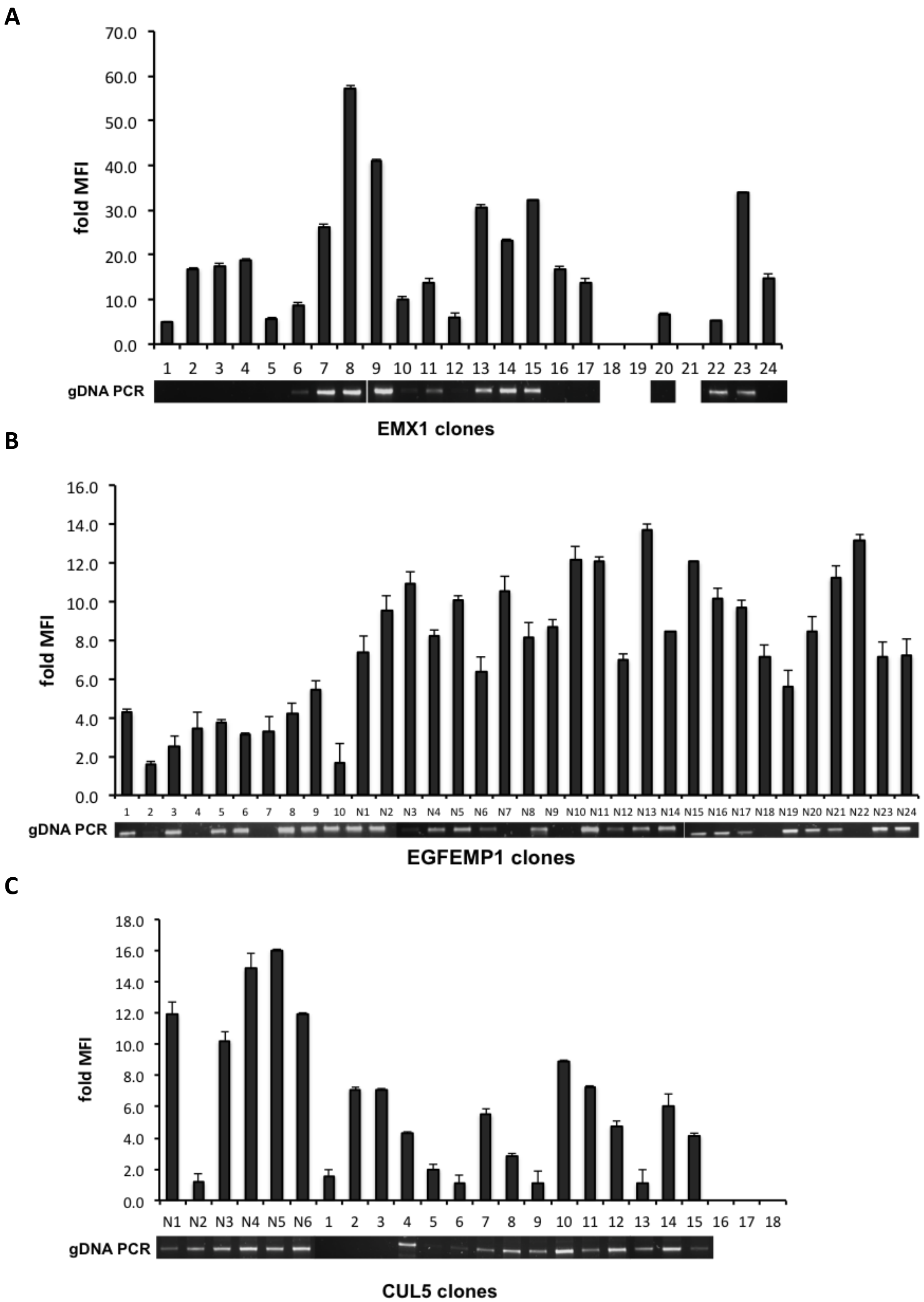


Figure 5.5. Mean fluorescent intensity of CRISPR-Cas9-edited clones and confirmation by junction PCR.

MFI values are expressed as fluorescence intensity relative to GFP-negative cells. Successful homologous recombination at the target sites (A) EMX1, (B) EGFEM1P and (C) CUL5 was confirmed by genomic PCR as described in Section 2.2.6. N (in the x-axis) stands for the clones that were extracted in a second batch, although all clones were treated equally.

5.4.3 Assessment of Cas9/sgRNA transfection efficiency

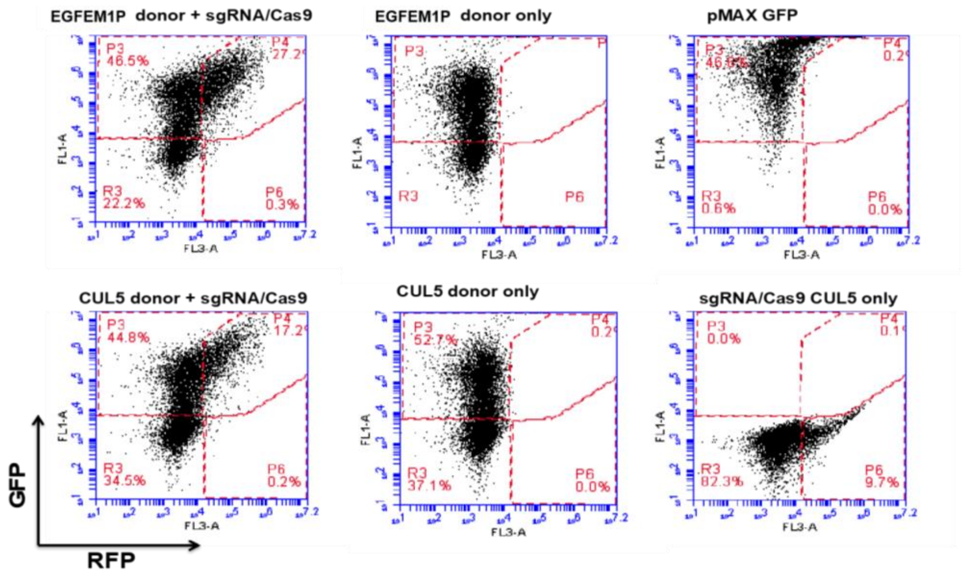


Figure 5.6. Flow cytometry analysis of Cas9/sgRNA and donor construct co-expression in HEK 293 cells.

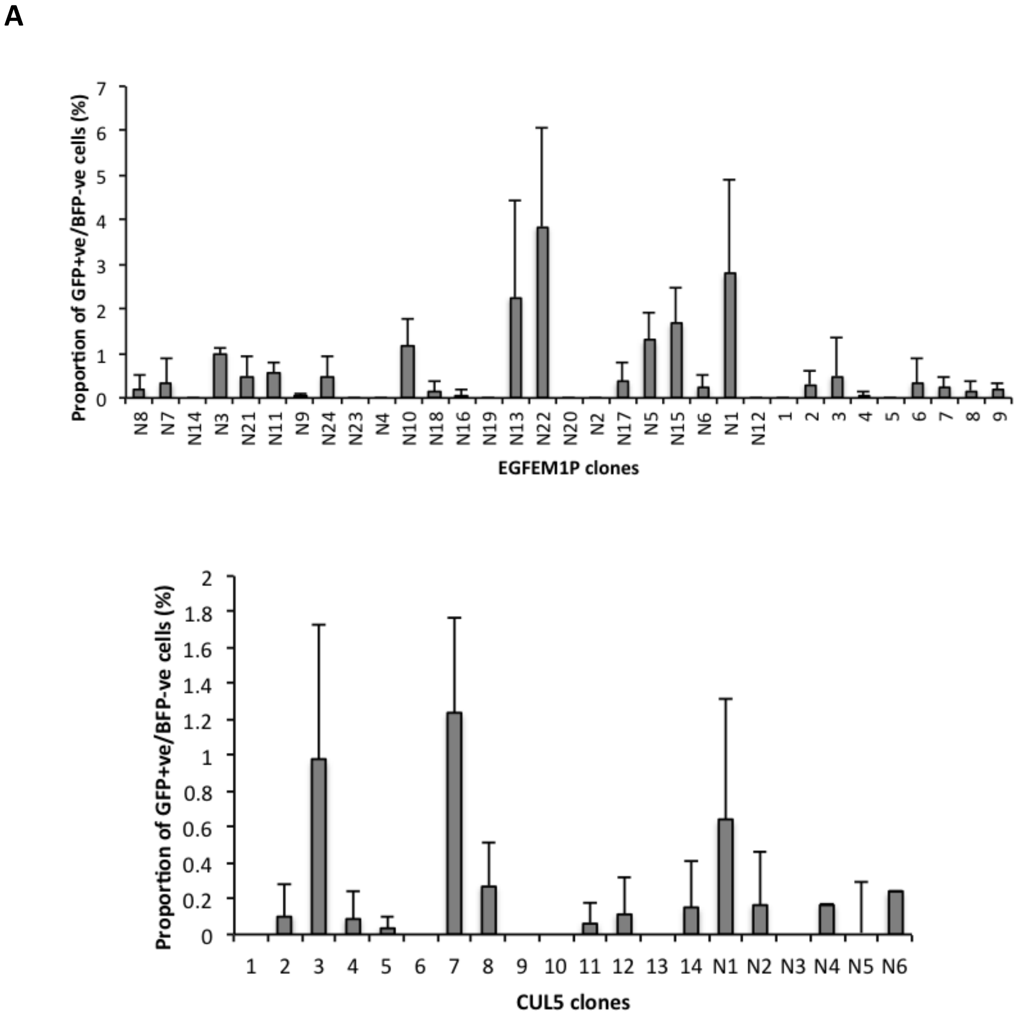
Cells were transfected with pCMV-Cas9-U6-sgRNA-RFP plasmid and donor construct (Figure 5.2). Marker gene expression was determined by flow cytometry after 3 days post-transfection. GFP and RFP fluorescence indicates expression of transfer vector and Cas9, respectively. Results for sgRNA/EGFEM1P only are comparable to those of sgRNA/EGFEM1P.

Cas9 plasmid transfection efficiency was confirmed by flow cytometry for RFP fluorescence 3 days post-transfection in EGFEM1P and CUL5 transfections. Transfection efficiency rose up to 75% for cells co-transfected with the sgRNA/Cas9 + donor plasmids and was 50% for cells transfected only with the donor construct. The latter could be attributed to non-specific integration as well as transient expression. 17% and 27% in HEK 293 cells co-transfected with sgRNA/Cas9 and donor plasmids were RFP and GFP positive, which indicates that both the nuclease proteins (fused to RFP with a 2A peptide) and the transfer vector (encoding for GFP) was being expressed in EGFEM1P and CUL5 clones (). RFP fluorescence was not detected in edited clones after 1-2 weeks post-transfection by cell imaging indicating the expression was transient as expected (data not shown). pmaxGFP was used as a GFP single fluorochrome positive control and plasmids containing EGFEM1P and CUL5 pCMV-Cas9-U6-sgRNA-RFP were used as a RFP single fluorochrome positive control. Western Blot analysis of Cas9 expression on transfected cells would confirm these results.

5.5 Screening of specificity, copy number and titer on targeted packaging cell lines

5.5.1 Assessment of background integration

To evaluate the rate of non-HR mediated donor construct integration in EGFEM1P and CUL5 clones (EMX1 clones lack the BFP marker in the donor construct), the GFP and BFP fluorescence levels were measured on all clones using an IN Cell 2000 imaging system. Successful homology directed repair would result in the integration of a GFP cassette only. In the case of a single random integration, cells are expected to appear double positive for green and blue fluorescent protein. pmaxGFP (supplied with the Invitrogen Nucleofection kit V, VCA-1003, Appendix A) and pMA-RQ HA-MCS-HA2-BFP plasmid (Appendix A) were used as green and blue single fluorescence controls.



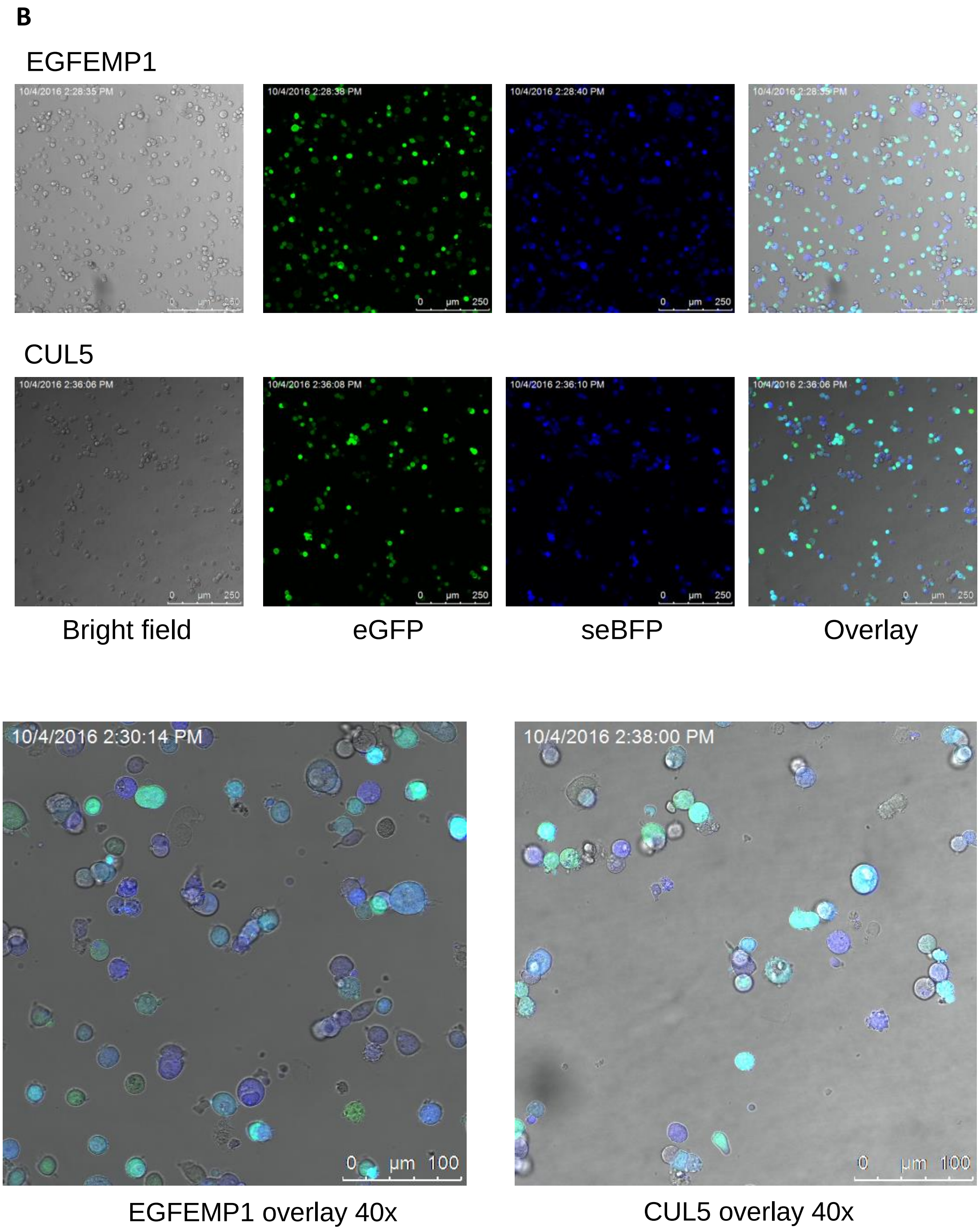


Figure 5.7. Screening of successful homologous recombination integration events in EGFEM1P and CUL5 edited clones by cell imaging.

(A) Quantification of the proportion of GFP+ve/BFP-ve cells within the total population of transduced cells using a custom script in Columbus imaging software (Section 2.2.42 and Appendix B). (B) Images were taken using a fluorescence confocal microscope at a 20x or 40x magnification. eGFP, enhanced green fluorescent protein (430-480nm); seBFP strongly enhanced blue fluorescent protein (510-600nm). No overlap between the detection spectra was observed between seBFP and eGFP. EGFEM1P clone 2 and CUL5 clone 2 are shown; results were comparable for all clones (data not shown). N (in the x-axis) stands for the clones that were extracted in a second batch, although all clones were treated equally.

Co-expression of BFP and GFP was observed in more than 95% of cells (higher in CUL5 clones) in all clones of EGFEM1P and CUL5 indicating random integration (Figure 5.7). No clone with high percentage of eGFP only was observed, indicating that homologous recombination at these loci did not occur.

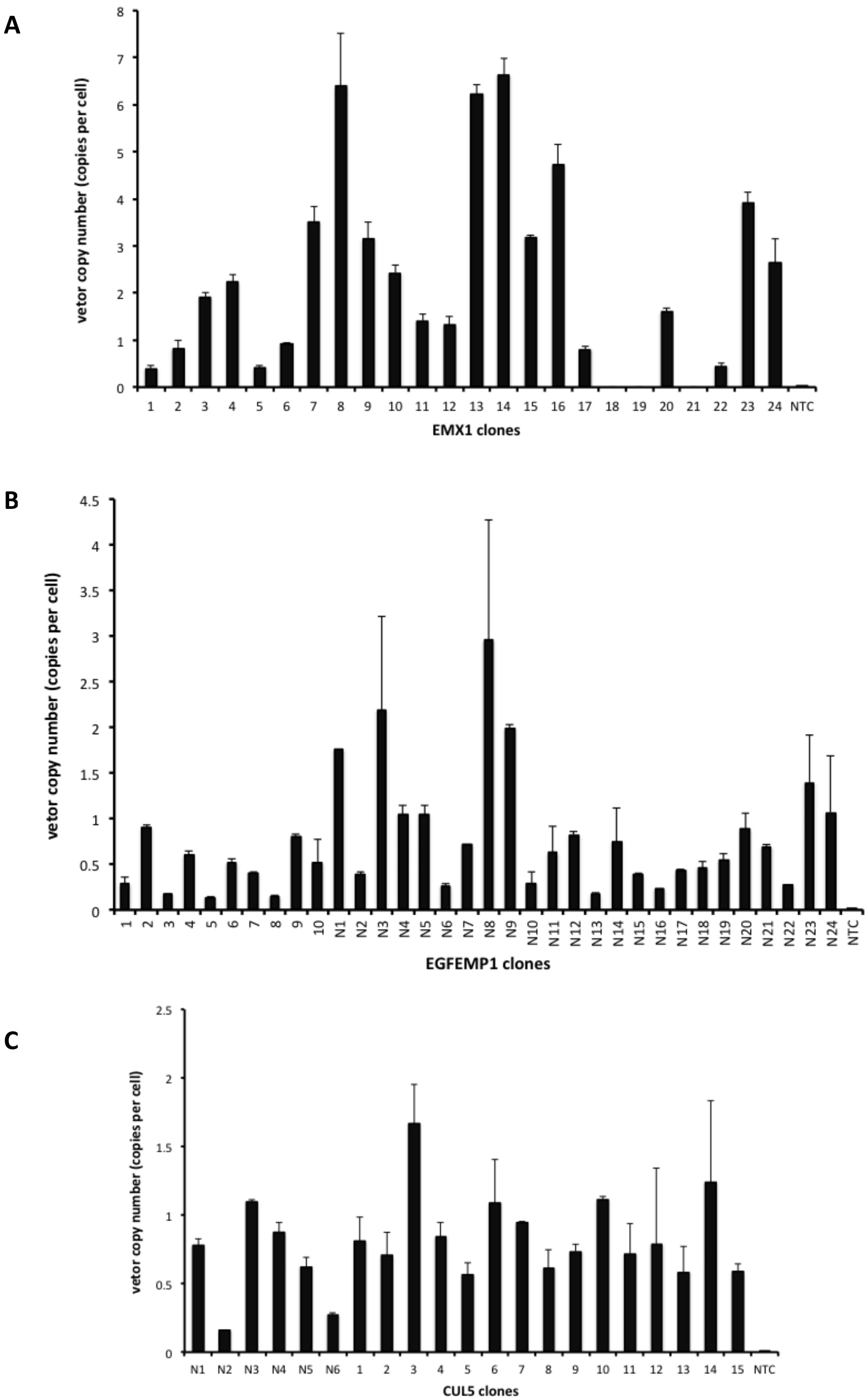
5.5.2 Lentiviral vector genome copy number of packaging cell lines

Quantification of the number of copies of transfer vector was performed on all clones by qPCR. Absolute quantification accounting for the mass of a cell genome (based on an average number of chromosomes[781]) was used to normalize the number of copies obtained per number of cells (calculations explained in Section 2.2.25).

Most EMX1 clones harboured between 1 and 3 copies of the transfer lentiviral vector. Clones 8, 13 and 14 reported up to 6 integrated copies of transfer vector, which resulted in proportionally increased GFP fluorescence. In EGFEM1P and CUL5 clones, most clones show around 1 copy and the highest expressers report up to 3 and 1.5-2, respectively.

The presence of multiple copies of integrated lentiviral transfer vector can be explained either by off-target integration, variable penetrance of the insertion within the multiple potential alleles of HEK293 cell lines or the variability of the assay. The HEK 293 cell line is originally hypotriploid with a modal number of 64 chromosomes occurring in 30% of the cells (ATCC® CRL-1573™); therefore, copy numbers greater than 2 are biologically possible. However, the rate of multiple HR events is expected to be lower than a single HR, which suggests off-target integration has occurred.

As can be seen in Figure 5.5A, C and A, C, there is a strong correlation (Pearson's R^2=0.643) between GFP intensity and the number of integrated copies of lentiviral transfer vector in EMX1 clones (Figure 5.8D) unlike EGFEM1P and CUL5 clones.


A
vetor copy number (copies per cell)
0
1
2
3
4
5
6
7
8
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 NTC
EMX1 clones
B
vetor copy number (copies per cell)
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
1 2 3 4 5 6 7 8 9 10 N1 N2 N3 N4 N5 N6 N7 N8 N9 N10 N11 N12 N13 N14 N15 N16 N17 N18 N19 N20 N21 N22 N23 N24 NTC
EGFEMP1 clones
C
vetor copy number (copies per cell)
0
0.5
1
1.5
2
2.5
N1 N2 N3 N4 N5 N6 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 NTC
CUL5 clones

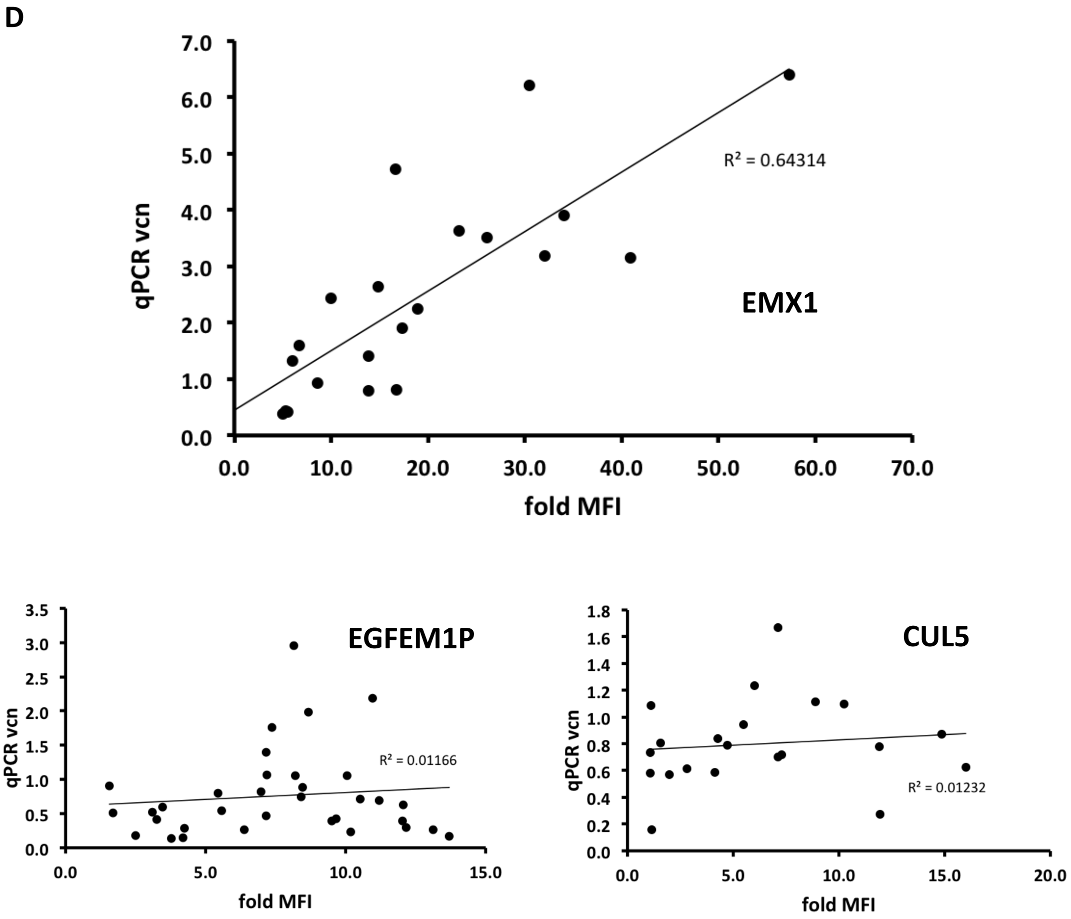


Figure 5.8. CRISPR-Cas9-integrated lentiviral vector copy number on clones and correlation with the fold MFI.

(A, B and C) Vector copy number was assessed by RT-qPCR (2.2.25). Samples labelled N* indicate clones isolated without cloning rings but by pipetting after incubation with 10%(v/v) TrypLE dissociating agent in PBS. NTC, non-template control 1. Results expressed as means ± SD. (D) Pearson's correlation between qPCR vector copy number and fold MFI for the three genomic candidates

5.5.3 Lentiviral vector titer of packaging cell lines

EMX1 clones 2, 17 and 23 and the EGFEM1 and CUL5 polyclonal pools were chosen for lentiviral vector production. The criteria behind the choice of EMX1-clone 2 and clone 17 comprised the detection of CRISPR-Cas9 mediated integration and a copy number of 1 indicating no off-target integration. EMX1-clone 23 was chosen because it reported the highest MFI (within the clones that successfully readapted to suspencion culture) despite the copy number being greater than 1. Clones with different MFI values were chosen in order not to link the vector production yield to enhanced GFP expression driven by other reasons (interclonal intrinsic variation, clonal fitness). Vector titers from EMX1 clones 2 and 17, EGFEM1P and CUL5 packaging cell lines were undetectable (Figure 5.9). HEK 293 6E cells transduced with viral supernatant from EMX1 clone 23 did show low levels of transduction, translated in titers of 10^4 TU/mL. However, titers were significantly lower compared to a standard lentiviral preparation (10^7 TU/mL) using non-modified HEK293 6E cells cotransfected with a 4 plasmid system (gag-pol, rev, VSV-G and transfer vector).

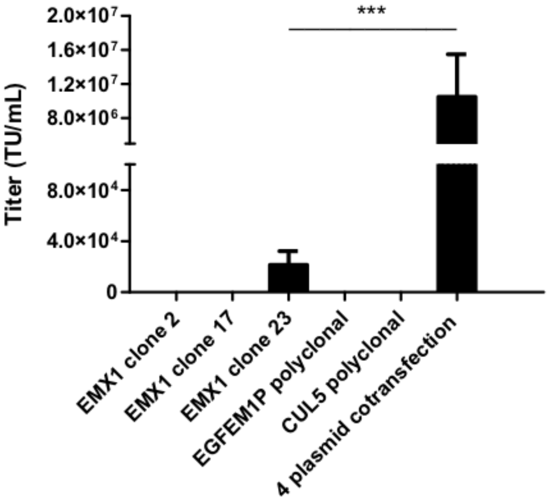


Figure 5.8. Functional lentiviral vector titration by flow cytometry on CRISPR-Cas9-modified clones.

Packaging cell lines were transiently co-transfected with gag-pol, rev and VSV-G for the production of lentiviral vectors using the calcium phosphate technique. Titers were calculated as indicated in Section 2.2.18. All results presented (means ± SD; * $p<0.05$, ** $p<0.001$, *** $p<0.0001$, grouped per cell type, Friedman's test analysis of variance) correspond to 3 technical replicates. Titers were assessed on HEK 293 6E cells.

5.6 Summary of results and concluding remarks

- A 5.5kb lentiviral transfer vector donor construct was specifically integrated into the expected loci at the EMX1 gene using CRISPR-Cas9 genome editing technology.
- Targeting efficiencies observed were high (57%) at the EMX1 locus. The integration of donor template was mediated via homologous directed repair pathway as the junction sequence (with no indels) indicates.
- Random integration was detected when a donor plasmid containing a transfer vector was used to target EGFEM1P and CUL5 loci, indicating unsuccessful CRISPR-mediated homologous recombination.
- Functional lentiviral production titers derived from the integration of a transfer vector in the EMX1 loci reported 10^4 TU/mL.

Successful targeting in EMX1 clones

The results presented in this chapter provided evidence that CRISPR-Cas9 can be used as a genome editing tool to mediate knock-in of a >5kb functional transfer vector cassette into HEK 293 6E host cells. A lentiviral transfer vector was integrated into the EMX1 gene, reporting a targeting efficiency of 57% (using a 3 plasmid system). These targeting efficiencies are comparable to those reported in recent articles describing 'knock-in' strategies using antibiotic selection even though the size of the donor construct is relatively higher (see Appendix A). Successful integration was confirmed by Sanger sequencing of one of both junctions.

Correlation between barcode counts and titer

Although targeted integration of donor construct containing transfer lentiviral vectors was successful, transduction of HEK 293 6E cells with supernatant from EMX1 clones 2 and 17 resulted in no titers. Despite being undetectable, EMX1 clone 23 (with a vector copy number of 4) showed low levels of transduction (10-15% transduction in the 1/10 dilution). Taking into account that vector copy numbers of 14-59 reported by Hu *et al.*,[463] or up 200 with the concatemeric array[292] are necessary to achieve 10^7 TU/mL titers, it is understandable that a vector copy number of 4 is not sufficient. In this study, individual loci were targeted with donor plasmid containing the transfer vector genome in order to test the hypothesis of the contribution of the a genomic environment on the expression of a particular barcoded vector transcript. Targeting multiple genomic locations would have made more difficult to assign the contribution of a particular position to the overall expression of vector genome.

Several factors could explain the lack of correlation between the high number of barcode counts associated to a particular locus and a low or inexistent vector titer from a transfer vector integrated in that position:

- Firstly, the stable integration of a lentiviral transfer vector might compromise the stoichiometry of the 4 necessary components required to produce 3rd generation lentiviral vector. In EMX1, EGFEM1P and CUL5 clones, the remaining plasmids containing *vsv-g*, *gag-pol* and *rev* were provided following the stoichiometry used for transient transfection (Section 2.2.16). However, the number of mRNA molecules in cell cytoplasm was not analysed. RT-PCR of the viral transcripts could provide insights on the actual messenger ratios although amplification bias could skew the results. Transcriptomic analysis could contribute to optimise the stoichiometry in addition to help explain the differences seen between clones and copy numbers.

- In line with the lentiviral integration preferences exploited in this work, Cas9 activity was reported to be higher in open chromatin regions[782]. Nonetheless, efficiencies could substantially vary within a particular locus. The high-transcribing position of vector integration retrieved by LM-PCR does not exactly correspond to the position where the nuclease excised the DNA. The DSB produced after the 17th base pair of the sgRNA and the choice of site of integration are subject to the presence of a PAM sequence (NGG) from *S. pyogenes*. Restriction to the *S.pyogenes* PAM limits genome accessibility to 1/42bp (the average frequency of a GG dinucleotide in a DNA sequence)[782]. The use of other PAM sequences such as NNAGAA and NGGNG for *S. thermophiles*[459] or NNNNGATT for

N.meningiditis[783,784] could answer that question although such systems are not standardised at a commercial level. Alternatively, targeting efficiencies could be tested using different sgRNA sequences (protospacers).

- On a technical note, optimisation of co-transfection parameters showed that delivery of higher amounts of plasmid also contributes to increase the transfection efficiency. CRISPR-Cas9 is a relatively new genome editing technology and further investigation is required to determine the optimal parameters for addition of foreign DNA into cells. Although most studies use homology arms 500bp-1kb and repair templates of up to 6kb, the interdependencies of these characteristics and their direct effects on the targeting efficiency of the system remains to be explored.

Transfection of linear plasmid DNA has been known to yield more stable clones compared to supercoiled circular plasmid[501]. In fact, *PI-SceI* yeast restriction sites had been designed opposite the exchangeable components of the donor plasmid in order to explore that possibility. However, despite being more recombinogenic, linearised plasmid DNA has also been described to be taken up less efficiently by the cell[785]. In addition, endogenous exonucleases could degrade linearised DNA[785]. Since, the delivery of the CRISPR-Cas9 plasmids did not become a limitation (as can be seen in) and posterior selection was going to be applied to cells, this measure was not applied.

Interclonal variation in mean fluorescence intensity and vector copy number

We also observed variability in GFP expression and titer within clones with the same amount of vector copies. Interclonal variation is a common phenomenon observed in cell line development consisting in variability in the performance of clones that theoretically share the same genotype. This heterogeneity manifests in measurable variation in terms of cell densities, growth rate and protein secretion[786]. In 2005, Barnes *et al.*, studied the causes of this phenotypic drift and demonstrated this phenomenon can be observed in the absence of selective pressure[787]. In their study, they attributed interclonal divergence to intrinsic genetic differences due to the natural stability of mammalian cell lines. Random mutations in cell cycle-regulator or proteins involved in the nutrient uptake can affect the individual yield of each cell. They also showed phenotypic drift is more likely to occur if parental cell lines have been cultured for prolonged periods as accumulation of changes. However, it is unclear if this heterogeneity is intrinsic to the clone or arises after isolation of the single cell. Other factors such us epigenetics have been suggested to play a role in clonal variability through binary switching of endogenous metabolic genes. In addition, stochastic fluctuations of endogenous genes in cell cultures have been described[622]. Expression of p27 protein, a member of the cyclin–dependent kinase family, is often screened as this protein cell cycle regulator inhibits cell cycle progression. High levels of p27 have been shown to correlate with decreased cell growth rates[786]. However, (and in order to understand the extent of this phenomenon) heterogeneous expression is not limited to different clones. Pilbrough *et al*., showed that expression noise can also occur within cells of the same clone[788]. Stochastic bursts of promoter activation linked with fluctuations in chromatin folding dynamics generates a graded repertoire of expressions[789]. This burst is then subject to amplification by protein and mRNA turnover in the timescale of hours for higher eukaryotes[790].

Acquired phenotypic drift has also been observed and could explain why the rate of variation for different parameters remains low (7% in Barnes *et al*.,) in early passages and increases over generations. In any case, Kim *et al*., remarked that the term ‘clones’ was not sufficiently accurate and instead the idea of a “clonally-derived population” was more realistic[791]. Another source of acquired heterogeneity is the lack of control of the effects of integrated copies of donor plasmid in host chromosomes. Deregulation or disruption of endogenous genes can result in interclonal changes in response to environmental factors such as temperature or pH[792,793]. In this study, all clones had the same number of passages, which discards the accumulated drift hypothesis. Therefore, differences in GFP expression of cell generated could be explained by intrinsic heterogeneity or as a result of the effect of random integration. Precisely because of this reason, clones displaying different levels of GFP (with low/no random integration and low copy number) were tested for lentiviral production.

Targeting specificity

Tightly coupled with the efficiency of the editing is the specificity of the DSB. In this study, we showed specific integration of a donor construct in the expected EMX1 loci. However, copy numbers higher than one (and especially higher than three, being HEK293 considered hypotriploid) observed in some clones do not allow discernment between off-target integration and targeted integration in other alleles. Contrarily, specificity results for EGFEM1P and CUL5 showed that integration of the transfer vector construct took place in a random manner. In order to complement the specificity results obtained by PCR and quantify the potential off-target integrations, lentiviral transfer vector copy number was analysed on selected/isolated/amplified/re-adapted clones. Its correlation with the GFP intensity (particularly seen in CUL5 and EMX1 clones) and also observed in Charrier *et al.*,[724], reinforces the consistency of the outcome. Donor copy numbers of EMX1 were found to be higher than those of EGFEM1P and CUL5 (with values around 1 donor copy per cell). Nevertheless, values superior to 1 do not necessarily imply off-target integration events. Although it occurs less frequently, the integration of up to two separate copies (not in tandem) of donor plasmid could be explained by biallelic targeting (or even triallelic, given the hypotriploid nature of this cell line). Access to the karyotype of this particular cell line would provide insight into this aspect. As mentioned before, cell lines are known to undergo genomic rearrangements to overcome metabolic limitations and their genetic stability is compromised. Alternatively, the correct composition of the junctions and homology arms could be identified by Fluorescent In Situ Hibridisation (FISH)[794]. Next-generation sequencing executed with the MiSeq system (Illumina) would provide representativity of the rate of indels/integrations obtained although the read length could suppose a limitation to identify junctions with long homology arms. Single-molecule real-time (SMRT) sequencing allows for read lengths of up to 3kb and has been used to screen edited human cells[795]. A reasonable complementary approach to confirm these results could be to attempt amplification of donor-genome junctions on the off-sites predicted by the sgRNA design tool (http://crispr.mit.edu/)[581].

Possible explanations for random integration observed in EGFEM1P and CUL5 candidate loci

As previously stated, targeting efficiency and specificity are partly locus-dependent factors. However, 57% on target efficiency obtained in EMX1 clones contrast with the results observed in other candidate loci (EGFEM1P and CUL5), where targeted integration was not observed. Despite different clones showed a correlation between level of GFP expression and vector copy number, they failed to show targeted integration by PCR of integration junctions from genomic DNA. Therefore, the positional effect of the viral integration could not be assessed in these candidate positions. Polyclonal pools heterogeneously expressing transfer vector were assessed for EGFEM1P and CUL5 but functional titers were not detected. Possible reasons for failure to target candidate loci may include:

- The difference in targeting efficiency could be due to the co-transfection of a different number of plasmids. While EMX1 locus was targeted using 3 plasmids, EGFEM1P and CUL5 loci were targeted with 2 plasmids. Co-transfection of the EMX1 locus with the 2 plasmids system would answer that question.

- The design of the sgRNA was also found to influence its targeting efficiency. A study by Wang *et al.*, using a library of 73,000 sgRNAs and massive parallel sequencing helped determine the parameters for the design of effective sgRNA[796]. Purine rich sgRNA PAM-proximal regions as well as a balanced GC content and sgRNA were found to favour Cas9 activity. Taking that into consideration, the composition of the sgRNA sequences utilized in this work (EGFEM1P 5'-TACCTGGGGGT**GG**T**GG**T**G**T**A**-3' 60% GC content and CUL5 5'-

TACCTTCTTCCCT**A**C**AGG**TC-3' 50% GC content) would indicate that higher targeting efficiency should be expected in EGFEM1P and CUL5 loci. However, successful targeting was achieved in EMX1 clones with 15% GC content (EMX1 sgRNA 5'-TTAATGCTTATTT**A**TTTT**G**T-3'). Enhanced efficiency has also been reported if the sgRNA targets the transcribing strand. However, in this study, all three designed sgRNA targeted the non-transcribing strand where the closest PAM motif to the viral integration was.

- Another aspect that might have contributed to non-specific integration is the application of increasing concentrations of antibiotic post-transfection. Although antibiotic selection can help reduce selection timelines, this might also exert a negative effect on the specificity of the integration and the donor copy number. Similarly to genomic amplification strategies (e.g. DHFR/MTX system[624]) used for biopharmaceutical protein production, the metabolic burden imposed by the presence of zeocin may promote random integration, amplification or even genomic rearrangement instead of HR to adapt to such conditions. For that reason and in order to assess the targeting efficiencies without selection, a non-selection control could be added to the study.

In EGFEM1P and CUL5 clones, the occurrence of off-target effects might justify the presence of more copies and consequent higher reporter gene expression. A relatively high frequency (50%) of GFP+ cells was detected when transfecting donor construct (in absence of sgRNA/Cas9 plasmids). Besides transient GFP expression, screening of random integration was required to discern whether integration events were legitimate. The inclusion of a BFP selectable marker revealed that a large proportion of cells (>90%) had undergone random integration, independently of the locus. An alternative way to examine the off-target integrations would be to amplify potential junctions of donor plasmid with regions with a certain number of mismatches in the sgRNA sequence (predicted by sgRNA design tools). Alternatively, other studies have used negative selectable markers (such as the HSV thymidine kinase or diphtheria toxin A) located outside the homology arms to kill cells with randomly integrated donor constructs[797].

- Potential recombination between SV40 polyA. The donor plasmid harbours three SV40 polyadenylation signals that enable termination of the transcripts containing the gene of interest and zeocin and BFP selectable marker. Potential recombinations occurring between the three polyadenylation signals, which share exactly the same sequence, could explain the presence of a BFP marker between homology arms and thus its expression upon integration (either random or targeted). At a plasmid level, donor plasmids were completely sequenced before transfection. Therefore, in the event of a recombination within the donor plasmids, this must have occurred in the host cell line, not during the cloning process. However, successful targeted integration was achieved using the 3-plasmid system with a donor plasmid containing two exact polyA sequences (devoid of the BFP marker).

- The absence of integration of donor construct in the candidate loci might be explained by the lack of Cas9 cutting activity. An alternative approach to test this hypothesis could be to transfect cells with the plasmid containing the Cas9 and examine the cutting site with primers flanking it to see if there are any indels. Plasmid rescue is another alternative to retrieve bacterial backbone randomly integrated by digesting gDNA and circularising, transforming and selecting with the antibiotic resistance. However, although the fusion of RFP with a Cas9 is not sufficient to demonstrate cutting efficiency, it does confirm with the Cas9 expression () and therefore weakens this theory.

- Stoichiometry. The EMX1 locus was targeted using a 3-plasmid system (donor+sgRNA+Cas9) that worked compared to the 2-plasmid system (donor+Cas9/sgRNA) that did not work with EGFEM1P and CUL5. As a result, this might have implications on the stoichiometry of the system. The two-plasmid system drives expression of sgRNA and Cas9 from two different promoters within the same plasmid. As the same amount (in mass, 2µg) of each plasmid were used per single transfection in a single 6-well plate, the stoichiometry might be slightly different from a 8,236bp plasmid expressing both compared to two plasmids of 7,037bp and 2,349bp (Figure 5.2). In line with the stoichiometry point, the 3-plasmid system with a BFP marker was placed downstream of the homology arms in order to detect non-specific integration. That makes the donor construct larger (11.3kb vs 9.6kb), which means that with the same mass of plasmid DNA transfected there would be less molecules of plasmid vector. In order to test that hypothesis, the donor construct of EGFEM1P and CUL5 could be tested following the transfection conditions of EMX1 with the 3-plasmid system.

- Compatible origin of replication. Firstly, HEK 2936E cells have the EBNA1 antigen and can maintain expression of plasmids with an EBV origin of replication. Despite not possessing such origin of replication, donor plasmid sequences potentially similar to it could sustain stable episomal expression of the donor plasmid including transfer vector and the BFP selectable marker. However, no similarities were found between EBV origin of replication (GenBank: DQ279927.1) and any of the elements in the donor plasmid.

- Lastly, low levels (17 and 27%) of co-transfection of donor construct and sgRNA-Cas9 plasmid observed in might explain the low targeting efficiencies observed in EGFEM1P and CUL5. However, EMX1 clones, which showed 57% efficiency, were transfected following the same procedure and in addition, sgRNA and Cas9 were split in two different plasmids.

As previously stated, to our knowledge this is the first approach that uses a genome editing technology for lentiviral packaging cell line development. This study opens the door to the introduction of packaging plasmids and transfer vectore genomes into optimal positions as opposed to integrated via random integration and also preventing any concerns arising from integrated viral sequences.

A reasonable criticism to this work would be the limited number of sgRNAs tested for each genomic loci. Typically, three different sgRNAs targeting a particular region are designed in order to find the one that shows higher cutting efficiency. Due to the time limitation of the project, only one sgRNA per loci was tested. A relatively easy way of testing the cutting efficienty of several sgRNA would be to transfect the Cas9-sgRNA plasmid (and not the donor) and amplify expected target sites in the genome seeking for double strand breaks resolved by NHEJ (with insertions and deletions in their sequence).

Chapter 6

DISCUSSION

Typical strategies for the generation of lentiviral packaging cell lines are based on sequential stable transfection and selection for populations containing lentiviral packaging genes. The random and multi-step nature of this process requires arduous screening and limits the performance of higher producers, respectively. In the last 10-15 years, viral transduction was introduced as a means of delivery that can efficiently target actively transcribed sites with higher associated expression compared to conventional stable transfection. This represents a useful tool for the expression of packaging plasmids. However, state of the art transfer SIN LVV widely used in the clinic cannot be produced using non-SIN for safety reasons and are incompatible with SIN vector delivery systems. Although some solutions have been proposed (cSIN), titers are slightly lower and genotoxicity has not been extensively assessed[401,463].

The rationale behind this project lies on the idea of optimising the screening capacity of a semi-random insertion of a lentiviral transfer vector based on viral integration preferences. However, citing the words of Prof David James (University of Sheffield – advance biomanufacturing centre) in an oral communication at GSK, *"screening is the admission of cellular screening* *incapability. Design, don't screen".* He presented the rational design vs irrational screening paradigm and claims that while the engineering or analytical tools to analyse phenotype (bioinformatics, genome editing, synthetic biology, "-omics") are available, efforts still needs to be made in understanding the biological mechanism and interactions that govern production systems such as lentiviral production so that we can develop predictive models. These predictive models would then serve to iteratively test (design-build-validate) different expression configurations. Although this strategy was thought and executed for the previous wave of therapeutic products in the 1980s-2000s (hybridomas, recombinant proteins in *E.coli* and mammalian cell lines, protein engineering), it is reasonable to think that gene and cell therapies and regenerative medicine follow a similar path now that academic clinical trials meet industrial production. This project aimed to use irrational screening of integration sites to allow rational design of PCLs in the future.

Following a similar rationale to Sanber *et al.,*[381] lentiviral targeting of actively transcribed sites and nucleased-based genome editing (instead of retargeting via recombinase-mediated cassette exchange) could present a solution to that issue and is explored in this project. In addition, a barcode system was implemented to quantitatively evaluate the expression derived from each integration site and enhance the screening of naturally actively expressed sites.

Key findings and observations

The lentiviral barcoded library

In this study, we developed a simple and cost effective method that allows simultaneous genetic cell marking and screening of thousands of integration sites for cell line development. This is the first example of barcoding applied in lentiviral packaging cell line development. Apart from the aforementioned advantages, the barcode system offers attractive features summarised below:

The titers of the lentiviral vector library (vSYNT) are comparable to standard lentiviral preparations, indicating incorporation of a 70nt fragment of DNA containing a genetic tag is not deleterious for viral replication. Low 10^9 TU/mL titers were achieved after ultracentrifugation, which established the first empirical threshold in the maximum library complexity (4^{14} x 2^2 = 1,073,741,824 ~10^9).

During the library construction process, the scale up of the ligation reaction reported the highest improvement in cloning efficiency. Around 120,000 colonies were obtained from a single ligation experiment. 90% of them contained a barcode and NGS sequencing revealed the library composition was balanced, with no predominant clones/variants outgrowing the population. Experimental optimization of the oligonucleotide cloning was critical and minimised the impact of backbone re-ligation, backbone:insert ratios, ligation temperature/time conditions and oligonucleotide annealing mismatches in the overall cloning efficiency .

The applicability of this barcode method lies in the ability of the vector library to transduce cells. However, this is not a limiting factor since VSV-G pseudotype can efficiently transduce a wide range of cell types to meet different applications. On a different level, host cell line restriction factors dictate its permissiveness to the vector library. The versatility of this method allows assessment of the (therapeutic) transgene of interest driven by a promoter of choice and thus does not require a reporter gene. Another advantage is that selection does not need to be applied to maintain the cell tag. In addition and contrary to antibody/secretion selection methods, genetic marking allows for cells to be manipulated (passaged, frozen, thawed) and high transcribing integration sites will still be tagged.

Integration site analysis and barcode abundance

Notably, a modification of LM-PCR was used to retrieve integration sites; instead of being performed in the 3'LTR, primers were designed to anneal sequences immediately downstream the 5'LTR U3 region, allowing higher read lengths and thus longer junctions.

Integration preferences did not diverge from those reported in lentiviruses[232]. Integration sites were found to be more abundant in gene-rich regions and chromosomes and LEDGF/p75 protein tethered insertions along the transcription unit[798] of the gene as opposed to gamma-retroviral vector preferences (TSS). Primary weak consensus sequences[799] in the integration junctions were also identified.

The system is also compatible to any sequencing platform. Although analysis was performed with Illumina using a 300bp PE, retrieval of longer integration junctions would benefit from techniques that push NGS read length to 2x500bp[800]. Similarly, single molecule real time sequencing technology such as PacBio (that offers 100,000 reads of >1kb) or Roche 454 newest system (GS FLX Titanium XL that allows for sequencing of 1,000bp for 700,000reads) could help[801,802].

In order to process LM-PCR and RNA-Seq reads into annotated barcoded integration sites and sorted barcode counts, respectively, bioinformatics pipelines were designed. For LM-PCR custom scripts were written to filter sequences >20nt with high homology with the linker (mainly in the last 5 bases) and map them using BLAT[692] against the hg19 genome taking into account that indels and gaps are not expected, discarding ambiguous alignments and promoting regions with a high degree of intensity. Although the RNA-Seq processing pipeline was simpler, the RNA-Seq protocol itself was also optimised by extending fragmentation times and thus enabling synthesis of longer RNA –seq libraries, identified as critical factors for barcode retrieval. Out of the top 15 barcode variants with a higher number of barcode counts, 6 and 0 variants for 10^4-cell and 10^3-cell transductions were found to have a correspondence with a genomic position. The difference could be explained due to the fact that ten times more potentially high expressing positions are screened with the 10^4 TU library.

The complexity challenge

The complexity of the library reached 4x10^5 barcode variants according to the Lincoln-Petersen estimate[709]. In a single aliquot, 5x10^4 different barcodes were retrieved and 10% of them were overlapping events. Sequencing errors account for false barcode generated as misread events. In order to quantify their contribution to the library diversity, Starcode clustering analysis was performed throughout all the steps of this project[691]. Conversely, mean number of 11bp dissimilarities among barcodes was found to correlate with a dramatic increase of the library complexity if more than 3bp (out of the 14 nucleotides of the barcode) are edited.

The diversity of the library dictates the throughput of the system. Accordingly, the complexity observed at a plasmid and vector level, 10^3, 10^4 and 10^5 integration sites were screened using the lentiviral barcoded library method. 1,245 (for 10^3 TU applied) and 8,261 (for 10^4 TU applied) integration sites with distinct barcode variants were identified by high-throughput sequencing of LM-PCR reads. Therefore, the system presents an adequate scalability to a number of clones greater than current screening platforms. The lack of correlation between 10^5 TU applied and the IS retrieved was likely due to a change in the culture format. Besides that, the limiting factor in the screening capacity could be attributed to the complexity of the library. The 10^5 barcode variants threshold achieved during the library construction process is below the theoretical sampling space of the barcode design (10^9 variants) the 10^9 TU/mL of a lentiviral preparation and the 2M reads/sample sequencing capacity of a MiSeq run (RNA-Seq capacity is higher), as long as the computational power/time is not a limitation.

CRISPR-Cas9 Genome editing for packaging cell line development

Once EGFEM1P, CUL5 and 'intergenic' high-expressing positions were selected, site-specific integration of a transfer vector was attempted using the CRISPR-Cas9 technology. Sanger sequencing confirmed the insertion of a 5.5kb donor construct containing a lentiviral transfer vector and a zeocin selectable marker into the control EMX1 loci with a 57% targeting efficiency (similar to those reported in the literature with selection). Despite displaying Cas9 activity and the same protocol as the control position (although different CRISPR-Cas9 plasmids), EGFEM1P and CUL5 ('intergenic' was discarded) showed off-target integration. Among the targeted EMX1 clones, only EMX1 clone 23 with vector copy number of 4 reported measurable titers (10^4 TU/mL), highlighting the importance of the expression of transfer vector for the stoichiometry of lentiviral vector production.

Criticisms and potential improvements

The barcode system also presents some intrinsic limitations. Beyond technical difficulties related to library generation, including the cloning bottleneck to library diversity and the possibility of introducing multiple copies per cell (already discussed in Chapter 3 and 4), sequencing errors may become a problem for the library complexity and thus the throughput of the approach or application. False barcodes can be generated as a result of misreading events or amplification of those generated by the polymerases' biases. In order to control that variable, and apart from barcode clustering, future work should be oriented towards calibration of the library to compensate for those biases and accurately predict the library size. An internal control library with known number of manually cloned variants could help to improve the accuracy of library complexity determination. The downside of this would be the throughput of the library since the cloning process would be laborious and time-consuming. An alternative to this could be to use spike-in controls as in Brugman *et al*[735]. Different proportions (i.e. 1%, 5%, 10%, 50% and 100%) of a plasmid with a known variant (or another distinguishable sequence tag) could be added into the library plasmid pool to normalise the barcode counts by the number of molecules present in the library.

However, this is time-consuming and may be only feasible for low complexity libraries. To avoid enzyme bias, heat and divalent metal cation[803], acoustic and hydrodynamic DNA shearing[804] as well as sonication or DNaseI non-specific nuclease[805], phage Mu (Buschman's laboratory)[806] or nrLM-PCR (Schmidt's laboratory)[807] can be used as alternatives to fragment DNA.

Linked to the complexity of the library, a balance needs to be found between the throughput requirements and Leveinshtein or editing distance for each individual application. A high complexity library is not always recommended as it can compromise the accuracy of its variant distinction; it depends on the complexity required or the proportion of the theoretical space occupied by the actual library. Library complexity is a critical parameter in studies involving barcoded experiments. However, a uniform criteria does not exist for the estimation of library complexities; diversity indexes such as the Schnabel, Lincoln Petersen (used in this study), Shannon-Weaver (used in Porter *et al.*,)[736], Simpson, Berger Parker and their respective modifications contrast with descriptive frequency plots or simply no control over this variable.

Another potential major objection to this novel marking approach might be its inability to screen for genomic positions with high associated expression due to post-translational modifications. Therefore, this system would not be sensitive enough to detect epigenetic changes that could affect clonal fitness understood as growth rate, sensitivity to lactate/ammonia and tolerance to the heterologous expression burden. However, environmental factors are not the only parameters to affect expression. Internal processes like mRNA export and its processing (capping, splicing and polyadenylation) regulate the stability and decay of the mRNA. Translation regulation mechanisms include the availability of eIF2-GTP-Met-tRNA$_i^{Met}$ ternary complexes (necessary for translation initiation)[94], upregulation by upstream ORFs (uORFS)[808], AU-rich elements[809], or downregulation by interference RNAs or inhibitory proteins (or combinations of them). In the context of lentiviral vectors there are other steps in assembly and budding that can influence the infectivity of the vector. However, the main objective of this project was to increase the transcription of viral genome and thus post-transcriptional modifications are not critical. In the event of using this system to improve the expression of packaging genes, post-transcriptional modifications should be considered.

While analysis of clonal dynamics is critical for the characterisation of different lineages, in this study barcode proportions are analysed at a particular time point (cells are harvested one week after transduction). However, expression dynamics could be assessed to determine whether barcode abundance (and thus gene expression) is stable over time and also to discard false barcodes (background sequencing noise). A potential alternative method to be explored to determine site-specific expression would be to examine read through barcoded transcripts that have extended into the genome from RNA-Seq and have not terminated at the polyadenylation site. Another advantage of tracking clonal dynamics of a particular barcode variant would be the information provided about the stability of transgene expression. In this study, despite being cultured for 4-5 weeks and during the genome editing process, the lack of stability assessment could be a possible criticism to the strategy followed.

As a quantitative method for transcript expression, the random integration events observed in the candidate positions preclude drawing any conclusions about the correlation between the barcode counts obtained in EGFEM1P and CUL5 positions with their basal expression and eventually the titers resulting from the integration of the transfer vector. Noise discrimination experiments should be planned to determine its specificity, recovery rate, precision (repeatability) and establish a linearity range in which the method is fit for purpose.

One of the objectives pursued with the barcode screening was the reduction of cell line development timelines. However, the intrinsic variability in phenotype (copy number and consequently GFP intensity) observed between clones after the genome edition process and the need for screening questions that argument. In that matter, the advantage of the CRISPR-Cas9 technology compared to other editing technologies is that several positions can be targeted at the same time in the event of simultaneous integration of several lentiviral components.

A further criticism is that transcription was measured from the internal EFS promoter instead of the 5’LTR promoter, which might be more predictive of a producer cell line. To explore that possibility, the barcode should have been located immediately downstream of the RSV promoter, which might be incompatible with the sequencing of vector-host DNA junctions for integration site analysis.

Overall, despite the criticisms, the described method could represent an improved method of cell line development in terms of screening and amplification, decreasing time lines and adjusting costs. In addition, the application of these findings would not only be limited to seek high-producers. Low-producer clones could be used to identify where to integrate host-cell modifying enzymes that might confer a faster growth, increased protein production and secretion, lower rates of cell death (apoptosis resistance), resistance to any toxicity observed from overexpression of proteins (or viral vectors), serum independence or an innate resistance to bacterial or fungal contamination. In this way, these factors could be constitutively expressed in packaging cell lines at a level that is not toxic to the cells and would not reduce vector production.

Viral integration and packaging/producer cell line development

An alternative approach to the rational screening strategy described in this study could employ viral transduction as opposed to random transfection for the separate delivery of lentiviral components and has been used by several authors in the last years[385,395,396,398,400,456,460]. Separate delivery of the packaging constructs (gag-pol, rev, VSV-G) into the host cell line is advised to avoid generation of RCL. Providing that selection and screening needs to be performed in any case to select for successful integration events, viral delivery provides an efficient means of delivery and semi-random integration into actively transcribed regions. MOI can be adjusted if more copies of the gene are required (Figure 3.11). Integration profiles (alpha, gamma- lenti-), other elements (WPRE, S/MARs) as well as different promoters can help modulate expression of the transgene.

In terms of envelope proteins, although an eventual lentiviral platform should be open to accommodate different envelope proteins, the wide usage and multiple advantages of VSV-G makes it a good candidate for an inducible PCL.

Among inducible systems, "inducer-on" present advantages over "inducer-off" (explained in Section 1.1.6) and in particular double switch systems such as the cumate switch reported by Broussau *et al.*, have demonstrated relatively high titers (10^7 TU/mL) and sustained expression for 18 weeks.

The delivery of the SIN lentiviral transfer vector is a more cumbersome aspect. As highlighted in the introduction of this study, stoichiometric limitation of transfer vector is a critical factor for lentiviral production[776]. In this study, that requirement was confirmed by the fact that only measurable titers from cells with >1 proviral copies/cell were obtained. Production of clinical grade SIN lentiviral transfer vectors would benefit from stable PCL that enable generation of large volumes of vector. Viral delivery of transfer vectors enable generation of populations with stable expression derived from actively transcribed sites, which offers advantages compared to stable transfection. However, SIN LVV cannot be delivered using SIN-LVV due to the inactivated U3 in the 3'LTR (one round of replication). Initial attempts of lentiviral delivery of transfer vectors such as Kafri *et al.*, Klages *et al.*, or Kuate *et al.*, used non-SIN vectors tat dependent (similar to Ikeda *et al.*, and Ni *et al.*, [385,400]), chimeric *tat* independent LTRs (HIV R and U5 and RSV U3/CMV promoter in 5'/3' LTR) or LTR from other species (SIV), respectively[398,399,459]. However, mobilisation of vector into RCL in target cells upon delivery of *gag-pol* and *env* genes supposes a safety concern[811,812]. Although this problem was partly solved with cSIN vectors (with TRE regulatory regions in the 3'LTR U3 region), the genotoxicity of these vectors have not been extensively tested[401,463]. The alternative is stable co-transfection of SIN transfer vectors although further optimisation of the method is required and productivity drops 40-fold[385]. In 2015, Sanber *et al.*, used a two-step approach to target recombinase recognition sites into actively transcribed regions using viral transduction and subsequently induce recombination of transfected transfer vector constructs[381]. However, that strategy involves two more rounds of selection, which might delay production timelines.

Delivery of full SIN vectors using SIN (non-cSIN) LVV transduction remains a challenge. SIN lenti-delivery of SIN vectors could provide a more efficient, stable and reproducible way to increase transfer vector expression than Throm's concatemeric array (currently the approach of sustained production with higher titers). Integration preferences of lentiviral vectors would offer a one-step delivery of the vector into a stable, high-expressing locus and would also meet all the safety requirements. The eventual design should overcome the limitations of inactivated vectors encountered during reverse transcription and enable two rounds of replication.

The promising results of gene therapy using lentiviral vectors as a safe integrating transgene delivery method envisage a future requirement for larger volumes of vector when the treatment of diseases with increasing incidences are developed as commercial biopharmaceutical products. The generation of lentiviral producer cell lines that yield high titers is thus crucial to enable efficient, long-lasting and accesible therapies to the patients.

References

1. International Congress of Genetics. *Proceedings of the Sixth International Congress of Genetics*. (1932).
2. McCarty, M. & Avery, O. T. STUDIES ON THE CHEMICAL NATURE OF THE SUBSTANCE INDUCING TRANSFORMATION OF PNEUMOCOCCAL TYPES : II. EFFECT OF DESOXYRIBONUCLEASE ON THE BIOLOGICAL ACTIVITY OF THE TRANSFORMING SUBSTANCE. *J. Exp. Med.* 83, 89–96 (1946).
3. WATSON, J. D. & CRICK, F. H. Genetical implications of the structure of deoxyribonucleic acid. *Nature* 171, 964–967 (1953).
4. Franklin RE, G. R. Molecular structure of nucleic acids. Molecular configuration in sodium thymonucleate. *Nature* 171, 740–741 (1953).
5. ZINDER, N. D. & LEDERBERG, J. Genetic exchange in Salmonella. *J. Bacteriol.* 64, 679–699 (1952).
6. Rogers, S., Lowenthal, A., Terheggen, H. G. & Columbo, J. P. Induction of arginase activity with the Shope papilloma virus in tissue culture cells from an argininemic patient. *J. Exp. Med.* 137, 1091–1096 (1973).
7. Friedmann, T. & Roblin, R. Gene Therapy for Human Genetic Disease? *Science* 175, 949–955 (1972).
8. Terheggen, H. G., Lowenthal, A., Lavinha, F., Colombo, J. P. & Rogers, S. Unsuccessful trial of gene replacement in arginase deficiency. *Z. Kinderheilkd.* 119, 1–3 (1975).
9. Graham, F. L. & van der Eb, A. J. A new technique for the assay of infectivity of human adenovirus 5 DNA. *Virology* 52, 456–467 (1973).
10. Mulligan, R. C. & Berg, P. Selection for animal cells that express the Escherichia coli gene coding for xanthine-guanine phosphoribosyltransferase. *Proc. Natl. Acad. Sci. U. S. A.* 78, 2072–2076 (1981).
11. Anderson, W. F. & Fletcher, J. C. Sounding boards. Gene therapy in human beings: when is it ethical to begin? *N. Engl. J. Med.* 303, 1293–1297 (1980).
12. Cline, M. J. *et al.* Gene transfer in intact animals. *Nature* 284, 422–425 (1980).
13. Wade, N. UCLA gene therapy racked by friendly fire. *Science (New York, N.Y.)* 210, 509–511 (1980).
14. Wade, N. Gene therapy caught in more entanglements. *Science (New York, N.Y.)* 212, 24–25 (1981).
15. Baltimore, D. RNA-dependent DNA polymerase in virions of RNA tumour viruses. *Nature* 226, 1209–1211 (1970).
16. Temin, H. M. & Mizutani, S. RNA-dependent DNA polymerase in virions of Rous sarcoma virus. *Nature* 226, 1211–1213 (1970).
17. Willis, R. C. *et al.* Partial phenotypic correction of human Lesch-Nyhan (hypoxanthine-guanine phosphoribosyltransferase-deficient) lymphoblasts with a transmissible retroviral vector. *J. Biol. Chem.* 259, 7842–7849 (1984).
18. Miller, A. D., Jolly, D. J., Friedmann, T. & Verma, I. M. A transmissible retrovirus expressing human hypoxanthine phosphoribosyltransferase (HPRT): gene transfer into cells obtained from humans deficient in HPRT. *Proc. Natl. Acad. Sci. U. S. A.* 80, 4709–4713 (1983).
19. Kantoff, P. W. *et al.* Correction of adenosine deaminase deficiency in cultured human T and B cells by retrovirus-mediated gene transfer. *Proc. Natl. Acad. Sci. U. S. A.* 83, 6563–6567 (1986).
20. Rosenberg, S. A. *et al.* Gene transfer into humans--immunotherapy of patients with advanced melanoma, using tumor-infiltrating lymphocytes modified by retroviral gene transduction. *N. Engl. J. Med.* 323, 570–578 (1990).
21. Blaese, R. M. *et al. T lymphocyte-directed gene therapy for ADA- SCID: initial trial results after 4 years. Science (New York, N.Y.)* 270, (1995).
22. Bordignon, C. *et al.* Gene therapy in peripheral blood lymphocytes and bone marrow for ADA- immunodeficient patients. *Science* 270, 470–475 (1995).
23. Fox, J. L. Gene-therapy death prompts broad civil lawsuit. *Nature biotechnology* 18, 1136 (2000).
24. Hacein-Bey-Abina, S. *et al.* Efficacy of gene therapy for X-linked severe combined immunodeficiency. *N. Engl. J. Med.* 363, 355–364 (2010).

25. Gaspar, H. B. *et al.* Gene therapy of X-linked severe combined immunodeficiency by use of a pseudotyped gammaretroviral vector. *Lancet (London, England)* 364, 2181–2187 (2004).
26. Aiuti, A. *et al.* Correction of ADA-SCID by stem cell gene therapy combined with nonmyeloablative conditioning. *Science* 296, 2410–2413 (2002).
27. Aiuti, A. *et al. Gene therapy for immunodeficiency due to adenosine deaminase deficiency. The New England journal of medicine* 360, (2009).
28. Gaspar, H. B. *et al.* Successful reconstitution of immunity in ADA-SCID by stem cell gene therapy following cessation of PEG-ADA and use of mild preconditioning. *Mol. Ther.* 14, 505–513 (2006).
29. Ott, M. G. *et al.* Correction of X-linked chronic granulomatous disease by gene therapy, augmented by insertional activation of MDS1-EVI1, PRDM16 or SETBP1. *Nat. Med.* 12, 401–409 (2006).
30. Cartier, N. *et al.* Hematopoietic stem cell gene therapy with a lentiviral vector in X-linked adrenoleukodystrophy. *Science* 326, 818–823 (2009).
31. Aiuti, A. *et al.* Lentiviral hematopoietic stem cell gene therapy in patients with Wiskott-Aldrich syndrome. *Science* 341, 1233151 (2013).
32. Biffi, A. *et al.* Lentiviral hematopoietic stem cell gene therapy benefits metachromatic leukodystrophy. *Science* 341, 1233158 (2013).
33. Bartus, R. T., Weinberg, M. S. & Samulski, R. J. Parkinson's disease gene therapy: success by design meets failure by efficacy. *Mol. Ther.* 22, 487–97 (2014).
34. Schimmer, J. & Breazzano, S. Investor Outlook: Rising from the Ashes; GSK's European Approval of Strimvelis for ADA-SCID. *Hum. Gene Ther. Clin. Dev.* 27, 57–61 (2016).
35. Deev, R. V *et al.* pCMV-vegf165 Intramuscular Gene Transfer is an Effective Method of Treatment for Patients With Chronic Lower Limb Ischemia. *J. Cardiovasc. Pharmacol. Ther.* 20, 473–482 (2015).
36. Yin, H. *et al.* Non-viral vectors for gene-based therapy. *Nat. Rev. Genet.* 15, 541–555 (2014).
37. Putnam, D. Polymers for gene delivery across length scales. *Nat. Mater.* 5, 439–451 (2006).
38. Yeh, P. & Perricaudet, M. Advances in adenoviral vectors: from genetic engineering to their biology. *FASEB J.* 11, 615–623 (1997).
39. Raper, S. E. *et al.* Developing adenoviral-mediated in vivo gene therapy for ornithine transcarbamylase deficiency. in *Journal of Inherited Metabolic Disease* 21, 119–137 (1998).
40. Kim, M. Replicating poxviruses for human cancer therapy. *J. Microbiol.* 53, 209–218 (2015).
41. Breitbach, C. J., Thorne, S. H., Bell, J. C. & Kirn, D. H. Targeted and armed oncolytic poxviruses for cancer: the lead example of JX-594. *Curr. Pharm. Biotechnol.* 13, 1768–1772 (2012).
42. Boehmer, P. E. & Lehman, I. R. Herpes simplex virus DNA replication. *Annu. Rev. Biochem.* 66, 347–384 (1997).
43. Miller, A. D. Identification and Elimination of Replication-Competent Adeno-Associated Virus (AAV) That Can Arise by Nonhomologous Recombination during AAV Vector Production. *Microbiology* 71, 6816–6822 (1997).
44. Flotte, T. R. *et al. Phase I trial of intranasal and endobronchial administration of a recombinant adeno-associated virus serotype 2 (rAAV2)-CFTR vector in adult cystic fibrosis patients: a two-part clinical study. Human gene therapy* 14, (2003).
45. Simonelli, F. *et al. Gene therapy for Leber's congenital amaurosis is safe and effective through 1.5 years after vector administration. Molecular therapy : the journal of the American Society of Gene Therapy* 18, (2010).
46. MacLaren, R. E. *et al.* Retinal gene therapy in patients with choroideremia: initial findings from a phase 1/2 clinical trial. *Lancet* 383, 1129–37 (2014).
47. Hasbrouck, N. C. & High, K. A. AAV-mediated gene transfer for the treatment of hemophilia B: problems and prospects. *Gene Ther.* 15, 870–875 (2008).
48. Louis Jeune, V., Joergensen, J. A., Hajjar, R. J. & Weber, T. Pre-existing anti-adeno-associated virus antibodies as a challenge in AAV gene therapy. *Hum. Gene Ther. Methods* 24, 59–67 (2013).
49. Nault, J.-C. *et al.* Recurrent AAV2-related insertional mutagenesis in human hepatocellular carcinomas. *Nat. Genet.* 47, 1–15 (2015).
50. Kraunus, J. *et al.* Murine leukemia virus regulates alternative splicing through sequences upstream of the 5??? splice site. *J. Biol. Chem.* 281, 37381–37390 (2006).

51. Pessel-Vivares, L., Houzet, L., Lainé, S. & Mougel, M. Insights into the nuclear export of murine leukemia virus intron-containing RNA. *RNA Biol.* 12, 942–9 (2015).
52. Cavazza, A., Moiani, A. & Mavilio, F. Mechanisms of retroviral integration and mutagenesis. *Hum. Gene Ther.* 24, 119–31 (2013).
53. Li, C. L., Xiong, D., Stamatoyannopoulos, G. & Emery, D. W. Genomic and functional assays demonstrate reduced gammaretroviral vector genotoxicity associated with use of the cHS4 chromatin insulator. *Mol. Ther.* 17, 716–724 (2009).
54. Meiering, C. D. & Linial, M. L. Historical perspective of foamy virus epidemiology and infection. *Clin. Microbiol. Rev.* 14, 165–76 (2001).
55. Jackson, D. L., Lee, E.-G. & Linial, M. L. Expression of prototype foamy virus pol as a Gag-Pol fusion protein does not change the timing of reverse transcription. *J. Virol.* 87, 1252–4 (2013).
56. Trobridge, G. & Russell, D. W. Cell cycle requirements for transduction by foamy virus vectors compared to those of oncovirus and lentivirus vectors. *J. Virol.* 78, 2327–35 (2004).
57. Bodem, J., Schied, T., Gabriel, R., Rammling, M. & Rethwilm, A. Foamy virus nuclear RNA export is distinct from that of other retroviruses. *J Virol* 85, 2333–2341 (2011).
58. Löchelt, M. *et al.* The antiretroviral activity of APOBEC3 is inhibited by the foamy virus accessory Bet protein. *Proc. Natl. Acad. Sci. U. S. A.* 102, 7982–7 (2005).
59. Venkatesh, L. K., Theodorakis, P. A. & Chinnadurai, G. Distinct cis-acting regions in U3 regulate trans-activation of the human spumaretrovirus long terminal repeat by the viral bel1 gene product. *Nucleic Acids Res.* 19, 3661–3666 (1991).
60. Olszko, M. E. & Trobridge, G. D. Foamy virus vectors for HIV gene therapy. *Viruses* 5, 2585–600 (2013).
61. Trobridge, G., Josephson, N., Vassilopoulos, G., Mac, J. & Russell, D. W. Improved foamy virus vectors with minimal viral sequences. *Mol.Ther* 6, 321–328 (2002).
62. Suerth, J. D., Maetzig, T., Galla, M., Baum, C. & Schambach, A. Self-inactivating alpharetroviral vectors with a split-packaging design. *J. Virol.* 84, 6626–35 (2010).
63. Stewart, H. J., Leroux-Carlucci, M. a, Sion, C. J. M., Mitrophanous, K. a & Radcliffe, P. a. Development of inducible EIAV-based lentiviral vector packaging and producer cell lines. *Gene Ther.* 16, 805–814 (2009).
64. Leroux, C., Cadoré, J. L. & Montelaro, R. C. Equine Infectious Anemia Virus (EIAV): What has HIV's country cousin got tell us? *Veterinary Research* 35, 485–512 (2004).
65. Shimojima, M. *et al.* Use of CD134 as a primary receptor by the feline immunodeficiency virus. *Science* 303, 1192–5 (2004).
66. Oberste, M. S., Greenwood, J. D. & Gonda, M. A. Analysis of the transcription pattern and mapping of the putative rev and env splice junctions of bovine immunodeficiency-like virus. *J Virol* 65, 3932–7. (1991).
67. Dietrich, I. *et al.* Feline tetherin efficiently restricts release of feline immunodeficiency virus but not spreading of infection. *J. Virol.* 85, 5840–5852 (2011).
68. De Rijck, J. & Debyser, Z. The central DNA flap of the human immunodeficiency virus type 1 is important for viral replication. *Biochem. Biophys. Res. Commun.* 349, 1100–1110 (2006).
69. Mamede, J. I., Sitbon, M., Battini, J.-L. & Courgnaud, V. Heterogeneous susceptibility of circulating SIV isolate capsids to HIV-interacting factors. *Retrovirology* 10, p1–15. 15p. (2013).
70. Naldini, L. *et al.* In vivo gene delivery and stable transduction of nondividing cells by a lentiviral vector. *Science* 272, 263–7 (1996).
71. Naldini, L. Lentiviruses as gene transfer agents for delivery to non-dividing cells. *Curr. Opin. Biotechnol.* 9, 457–63 (1998).
72. Coffin, J. M., Hughes, S. H. & Varmus, H. E. in *Retroviruses* (1997).
73. Huthoff, H., Bugala, K., Barciszewski, J. & Berkhout, B. On the importance of the primer activation signal for initiation of tRNA(lys3)-primed reverse transcription of the HIV-1 RNA genome. *Nucleic Acids Res.* 31, 5186–94 (2003).
74. Verma, I. M., Meuth, N. L., Bromfeld, E., Manly, K. F. & Baltimore, D. Covalently linked RNA-DNA molecule as initial product of RNA tumour virus DNA polymerase. *Nat. New Biol.* 233, 131–4 (1971).
75. Dahlberg, J. E. *et al.* Transcription of DNA from the 70S RNA of Rous sarcoma virus. I. Identification of a specific 4S RNA which serves as primer. *J. Virol.* 13, 1126–1133 (1974).

76. Mann, R. & Baltimore, D. Varying the position of a retrovirus packaging sequence results in the encapsidation of both unspliced and spliced RNAs. *J. Virol.* 54, 401–7 (1985).

77. Charneau, P. *et al.* HIV-1 reverse transcription. A termination step at the center of the genome. *Journal of molecular biology* 241, 651–662 (1994).

78. Finston, W. I. & Champoux, J. J. RNA-primed initiation of Moloney murine leukemia virus plus strands by reverse transcriptase in vitro. *J. Virol.* 51, 26–33 (1984).

79. Emerman, M. & Malim, M. H. HIV-1 regulatory/accessory genes: keys to unraveling viral and host cell biology. *Science* 280, 1880–4 (1998).

80. Frankel, A. D. & Young, J. A. HIV-1: fifteen proteins and an RNA. *Annu. Rev. Biochem.* 67, 1–25 (1998).

81. Zhou, M. *et al.* The Tat/TAR-dependent phosphorylation of RNA polymerase II C-terminal domain stimulates cotranscriptional capping of HIV-1 mRNA. *Proc. Natl. Acad. Sci. U. S. A.* 100, 12666–71 (2003).

82. Varmus H, S. R. *Replication of retroviruses. Tumour Viruses: Molecular Biology of Tumor Viruses.* (Cold Spring Harbor Laboratory, 1984).

83. Van Lint, C., Bouchat, S. & Marcello, A. HIV-1 transcription and latency: an update. *Retrovirology* 10, 67 (2013).

84. Pierson, T. C. *et al.* Molecular Characterization of Preintegration Latency in Human Immunodeficiency Virus Type 1 Infection. *J. Virol.* 76, 8518–8531 (2002).

85. Mbonye, U. & Karn, J. Control of HIV latency by epigenetic and non-epigenetic mechanisms. *Curr. HIV Res.* 9, 554–67 (2011).

86. Gilmartin, G. M., Fleming, E. S., Oetjen, J. & Graveley, B. R. CPSF recognition of an HIV-1 mRNA 3′ -processing enhancer: Multiple sequence contacts involved in poly(A) site definition. *Genes Dev.* 9, 72–83 (1995).

87. Das, A. T., Klaver, B. & Berkhout, B. A hairpin structure in the R region of the human immunodeficiency virus type 1 RNA genome is instrumental in polyadenylation site selection. *J Virol* 73, 81–91 (1999).

88. Ashe, M. P., Furger, A. & Proudfoot, N. J. Stem-loop 1 of the U1 snRNP plays a critical role in the suppression of HIV-1 polyadenylation. *RNA* 6, 170–7 (2000).

89. Cullen, B. R. Mechanism of action of regulatory proteins encoded by complex retroviruses. *Microbiol. Rev.* 56, 375–394 (1992).

90. Cullen, B. R. Nuclear mRNA export: insights from virology. *Trends Biochem. Sci.* 28, 419–24 (2003).

91. Fritz, C. C. & Green, M. R. HIV Rev uses a conserved cellular protein export pathway for the nucleocytoplasmic transport of viral RNAs. *Curr. Biol.* 6, 848–54 (1996).

92. Elfgang, C. *et al.* Evidence for specific nucleocytoplasmic transport pathways used by leucine-rich nuclear export signals. *Proc. Natl. Acad. Sci. U. S. A.* 96, 6229–34 (1999).

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

94. Jackson, R. J., Hellen, C. U. T. & Pestova, T. V. The mechanism of eukaryotic translation initiation and principles of its regulation. *Nat. Rev. Mol. Cell Biol.* 11, 113–127 (2010).

95. Chamond, N., Locker, N. & Sargueil, B. The different pathways of HIV genomic RNA translation. *Biochem. Soc. Trans.* 38, 1548–52 (2010).

96. de Breyne, S., Soto-Rifo, R., López-Lastra, M. & Ohlmann, T. Translation initiation is driven by different mechanisms on the HIV-1 and HIV-2 genomic RNAs. *Virus Research* (2012). doi:10.1016/j.virusres.2012.10.006

97. Willey, R. L., Shibata, R., Freed, E. O., Cho, M. W. & Martin, M. A. Differential glycosylation, virion incorporation, and sensitivity to neutralizing antibodies of human immunodeficiency virus type 1 envelope produced from infected primary T-lymphocyte and macrophage cultures. *J. Virol.* 70, 6431–6 (1996).

98. Brierley, I. & Dos Ramos, F. J. Programmed ribosomal frameshifting in HIV-1 and the SARS-CoV. *Virus Res.* 119, 29–42 (2006).

99. Cassan, M., Delaunay, N., Vaquero, C. & Rousset, J. P. Translational frameshifting at the gag-pol junction of human immunodeficiency virus type 1 is not increased in infected T-lymphoid cells. *J. Virol.* 68, 1501–1508 (1994).

100. Shehu-xhilaga, M. & Crowe, S. M. Maintenance of the Gag / Gag-Pol Ratio Is Important for Human Immunodeficiency Virus Type 1 RNA Dimerization and Viral Infectivity. 75, 1834–

1841 (2001).

101. Miller, J. H., Presnyak, V. & Smith, H. C. The dimerization domain of HIV-1 viral infectivity factor Vif is required to block virion incorporation of APOBEC3G. *Retrovirology* 4, 81 (2007).

102. Fujita, M., Nomaguchi, M., Adachi, A. & Otsuka, M. SAMHD1-Dependent and -Independent Functions of HIV-2/SIV Vpx Protein. *Front. Microbiol.* 3, 297 (2012).

103. Bukrinsky, M. & Adzhubei, A. Viral protein R of HIV-1. *Rev. Med. Virol.* 9, 39–49

104. Bour, S., Schubert, U. & Strebel, K. The human immunodeficiency virus type 1 Vpu protein specifically binds to the cytoplasmic domain of CD4: implications for the mechanism of degradation. *J. Virol.* 69, 1510–20 (1995).

105. Das, S. R. & Jameel, S. Biology of the HIV Nef protein. *Indian Journal of Medical Research* 121, 315–332 (2005).

106. Douglas, J. L. *et al.* Vpu directs the degradation of the human immunodeficiency virus restriction factor BST-2/Tetherin via a {beta}TrCP-dependent mechanism. *J. Virol.* 83, 7931–47 (2009).

107. Ivanchenko, S. *et al.* Dynamics of HIV-1 assembly and release. *PLoS Pathog.* 5, (2009).

108. Göttlinger, H. G., Sodroski, J. G. & Haseltine, W. a. Role of capsid precursor processing and myristoylation in morphogenesis and infectivity of human immunodeficiency virus type 1. *Proc. Natl. Acad. Sci. U. S. A.* 86, 5781–5785 (1989).

109. Jin, J., Sturgeon, T., Weisz, O. A., Mothes, W. & Montelaro, R. C. HIV-1 matrix dependent membrane targeting is regulated by Gag mRNA trafficking. *PLoS One* 4, (2009).

110. Sundquist, W. I. & Kräusslich, H. G. HIV-1 assembly, budding, and maturation. *Cold Spring Harbor Perspectives in Medicine* 2, (2012).

111. Bryant, M. & Ratner, L. Myristoylation-dependent replication and assembly of human immunodeficiency virus 1. *Proc. Natl. Acad. Sci. U. S. A.* 87, 523–7 (1990).

112. Huang, Y. *et al.* Incorporation of excess wild-type and mutant tRNA(3Lys) into human immunodeficiency virus type 1. *J. Virol.* 68, 7676–83 (1994).

113. Cen, S. *et al.* Retrovirus-specific packaging of aminoacyl-tRNA synthetases with cognate primer tRNAs. *J. Virol.* 76, 13111–5 (2002).

114. Paxton, W., Connor, R. I. & Landau, N. R. Incorporation of Vpr into human immunodeficiency virus type 1 virions: requirement for the p6 region of gag and mutational analysis. *J. Virol.* 67, 7229–37 (1993).

115. Abeydeera, N. D. *et al.* Evoking picomolar binding in RNA by a single phosphorodithioate linkage. *Nucleic Acids Res.* 44, 8052–8064 (2016).

116. Wyatt, R. T. & Sodroski, J. The HIV-1 envelope glycoproteins: fusogens, antigens, and immunogens. *Science (80-.).* 280, 1884–1888 (1998).

117. Arthos, J. *et al.* HIV-1 envelope protein binds to and signals through integrin alpha4beta7, the gut mucosal homing receptor for peripheral T cells. *Nat. Immunol.* 9, 301–309 (2008).

118. Geijtenbeek, T. B. *et al.* DC-SIGN, a dendritic cell-specific HIV-1-binding protein that enhances trans-infection of T cells. *Cell* 100, 587–97 (2000).

119. Chen, B. *et al.* Structure of an unliganded simian immunodeficiency virus gp120 core. *Nature* 433, 834–841 (2005).

120. Rizzuto, C. D. *et al.* A conserved HIV gp120 glycoprotein structure involved in chemokine receptor binding. *Science* 280, 1949–1953 (1998).

121. Chan, D. C., Fass, D., Berger, J. M. & Kim, P. S. Core structure of gp41 from the HIV envelope glycoprotein. *Cell* 89, 263–73 (1997).

122. Fassati, A. & Goff, S. P. Characterization of intracellular reverse transcription complexes of human immunodeficiency virus type 1. *J. Virol.* 75, 3626–35 (2001).

123. Hulme, A. E., Perez, O. & Hope, T. J. Complementary assays reveal a relationship between HIV-1 uncoating and reverse transcription. *Proc. Natl. Acad. Sci. U. S. A.* 108, 9975–80 (2011).

124. Rasaiyaah, J. *et al.* HIV-1 evades innate immune recognition through specific cofactor recruitment. *Nature* 503, 402–5 (2013).

125. Lahaye, X. *et al.* 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, 1132–42 (2013).

126. Peng, K. *et al.* Quantitative microscopy of functional HIV post-entry complexes reveals association of replication with the viral capsid. *Elife* 3, e04114 (2014).

127. Hatziioannou, T., Perez-Caballero, D., Cowan, S. & Bieniasz, P. D. Cyclophilin interactions

with incoming human immunodeficiency virus type 1 capsids with opposing effects on infectivity in human cells. *J. Virol.* 79, 176–83 (2005).

128. Lukacs, G. L. *et al.* Size-dependent DNA mobility in cytoplasm and nucleus. *J. Biol. Chem.* 275, 1625–1629 (2000).

129. Vaughan, J. C., Brandenburg, B., Hogle, J. M. & Zhuang, X. Rapid actin-dependent viral motility in live cells. *Biophys. J.* 97, 1647–1656 (2009).

130. McDonald, D. *et al.* Visualization of the intracellular behavior of HIV in living cells. *J. Cell Biol.* 159, 441–52 (2002).

131. Suikkanen, S. *et al.* Exploitation of microtubule cytoskeleton and dynein during parvoviral traffic toward the nucleus. *J. Virol.* 77, 10270–9 (2003).

132. Ward, B. M. Visualization and characterization of the intracellular movement of vaccinia virus intracellular mature virions. *J. Virol.* 79, 4755–63 (2005).

133. Kohlstaedt, L. A., Wang, J., Friedman, J. M., Rice, P. A. & Steitz, T. A. Crystal structure at 3.5 A resolution of HIV-1 reverse transcriptase complexed with an inhibitor. *Science (80-.).* 256, 1783–1790 (1992).

134. Telesnitsky, A. & Goff, S. P. Two defective forms of reverse transcriptase can complement to restore retroviral infectivity. *EMBO J* 12, 4433–4438 (1993).

135. Hu, W.-S. & Hughes, S. H. HIV-1 reverse transcription. *Cold Spring Harb. Perspect. Med.* 2, a006882- (2012).

136. Kovaleski, B. J. *et al.* In vitro characterization of the interaction between HIV-1 Gag and human Lysyl-tRNA synthetase. *J. Biol. Chem.* 281, 19449–19456 (2006).

137. Fuentes, G. M., Rodriguez-Rodriguez, L., Fay, P. J. & Bambara, R. A. Use of an oligoribonucleotide containing the polypurine tract sequence as a primer by HIV reverse transcriptase. *J. Biol. Chem.* 270, 28169–28176 (1995).

138. Smith, C. M., Potts, W. B., Smith, J. S. & Roth, M. J. RNase H cleavage of tRNAPro mediated by M-MuLV and HIV-1 reverse transcriptases. *Virology* 229, 437–46 (1997).

139. Shaharabany, M., Rice, N. R. & Hizi, A. Expression and mutational analysis of the reverse transcriptase of the lentivirus equine infectious anemia virus. *Biochem Biophys Res Commun* 196, 914–920 (1993).

140. Renda, M. J. *et al.* Mutation of the methylated tRNA(Lys)(3) residue A58 disrupts reverse transcription and inhibits replication of human immunodeficiency virus type 1. *J. Virol.* 75, 9671–8 (2001).

141. Woodward, C. L., Prakobwanakit, S., Mosessian, S. & Chow, S. a. Integrase interacts with nucleoporin NUP153 to mediate the nuclear import of human immunodeficiency virus type 1. *J. Virol.* 83, 6522–6533 (2009).

142. Bouyac-Bertoia, M. *et al.* HIV-1 Infection Requires a Functional Integrase NLS. *Mol. Cell* 7, 1025–1035 (2001).

143. Jenkins, Y., McEntee, M., Weis, K. & Greene, W. C. Characterization of HIV-1 Vpr nuclear import: Analysis of signals and pathways. *J. Cell Biol.* 143, 875–885 (1998).

144. Haffar, O. K. *et al.* Two nuclear localization signals in the HIV-1 matrix protein regulate nuclear import of the HIV-1 pre-integration complex. *J. Mol. Biol.* 299, 359–68 (2000).

145. Zennou, V. *et al.* HIV-1 genome nuclear import is mediated by a central DNA flap. *Cell* 101, 173–185 (2000).

146. Dull, T. *et al.* A third-generation lentivirus vector with a conditional packaging system. *J. Virol.* 72, 8463–8471 (1998).

147. Rivière, L., Darlix, J.-L. & Cimarelli, A. Analysis of the viral elements required in the nuclear import of HIV-1 DNA. *J. Virol.* 84, 729–39 (2010).

148. Yamashita, M. & Emerman, M. The cell cycle independence of HIV infections is not determined by known karyophilic viral elements. *PLoS Pathog.* 1, 0170–0178 (2005).

149. Yamashita, M. & Emerman, M. Capsid Is a Dominant Determinant of Retrovirus Infectivity in Nondividing Cells. *J. Virol.* 78, 5670–5678 (2004).

150. Krishnan, L. *et al.* The requirement for cellular transportin 3 (TNPO3 or TRN-SR2) during infection maps to human immunodeficiency virus type 1 capsid and not integrase. *J. Virol.* 84, 397–406 (2010).

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

152. Gallay, P., Stitt, V., Mundy, C., Oettinger, M. & Trono, D. Role of the karyopherin pathway in human immunodeficiency virus type 1 nuclear import. *J. Virol.* 70, 1027–32 (1996).

153. Zaitseva, L. *et al.* HIV-1 exploits importin 7 to maximize nuclear import of its DNA genome. *Retrovirology* 6, 11 (2009).

154. Christ, F. *et al.* Transportin-SR2 Imports HIV into the Nucleus. *Curr. Biol.* 18, 1192–1202 (2008).

155. Schaller, T. *et al.* HIV-1 capsid-cyclophilin interactions determine nuclear import pathway, integration targeting and replication efficiency. *PLoS Pathog.* 7, (2011).

156. Zhang, Z. H. *et al.* A comparative study of techniques for differential expression analysis on RNA-seq data. *PLoS One* 9, (2014).

157. Ocwieja, K. E. *et al.* HIV integration targeting: A pathway involving transportin-3 and the nuclear pore protein RanBP2. *PLoS Pathog.* 7, (2011).

158. Engelman, A. & Cherepanov, P. The structural biology of HIV-1: mechanistic and therapeutic insights. *Nat. Rev. Microbiol.* 10, 279–290 (2012).

159. Fitzgerald, M. L., Vora, a C., Zeh, W. G. & Grandgenett, D. P. Concerted integration of viral DNA termini by purified avian myeloblastosis virus integrase. *J. Virol.* 66, 6257–6263 (1992).

160. Quinn, T. P. & Grandgenett, D. P. Genetic evidence that the avian retrovirus DNA endonuclease domain of pol is necessary for viral integration. *J. Virol.* 62, 2307–2312 (1988).

161. Schiff, R. D. & Grandgenett, D. P. Virus-coded origin of a 32,000-dalton protein from avian retrovirus cores: structural relatedness of p32 and the beta polypeptide of the avian retrovirus DNA polymerase. *J. Virol.* 28, 279–291 (1978).

162. Rice, P., Craigie, R. & Davies, D. R. Retroviral integrases and their cousins. *Current Opinion in Structural Biology* 6, 76–83 (1996).

163. Eijkelenboom, A. P. *et al.* The solution structure of the amino-terminal HHCC domain of HIV-2 integrase: a three-helix bundle stabilized by zinc. *Curr. Biol.* 7, 739–746 (1997).

164. Cai, M. *et al.* Solution structure of the N-terminal zinc binding domain of HIV-1 integrase [published erratum appears in Nat Struct Biol 1997 Oct;4(10):839-40]. *Nat. Struct. Biol.* 42854, 567–577 (1997).

165. Dyda, F. *et al.* Crystal structure of the catalytic domain of HIV-1 integrase: similarity to other polynucleotidyl transferases. *Science* 266, 1981–1986 (1994).

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

167. Calmels, C. *et al.* Biochemical and random mutagenesis analysis of the region carrying the catalytic E152 amino acid of HIV-1 integrase. *Nucleic Acids Res.* 32, 1527–1538 (2004).

168. Lodi, P. J. *et al.* Solution structure of the DNA binding domain of HIV-1 integrase. *Biochemistry* 34, 9826–9833 (1995).

169. Eijkelenboom, A. P. *et al.* The DNA-binding domain of HIV-1 integrase has an SH3-like fold. *Nat. Struct. Biol.* 2, 807–810 (1995).

170. van Gent, D. C., Vink, C., Groeneger, A. A. & Plasterk, R. H. Complementation between HIV integrase proteins mutated in different domains. *EMBO J.* 12, 3261–7 (1993).

171. Engelman, A., Bushman, F. D. & Craigie, R. Identification of discrete functional domains of HIV-1 integrase and their organization within an active multimeric complex. *EMBO J.* 12, 3269–75 (1993).

172. Maertens, G. N., Hare, S. & Cherepanov, P. The mechanism of retroviral integration from X-ray structures of its key intermediates. *Nature* 468, 326–9 (2010).

173. Bukrinsky, M. I. *et al.* Association of integrase, matrix, and reverse transcriptase antigens of human immunodeficiency virus type 1 with viral nucleic acids following acute infection. *Proc. Natl. Acad. Sci. U. S. A.* 90, 6125–6129 (1993).

174. Lee, Y. M. & Coffin, J. M. Relationship of avian retrovirus DNA synthesis to integration in vitro. *Mol. Cell. Biol.* 11, 1419–1430 (1991).

175. Lapadat-tapolsky, M. *et al.* Interactions between HIV-1 nucleocapsid protein and viral DNA may have important functions in the viral life cycle. *Nucleic Acids Research* 21, 2024 (1993).

176. Lee, M. S. & Craigie, R. Protection of retroviral DNA from autointegration: involvement of a cellular factor. *Proc. Natl. Acad. Sci. U. S. A.* 91, 9823–9827 (1994).

177. Farnet, C. M. & Bushman, F. D. HIV-1 cDNA integration: Requirement of HMG I(Y) protein

for function of preintegration complexes in vitro. *Cell* 88, 483–492 (1997).

178. Farnet, C. M. & Haseltine, W. A. Integration of human immunodeficiency virus type 1 DNA in vitro. *Proc Natl Acad Sci U S A* 87, 4164–4168 (1990).

179. Brown, P. O., Bowerman, B., Varmus, H. E. & Bishop, J. M. Correct integration of retroviral DNA in vitro. *Cell* 49, 347–356 (1987).

180. Bushman, F. D. & Craigie, R. Activities of human immunodeficiency virus (HIV) integration protein in vitro: specific cleavage and integration of HIV DNA. *Proc. Natl. Acad. Sci. U. S. A.* 88, 1339–1343 (1991).

181. Katzman, M., Katz, R. A., Skalka, A. M. & Leis, J. The avian retroviral integration protein cleaves the terminal sequences of linear viral DNA at the in vivo sites of integration. *J. Virol.* 63, 5319–5327 (1989).

182. Sherman, P. A. & Fyfe, J. A. Human immunodeficiency virus integration protein expressed in Escherichia coli possesses selective DNA cleaving activity. *Proc. Natl. Acad. Sci. U. S. A.* 87, 5119–5123 (1990).

183. Farnet, C. M. & Haseltine, W. A. Determination of viral proteins present in the human immunodeficiency virus type 1 preintegration complex. *J. Virol.* 65, 1910–1915 (1991).

184. Katz, R. A., Gravuer, K. & Skalka, A. M. A preferred target DNA structure for retroviral integrase in vitro. *J. Biol. Chem.* 273, 24190–24195 (1998).

185. Bor, Y. C., Miller, M. D., Bushman, F. D. & Orgel, L. E. Target-sequence preferences of HIV-1 integration complexes in vitro. *Virology* 222, 283–288 (1996).

186. Fujiwara, T. & Craigie, R. Integration of mini-retroviral DNA: a cell-free reaction for biochemical analysis of retroviral integration. *Proc. Natl. Acad. Sci. U. S. A.* 86, 3065–3069 (1989).

187. Kukolj, G., Katz, R. A. & Skalka, A. M. Characterization of the nuclear localization signal in the avian sarcoma virus integrase. *Gene* 223, 157–163 (1998).

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

189. Lobel, L. I., Murphy, J. E. & Goff, S. P. The palindromic LTR-LTR junction of Moloney murine leukemia virus is not an efficient substrate for proviral integration. *J. Virol.* 63, 2629–2637 (1989).

190. Misra, T. K., Grandgenett, D. P. & Parsons, J. T. Avian retrovirus pp32 DNA-binding protein. I. Recognition of specific sequences on retrovirus DNA terminal repeats. *J. Virol.* 44, 330–343 (1982).

191. Knaus, R. J. *et al.* Avian retrovirus pp32 DNA binding protein. Preferential binding to the promoter region of long terminal repeat DNA. *Biochemistry* 23, 350–359 (1984).

192. Heuer, T. S. & Brown, P. O. Photo-cross-linking studies suggest a model for the architecture of an active human immunodeficiency virus type 1 integrase-DNA complex. *Biochemistry* 37, 6667–6678 (1998).

193. Colicelli, J. & Goff, S. P. Sequence and spacing requirements of a retrovirus integration site. *J. Mol. Biol.* 199, 47–59 (1988).

194. Duyk, G., Leis, J., Longiaru, M. & Skalka, A. M. Selective cleavage in the avian retroviral long terminal repeat sequence by the endonuclease associated with the alpha beta form of avian reverse transcriptase. *Proc. Natl. Acad. Sci. U. S. A.* 80, 6745–6749 (1983).

195. Farnet, C. M. & Haseltine, W. A. Circularization of human immunodeficiency virus type 1 DNA in vitro. *J. Virol.* 65, 6942–6952 (1991).

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

197. Vink, C., van Gent, D. C., Elgersma, Y. & Plasterk, R. H. Human immunodeficiency virus integrase protein requires a subterminal position of its viral DNA recognition sequence for efficient cleavage. *J. Virol.* 65, 4636–4644 (1991).

198. Patel, P. H. & Preston, B. D. Marked infidelity of human immunodeficiency virus type 1 reverse transcriptase at RNA and DNA template ends. *Proc. Natl. Acad. Sci. U. S. A.* 91, 549–553 (1994).

199. Ellison, V. & Brown, P. O. A stable complex between integrase and viral DNA ends mediates human immunodeficiency virus integration in vitro. *Proc. Natl. Acad. Sci. U. S. A.* 91, 7316–7320 (1994).

200. Li, L. *et al.* Role of the non-homologous DNA end joining pathway in the early steps of

retroviral infection. *EMBO J.* 20, 3272–3281 (2001).

201. Junghans, R. P., Boone, L. R. & Skalka, a M. Products of reverse transcription in avian retrovirus analyzed by electron microscopy. *J. Virol.* 43, 544–554 (1982).

202. Kilzer, J. M. *et al.* Roles of host cell factors in circularization of retroviral DNA. *Virology* 314, 460–467 (2003).

203. Lee, M. S. & Craigie, R. A previously unidentified host protein protects retroviral DNA from autointegration. *Proc. Natl. Acad. Sci. U. S. A.* 95, 1528–1533 (1998).

204. Panet, A. & Cedar, H. Selective degradation of integrated murine leukemia proviral DNA by deoxyribonucleases. *Cell* 11, 933–940 (1977).

205. Vijaya, S., Steffen, D. L. & Robinson, H. L. Acceptor sites for retroviral integrations map near DNase I-hypersensitive sites in chromatin. *J. Virol.* 60, 683–692 (1986).

206. Rohdewohld, H., Weiher, H., Reik, W., Jaenisch, R. & Breindl, M. Retrovirus integration and chromatin structure: Moloney murine leukemia proviral integration sites map near DNase I-hypersensitive sites. *J. Virol.* 61, 336–343 (1987).

207. Marini, B. *et al.* Nuclear architecture dictates HIV-1 integration site selection. *Nature* (2015). doi:10.1038/nature14226

208. Chubb, J. R. & Bickmore, W. A. Considering nuclear compartmentalization in the light of nuclear dynamics. *Cell* 112, 403–406 (2003).

209. Bushman, F. D. & Miller, M. D. Tethering human immunodeficiency virus type 1 preintegration complexes to target DNA promotes integration at nearby sites. *J. Virol.* 71, 458–464 (1997).

210. Ciuffi, A. *et al.* A role for LEDGF/p75 in targeting HIV DNA integration. *Nat. Med.* 11, 1287–1289 (2005).

211. Brady, T. *et al.* Quantitation of HIV DNA integration: effects of differential integration site distributions on Alu-PCR assays. *J. Virol. Methods* 189, 53–7 (2013).

212. Serrao, E. *et al.* Integrase residues that determine nucleotide preferences at sites of HIV-1 integration: Implications for the mechanism of target DNA binding. *Nucleic Acids Res.* 42, 5164–5176 (2014).

213. Withers-Ward, E. S., Kitamura, Y., Barnes, J. P. & Coffin, J. M. Distribution of targets for avian retrovirus DNA integration in vivo. *Genes Dev.* 8, 1473–1487 (1994).

214. Shih, C. C., Stoye, J. P. & Coffin, J. M. Highly preferred targets for retrovirus integration. *Cell* 53, 531–537 (1988).

215. Pryciak, P. M. & Varmus, H. E. Nucleosomes, DNA-binding proteins, and DNA sequence modulate retroviral integration target site selection. *Cell* 69, 769–780 (1992).

216. Alexander G. Holman and John M. Coffin. Symmetrical base preferences surrounding HIV-1, avian sarcoma/leukosis virus, and murine leukemia virus integration sites. 102, 6103–6107 (2005).

217. Snásel, J., Rosenberg, I., Paces, O. & Pichová, I. Mapping of HIV-1 integrase preferences for target site selection with various oligonucleotides. *Arch. Biochem. Biophys.* 488, 153–62 (2009).

218. Jiang, C. & Pugh, B. F. Nucleosome positioning and gene regulation: advances through genomics. *Nat. Rev. Genet.* 10, 161–172 (2009).

219. Osipov, S. A., Preobrazhenskaia, O. V & Karpov, V. L. [Chromatin structure and transcription regulation in Saccharomyces cerevisiae]. *Mol Biol* 44, 966–979 (2010).

220. Schones, D. E. *et al.* Dynamic regulation of nucleosome positioning in the human genome. *Cell* 132, 887–898 (2008).

221. Segal, E. *et al.* A genomic code for nucleosome positioning. *Nature* 442, 772–778 (2006).

222. Pruss, D., Bushman, F. D. & Wolffe, A. P. Human immunodeficiency virus integrase directs integration to sites of severe DNA distortion within the nucleosome core. *Proc. Natl. Acad. Sci. U. S. A.* 91, 5913–5917 (1994).

223. Müller, H. P. & Varmus, H. E. DNA bending creates favored sites for retroviral integration: an explanation for preferred insertion sites in nucleosomes. *EMBO J.* 13, 4704–4714 (1994).

224. Pruss, D., Reeves, R., Bushman, F. D. & Wolffe, A. P. The influence of DNA and nucleosome structure on integration events directed by HIV integrase. *J. Biol. Chem.* 269, 25031–25041 (1994).

225. Wang, G. P. *et al.* Analysis of lentiviral vector integration in HIV+ study subjects receiving autologous infusions of gene modified CD4+ T cells. *Mol. Ther.* 17, 844–850 (2009).

226. Wang, G. P., Ciuffi, A., Leipzig, J., Berry, C. C. & Bushman, F. D. HIV integration site selection: Analysis by massively parallel pyrosequencing reveals association with epigenetic modifications. *Genome Res.* 17, 1186–1194 (2007).
227. Brady, T. *et al.* HIV integration site distributions in resting and activated CD4+ T cells infected in culture. *AIDS* 23, 1461–1471 (2009).
228. 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.* 72, 4005–4014 (1998).
229. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* 409, 860–921 (2001).
230. Felice, B. *et al.* Transcription factor binding sites are genetic determinants of retroviral integration in the human genome. *PLoS One* 4, (2009).
231. Mitchell, R. S. *et al.* Retroviral DNA integration: ASLV, HIV, and MLV show distinct target site preferences. *PLoS Biol.* 2, E234 (2004).
232. Schröder, A. R. W. *et al.* HIV-1 integration in the human genome favors active genes and local hotspots. *Cell* 110, 521–9 (2002).
233. Deichmann, A. *et al.* Vector integration is nonrandom and clustered and influences the fate of lymphopoiesis in SCID-X1 gene therapy. 117, (2007).
234. Xiaolin Wu, Yuan Li, Bruce Crise, S. M. B. Transcription Start Regions in the Human Genome Are Favored Targets for MLV Integration. *Science (80-.).* 300, 1749–1751 (2003).
235. Kim, S. *et al.* Fidelity of Target Site Duplication and Sequence Preference during Integration of Xenotropic Murine Leukemia Virus-Related Virus. *PLoS One* 5, (2010).
236. Kim, S. *et al.* Integration site preference of xenotropic murine leukemia virus-related virus, a new human retrovirus associated with prostate cancer. *J. Virol.* 82, 9964–9977 (2008).
237. Cattoglio, C. *et al.* High-definition mapping of retroviral integration sites identifies active regulatory elements in human multipotent hematopoietic progenitors. *Blood* 116, 5507–17 (2010).
238. Narezkina, A. *et al.* Genome-Wide Analyses of Avian Sarcoma Virus Integration Sites Genome-Wide Analyses of Avian Sarcoma Virus Integration Sites. 78, 11656–11663 (2004).
239. Moiani, A. *et al.* Genome-wide analysis of alpharetroviral integration in human hematopoietic stem/progenitor cells. *Genes (Basel).* 5, 415–429 (2014).
240. Ustek, D. *et al.* A genome-wide analysis of lentivector integration sites using targeted sequence capture and next generation sequencing technology. *Infect. Genet. Evol.* 12, 1349–1354 (2012).
241. Roth, S. L., Malani, N. & Bushman, F. D. Gammaretroviral integration into nucleosomal target DNA in vivo. *J. Virol.* 85, 7393–7401 (2011).
242. Wang, G. P. *et al.* DNA bar coding and pyrosequencing to analyze adverse events in therapeutic gene transfer. *Nucleic Acids Res.* 36, 1–12 (2008).
243. Marshall, H. M. *et al.* Role of PSIP 1/LEDGF/p75 in lentiviral infectivity and integration targeting. *PLoS One* 2, (2007).
244. Berry, C. C. *et al.* Estimating abundances of retroviral insertion sites from DNA fragment length data. *Bioinformatics* 28, 755–762 (2012).
245. Lewinski, M. K. *et al.* Retroviral DNA integration: viral and cellular determinants of target-site selection. *PLoS Pathog.* 2, e60 (2006).
246. Ciuffi, A. & Telenti, A. State of genomics and epigenomics research in the perspective of HIV cure. *Curr. Opin. HIV AIDS* 8, 176–81 (2013).
247. Barr, S. D. *et al.* HIV Integration Site Selection: Targeting in Macrophages and the Effects of Different Routes of Viral Entry. *Mol. Ther.* 14, 218–225 (2006).
248. Ferris, A. L. *et al.* Lens epithelium-derived growth factor fusion proteins redirect HIV-1 DNA integration. *Proc. Natl. Acad. Sci. U. S. A.* 107, 3135–3140 (2010).
249. Schröder, a. R. W. *et al.* HIV-1 integration in the human genome favors active genes and local hotspots. *Cell* 110, 521–529 (2002).
250. Lewinski, M. K. & Bushman, F. D. Retroviral DNA Integration-Mechanism and Consequences. *Adv. Genet.* 55, 147–181 (2005).
251. Stevens, S. W. & Griffith, J. D. Human immunodeficiency virus type 1 may preferentially integrate into chromatin occupied by L1Hs repetitive elements. *Proc. Natl. Acad. Sci. U. S. A.* 91, 5557–5561 (1994).

252. Stevens, S. W. & Griffith, J. D. Sequence analysis of the human DNA flanking sites of human immunodeficiency virus type 1 integration. *J. Virol.* 70, 6459–6462 (1996).

253. Leclercq, I. *et al.* Host sequences flanking the human T-cell leukemia virus type 1 provirus in vivo. *J. Virol.* 74, 2305–2312 (2000).

254. Weidhaas, J. B., Angelichio, E. L., Fenner, S. & Coffin, J. M. Relationship between retroviral DNA integration and gene expression. *J. Virol.* 74, 8382–8389 (2000).

255. Lunyak, V. V *et al.* Developmentally regulated activation of a SINE B2 repeat as a domain boundary in organogenesis. *Science* 317, 248–251 (2007).

256. Ferrigno, O. *et al.* Transposable B2 SINE elements can provide mobile RNA polymerase II promoters. *Nat. Genet.* 28, 77–81 (2001).

257. Cherepanov, P. *et al.* HIV-1 integrase forms stable tetramers and associates with LEDGF/p75 protein in human cells. *J. Biol. Chem.* 278, 372–381 (2003).

258. Llano, M., Delgado, S., Vanegas, M. & Poeschla, E. M. Lens epithelium-derived growth factor/p75 prevents proteasomal degradation of HIV-1 integrase. *J. Biol. Chem.* 279, 55570–55577 (2004).

259. Cherepanov, P. *et al.* Solution structure of the HIV-1 integrase-binding domain in LEDGF/p75. *Nat. Struct. Mol. Biol.* 12, 526–532 (2005).

260. Maertens, G. *et al.* LEDGF/p75 is essential for nuclear and chromosomal targeting of HIV-1 integrase in human cells. *J. Biol. Chem.* 278, 33528–39 (2003).

261. Busschots, K. *et al.* The interaction of LEDGF/p75 with integrase is lentivirus-specific and promotes DNA binding. *J. Biol. Chem.* 280, 17841–17847 (2005).

262. Llano, M. *et al.* Identification and Characterization of the Chromatin-binding Domains of the HIV-1 Integrase Interactor LEDGF/p75. *J. Mol. Biol.* 360, 760–773 (2006).

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

264. Shun, M. C. *et al.* LEDGF/p75 functions downstream from preintegration complex formation to effect gene-specific HIV-1 integration. *Genes Dev.* 21, 1767–1778 (2007).

265. Meehan, A. M. *et al.* LEDGF/p75 proteins with alternative chromatin tethers are functional HIV-1 cofactors. *PLoS Pathog.* 5, (2009).

266. Shun, M.-C. *et al.* Identification and characterization of PWWP domain residues critical for LEDGF/p75 chromatin binding and human immunodeficiency virus type 1 infectivity. *J. Virol.* 82, 11555–67 (2008).

267. Eidahl, J. O. *et al.* Structural basis for high-affinity binding of LEDGF PWWP to mononucleosomes. *Nucleic Acids Res.* 41, 3924–3936 (2013).

268. Biasco, L. *et al.* Integration profile of retroviral vector in gene therapy treated patients is cell-specific according to gene expression and chromatin conformation of target cell. *EMBO Mol. Med.* 3, 89–101 (2011).

269. Bannister, A. J. *et al.* Selective recognition of methylated lysine 9 on histone H3 by the HP1 chromo domain. *Nature* 410, 120–124 (2001).

270. Sharma, A. *et al.* BET proteins promote efficient murine leukemia virus integration at transcription start sites. *Proc. Natl. Acad. Sci. U. S. A.* 110, 12036–41 (2013).

271. El Ashkar, S. *et al.* BET-independent MLV-based Vectors Target Away From Promoters and Regulatory Elements. *Mol. Ther. Nucleic Acids* 3, e179 (2014).

272. Kalpana, G. V, Marmon, S., Wang, W., Crabtree, G. R. & Goff, S. P. Binding and stimulation of HIV-1 integrase by a human homolog of yeast transcription factor SNF5. *Science* 266, 2002–2006 (1994).

273. Miller, M. D. & Bushman, F. D. Inil for integration ? The newly discovered Inil cellular protein binds HIV-1 integrase and is. 5, (1995).

274. Bushman, F. *et al.* Genome-wide analysis of retroviral DNA integration. *Nat. Rev. Microbiol.* 3, 848–58 (2005).

275. Bartholomae, C. C. *et al.* Lentiviral vector integration profiles differ in rodent postmitotic tissues. *Mol. Ther.* 19, 703–10 (2011).

276. Bartholomae, C. C. *et al.* Lentiviral vector integration profiles differ in rodent postmitotic tissues. *Mol. Ther.* 19, 703–710 (2011).

277. Poeschla, E. *et al.* Identification of a human immunodeficiency virus type 2 (HIV-2) encapsidation determinant and transduction of nondividing human cells by HIV-2-based lentivirus vectors. *J. Virol.* 72, 6527–6536 (1998).

278. Kim, S. S. *et al.* Generation of replication-defective helper-free vectors based on simian immunodeficiency virus. *Virology* 282, 154–67 (2001).

279. Curran, M. A., Kaiser, S. M., Achacoso, P. L. & Nolan, G. P. Efficient transduction of nondividing cells by optimized feline immunodeficiency virus vectors. *Mol.Ther.* 1, 31–38 (2000).

280. Berkowitz, R. D., Ilves, H., Plavec, I. & Veres, G. Gene transfer systems derived from Visna virus: analysis of virus production and infectivity. *Virology* 279, 116–129 (2001).

281. Mselli-Lakhal, L., Guiguen, F., Greenland, T., Mornex, J. F. & Chebloune, Y. Gene transfer system derived from the caprine arthritis-encephalitis lentivirus. *J. Virol. Methods* 136, 177–184 (2006).

282. Mitrophanous, K. *et al.* Stable gene transfer to the nervous system using a non-primate lentiviral vector. *Gene Ther.* 6, 1808–1818 (1999).

283. Burns, J. C., Friedmann, T., Drievert, W., Burrascano, M. & Yee, J.-K. Vesicular stomatitis virus G glycoprotein pseudotyped retroviral vectors: Concentration to very high titer and efficient gene transfer into mammalian and nonmammalian cells (gene therapy/zebrafish). *Genetics* 90, 8033–8037 (1993).

284. Kahl, C. a, Marsh, J., Fyffe, J., Sanders, D. a & Cornetta, K. Human immunodeficiency virus type 1-derived lentivirus vectors pseudotyped with envelope glycoproteins derived from Ross River virus and Semliki Forest virus. *J. Virol.* 78, 1421–1430 (2004).

285. Trabalza, a *et al.* Venezuelan equine encephalitis virus glycoprotein pseudotyping confers neurotropism to lentiviral vectors. *Gene Ther.* 20, 723–32 (2013).

286. Zufferey, R., Nagy, D., Mandel, R. J., Naldini, L. & Trono, D. Multiply attenuated lentiviral vector achieves efficient gene delivery in vivo. *Nat. Biotechnol.* 15, 871–875 (1997).

287. Otto, E. *et al.* Characterization of a replication-competent retrovirus resulting from recombination of packaging and vector sequences. *Hum. Gene Ther.* 5, 567–75 (1994).

288. Zufferey, R. *et al.* Self-Inactivating Lentivirus Vector for Safe and Efficient In Vivo Gene Delivery. *J. Virol.* 72, 9873–9880 (1998).

289. Yu, S. F. *et al.* Self-inactivating retroviral vectors designed for transfer of whole genes into mammalian cells. *Proc. Natl. Acad. Sci. U. S. A.* 83, 3194–8 (1986).

290. Miyoshi, H., Blömer, U., Takahashi, M., Gage, F. H. & Verma, I. M. Development of a self-inactivating lentivirus vector. *J. Virol.* 72, 8150–7 (1998).

291. Schambach, A., Swaney, W. P. & van der Loo, J. C. M. Design and production of retro- and lentiviral vectors for gene expression in hematopoietic cells. *Methods Mol. Biol.* 506, 191–205 (2009).

292. Throm, R. E. *et al.* Efficient construction of producer cell lines for a SIN lentiviral vector for SCID-X1 gene therapy by concatemeric array transfection. *Blood* 113, 5104–5110 (2009).

293. Iglesias, C. *et al.* Residual HIV-1 DNA Flap-independent nuclear import of cPPT/CTS double mutant viruses does not support spreading infection. *Retrovirology* 8, 92 (2011).

294. Huang, J. & Liang, T. J. A novel hepatitis B virus (HBV) genetic element with Rev response element-like properties that is essential for expression of HBV gene products. *Mol. Cell. Biol.* 13, 7476–7486 (1993).

295. Donello, J. E., Beeche, A. A., Smith, G. J., Lucero, G. R. & Hope, T. J. The hepatitis B virus posttranscriptional regulatory element is composed of two subelements. *J. Virol.* 70, 4345–4351 (1996).

296. Higashimoto, T. *et al.* The woodchuck hepatitis virus post-transcriptional regulatory element reduces readthrough transcription from retroviral vectors. *Gene Ther.* 14, 1298–1304 (2007).

297. Donello, J. E., Loeb, J. E. & Hope, T. J. Woodchuck hepatitis virus contains a tripartite posttranscriptional regulatory element. *J. Virol.* 72, 5085–5092 (1998).

298. Zufferey, R., Donello, J. E., Trono, D. & Hope, T. J. Woodchuck hepatitis virus posttranscriptional regulatory element enhances expression of transgenes delivered by retroviral vectors. *J. Virol.* 73, 2886–92 (1999).

299. Schambach, a *et al.* Woodchuck hepatitis virus post-transcriptional regulatory element deleted from X protein and promoter sequences enhances retroviral vector titer and expression. *Gene Ther.* 13, 641–645 (2006).

300. Kingsman, S. M., Mitrophanous, K. & Olsen, J. C. Potential oncogene activity of the woodchuck hepatitis post-transcriptional regulatory element (WPRE). *Gene therapy* 12, 3–4 (2005).

301. Flajolet, M., Tiollais, P., Buendia, M. A. & Fourel, G. Woodchuck hepatitis virus enhancer I and enhancer II are both involved in N-myc2 activation in woodchuck liver tumors. *J. Virol.* 72, 6175–6180 (1998).

302. Nash, K. L., Jamil, B., Maguire, A. J., Alexander, G. J. M. & Lever, A. M. L. Hepatocyte-specific gene expression from integrated lentiviral vectors. *J. Gene Med.* 6, 974–983 (2004).

303. Sirma, H. *et al.* Hepatitis B virus X mutants, present in hepatocellular carcinoma tissue abrogate both the antiproliferative and transactivation effects of HBx. *Oncogene* 18, 4848–4859 (1999).

304. Tu, H. *et al.* Biological impact of natural COOH-terminal deletions of hepatitis B virus X protein in hepatocellular carcinoma tissues. *Cancer Res.* 61, 7803–7810 (2001).

305. Terradillos, O. *et al.* The hepatitis B virus X gene potentiates c-myc-induced liver oncogenesis in transgenic mice. *Oncogene* 14, 395–404 (1997).

306. Zanta-Boussif, M. a *et al.* Validation of a mutated PRE sequence allowing high and sustained transgene expression while abrogating WHV-X protein synthesis: application to the gene therapy of WAS. *Gene Ther.* 16, 605–619 (2009).

307. Swanstrom, R. & Wills, J. W. Synthesis, Assembly, and Processing of Viral Proteins. *Retroviruses* 263–334 (1997).

308. Wagner, R. *et al.* Rev-independent expression of synthetic gag-pol genes of human immunodeficiency virus type 1 and simian immunodeficiency virus: implications for the safety of lentiviral vectors. *Hum. Gene Ther.* 11, 2403–2413 (2000).

309. Kotsopoulou, E., Kim, V. N., Kingsman, A. J., Kingsman, S. M. & Mitrophanous, K. a. A Rev-Independent Human Immunodeficiency Virus Type 1 (HIV-1) -Based Vector That Exploits a A Rev-Independent Human Immunodeficiency Virus Type 1 (HIV-1) -Based Vector That Exploits a Codon-Optimized HIV-1 gag-pol Gene. *J. Virol.* 74, 4839–4852 (2000).

310. Nappi, F. *et al.* Identification of a novel posttranscriptional regulatory element by using a rev- and RRE-mutated human immunodeficiency virus type 1 DNA proviral clone as a molecular trap. *J. Virol.* 75, 4558–69 (2001).

311. Zolotukhin, A., Valentin, A., Pavlakis, G. & Felber, B. Continuous propagation of RRE(-) and Rev(-)RRE(-) human immunodeficiency virus type 1 molecular clones containing a cis-acting element of simian retrovirus type 1 in human peripheral blood lymphocytes. *J. Virol.* 68, 7944–7952 (1994).

312. Kim, V. N., Mitrophanous, K., Kingsman, S. M. & Kingsman, a J. Minimal requirement for a lentivirus vector based on human immunodeficiency virus type 1. *J. Virol.* 72, 811–816 (1998).

313. Takara, C.-. Fourth Generation Lentiviral Packaging Systems. Available at: http://www.clontech.com/US/Products/Viral_Transduction/Lentiviral_Packaging/Lentiviral_Packaging_Overview.

314. Urano, E. *et al.* Substitution of the myristoylation signal of human immunodeficiency virus type 1 Pr55Gag with the phospholipase C-??1 pleckstrin homology domain results in infectious pseudovirion production. *J. Gen. Virol.* 89, 3144–3149 (2008).

315. Hong, S. *et al.* Functional analysis of various promoters in lentiviral vectors at different stages of in vitro differentiation of mouse embryonic stem cells. *Mol Ther* 15, 1630–1639 (2007).

316. Hanawa, H., Yamamoto, M., Zhao, H., Shimada, T. & Persons, D. a. Optimized lentiviral vector design improves titer and transgene expression of vectors containing the chicken beta-globin locus HS4 insulator element. *Mol. Ther.* 17, 667–674 (2009).

317. Uchida, N., Washington, K. N., Lap, C. J., Hsieh, M. M. & Tisdale, J. F. Chicken HS4 insulators have minimal barrier function among progeny of human hematopoietic cells transduced with an HIV1-based lentiviral vector. *Mol. Ther.* 19, 133–139 (2011).

318. Real, G., Monteiro, F., Burger, C. & Alves, P. M. Improvement of lentiviral transfer vectors using cis-acting regulatory elements for increased gene expression. *Appl. Microbiol. Biotechnol.* 91, 1581–1591 (2011).

319. Leavitt, a D., Robles, G., Alesandro, N. & Varmus, H. E. Human immunodeficiency virus type 1 integrase mutants retain in vitro integrase activity yet fail to integrate viral DNA efficiently during infection. *J. Virol.* 70, 721–728 (1996).

320. Cornu, T. I. & Cathomen, T. Targeted genome modifications using integrase-deficient lentiviral vectors. *Mol. Ther.* 15, 2107–2113 (2007).

321. Miller, D. G., Petek, L. M. & Russell, D. W. Adeno-associated virus vectors integrate at

chromosome breakage sites. *Nat. Genet.* 36, 767–773 (2004).

322. Mátrai, J. *et al.* Hepatocyte-targeted expression by integrase-defective lentiviral vectors induces antigen-specific tolerance in mice with low genotoxic risk. *Hepatology* 53, 1696–1707 (2011).

323. Mates, L. *et al.* Molecular evolution of a novel hyperactive Sleeping Beauty transposase enables robust stable gene transfer in vertebrates. *Nat. Genet.* 41, 753–761 (2009).

324. Staunstrup, N. H. *et al.* Hybrid lentivirus-transposon vectors with a random integration profile in human cells. *Mol. Ther.* 17, 1205–14 (2009).

325. Li, L. *et al.* Genomic editing of the HIV-1 coreceptor CCR5 in adult hematopoietic stem and progenitor cells using zinc finger nucleases. *Mol. Ther.* 21, 1259–69 (2013).

326. Kochenderfer, J. N. & Rosenberg, S. A. Treating B-cell cancer with T cells expressing anti-CD19 chimeric antigen receptors. *Nat. Publ. Gr.* 10, 267–276 (2013).

327. Kumar, M., Keller, B., Makalou, N. & Sutton, R. E. Systematic determination of the packaging limit of lentiviral vectors. *Hum. Gene Ther.* 12, 1893–1905 (2001).

328. Segura, M. M., Mangion, M., Gaillet, B. & Garnier, A. New developments in lentiviral vector design, production and purification. *Expert Opin. Biol. Ther.* 13, 987–1011 (2013).

329. Pichlmair, A. *et al.* Tubulovesicular structures within vesicular stomatitis virus G protein-pseudotyped lentiviral vector preparations carry DNA and stimulate antiviral responses via Toll-like receptor 9. *J. Virol.* 81, 539–547 (2007).

330. Cavazzana-Calvo, M. *et al.* Transfusion independence and HMGA2 activation after gene therapy of human β-thalassaemia. *Nature* 467, 318–322 (2010).

331. Cartier, N. *et al.* Hematopoietic Stem Cell Gene Therapy with a Lentiviral Vector in X-Linked Adrenoleukodystrophy. *Science (80-.).* 326, 818–823 (2009).

332. Merten, O.-W., Hebben, M. & Bovolenta, C. Production of lentiviral vectors. *Mol. Ther. Methods Clin. Dev.* 3, 16017 (2016).

333. Slepushkin, V., Chang, N., Cohen, R., G. & Y., Jiang, B., Deausen, E. Large-scale purification of a lentiviral vector by size exclusion chromatography or Mustang Q ion exchange chromatography. *Bioprocess. J.* 89–95 (2003).

334. Tiscornia, G., Singer, O. & Verma, I. M. Production and purification of lentiviral vectors. *Nat. Protoc.* 1, 241–245 (2006).

335. Giry-Laterriere, M., Verhoeyen, E. & Salmon, P. Lentiviral vectors. *Methods Mol Biol* 737, 183–209 (2011).

336. Kuroda, H., Marino, M. P., Kutner, R. H. & Reiser, J. Production of lentiviral vectors in protein-free media. *Curr. Protoc. Cell Biol.* (2011). doi:10.1002/0471143030.cb2608s50

337. Segura, M. M., Mangion, M., Gaillet, B. & Garnier, A. New developments in lentiviral vector design , production and purification. 1–25 (2013).

338. Broussau, S. *et al.* Inducible packaging cells for large-scale production of lentiviral vectors in serum-free suspension culture. *Mol. Ther.* 16, 500–507 (2008).

339. Côté, J., Garnier, A., Massie, B. & Kamen, A. Serum-free production of recombinant proteins and adenoviral vectors by 293SF-3F6 cells. *Biotechnol. Bioeng.* 59, 567–575 (1998).

340. Ansorge, S. *et al.* Development of a scalable process for high-yield lentiviral vector production by transient transfection of HEK293 suspension cultures. *J. Gene Med.* 11, 868–876 (2009).

341. PUCK, T. T. The genetics of somatic mammalian cells. *Adv. Biol. Med. Phys.* 5, 75–101 (1957).

342. Jayapal, K., Wlaschin, K., Hu, W. & Yap, G. Recombinant protein therapeutics from CHO cells-20 years and counting. *Chem. Eng. Prog.* 103, 40–47 (2007).

343. Chu, L. & Robinson, D. K. Industrial choices for protein production by large-scale cell culture. *Current Opinion in Biotechnology* 12, 180–187 (2001).

344. Fan, L., Frye, C. C. & Racher, A. J. The use of glutamine synthetase as a selection marker: recent advances in Chinese hamster ovary cell line generation processes. *Pharm. Bioprocess.* 1, 487–502 (2013).

345. Hossler, P., Khattak, S. F. & Li, Z. J. Optimal and consistent protein glycosylation in mammalian cell culture. *Glycobiology* 19, 936–49 (2009).

346. Shulman, M., Wilde, C. D. & Köhler, G. A better cell line for making hybridomas secreting specific antibodies. *Nature* 276, 269–270 (1978).

347. Bebbington, C. R. *et al.* High-level expression of a recombinant antibody from myeloma cells using a glutamine synthetase gene as an amplifiable selectable marker. *Biotechnology. (N. Y).* 10, 169–175 (1992).

348. Macpherson, I. & Stoker, M. Polyoma transformation of hamster cell clones--an investigation of genetic factors affecting cell competence. *Virology* 16, 147–151 (1962).

349. Boeger, H. *et al.* Structural basis of eukaryotic gene transcription. in *FEBS Letters* 579, 899–903 (2005).

350. Viruses, A. *Viral Vectors for Gene Therapy*. *Viral Vectors for gene therapy: Methods and Protocols* 737, (2011).

351. Takeuchi, Y. *et al.* Sensitization of rhabdo-, lenti-, and spumaviruses to human serum by galactosyl(alpha1-3)galactosylation. *J Virol* 71, 6174–6178 (1997).

352. Pensiero, M. N., Wysocki, C. A., Nader, K. & Kikuchi, G. E. Development of amphotropic murine retrovirus vectors resistant to inactivation by human serum. *Hum. Gene Ther.* 7, 1095–1101 (1996).

353. Davis, J. L. *et al.* Retroviral particles produced from a stable human-derived packaging cell line transduce target cells with very high efficiencies. *Hum Gene Ther* 8, 1459–1467 (1997).

354. Mason, J. M. *et al.* Human serum-resistant retroviral vector particles from galactosyl (alpha1-3) galactosyl containing nonprimate cell lines. *Gene Ther.* 6, 1397–1405 (1999).

355. Jones, D. *et al.* High-level expression of recombinant IgG in the human cell line per.c6. *Biotechnol Prog* 19, 163–168 (2003).

356. Kirschweger, G. Crucell: biopharmaceuticals--as human as they get. *Mol. Ther.* 7, 5–6 (2003).

357. Graham, F. L., Smiley, J., Russell, W. C. & Nairn, R. Characteristics of a human cell line transformed by DNA from human adenovirus type 5. *J. Gen. Virol.* 36, 59–74 (1977).

358. Graham, F. L. & van der Eb, A. J. A new technique for the assay of infectivity of human adenovirus 5 DNA. *Virology* 52, 456–67 (1973).

359. Louis, N., Evelegh, C. & Graham, F. L. Cloning and sequencing of the cellular-viral junctions from the human adenovirus type 5 transformed 293 cell line. *Virology* 233, 423–9 (1997).

360. Lin, Y.-C. *et al.* Genome dynamics of the human embryonic kidney 293 lineage in response to cell biology manipulations. *Nat. Commun.* 5, 4767 (2014).

361. Park, M., Lee, M., Kim, S., Jo, E.-C. & Lee, G. Influence of culture passages on growth kinetics and adenovirus vector production for gene therapy in monolayer and suspension cultures of HEK 293 cells. *Appl. Microbiol. Biotechnol.* 65, (2004).

362. Shen, C. *et al.* The tumorigenicity diversification in human embryonic kidney 293 cell line cultured in vitro. *Biologicals* 36, 263–268 (2008).

363. Shaw, G., Morse, S., Ararat, M. & Graham, F. L. Preferential transformation of human neuronal cells by human adenoviruses and the origin of HEK 293 cells. *FASEB J.* 16, 869–71 (2002).

364. Merten, O.-W. *et al.* Large-scale manufacture and characterization of a lentiviral vector produced for clinical ex vivo gene therapy application. *Hum. Gene Ther.* 22, 343–56 (2011).

365. DuBridge, R. B. *et al.* Analysis of mutation in human cells by using an Epstein-Barr virus shuttle system. *Mol. Cell. Biol.* 7, 379–87 (1987).

366. Rio, D. C., Clark, S. G. & Tjian, R. A mammalian host-vector system that regulates expression and amplification of transfected genes by temperature induction. *Science* 227, 23–8 (1985).

367. Wu, X. *et al.* SV40 T antigen interacts with Nbs1 to disrupt DNA replication control. *Genes Dev.* 18, 1305–1316 (2004).

368. Gama-Norton, L. *et al.* Lentivirus Production Is Influenced by SV40 Large T-Antigen and Chromosomal Integration of the Vector in HEK293 Cells. *Hum. Gene Ther.* 22, 1269–1279 (2011).

369. Barbanti-Brodano, G. *et al.* Simian virus 40 infection in humans and association with human diseases: Results and hypotheses. *Virology* 318, 1–9 (2004).

370. Manilla, P. *et al.* Regulatory considerations for novel gene therapy products: a review of the process leading to the first clinical lentiviral vector. *Hum. Gene Ther.* 16, 17–25 (2005).

371. Engels, E. a *et al.* Cancer incidence in Denmark following exposure to poliovirus vaccine contaminated with simian virus 40. *J. Natl. Cancer Inst.* 95, 532–539 (2003).

372. Reisman, D. & Sugden, B. trans activation of an Epstein-Barr viral transcriptional enhancer by the Epstein-Barr viral nuclear antigen 1. *Mol. Cell. Biol.* 6, 3838–3846 (1986).

373. Young, J. M., Cheadle, C., Foulke, J. S., Drohan, W. N. & Sarver, N. Utilization of an Epstein-Barr virus replicon as a eukaryotic expression vector. *Gene* 62, 171–185 (1988).

374. Durocher, Y., Perret, S. & Kamen, A. High-level and high-throughput recombinant protein production by transient transfection of suspension-growing human 293-EBNA1 cells. *Nucleic Acids Res.* 30, E9 (2002).

375. Lambert, P. F., Baker, C. C. & Howley, P. M. The Genetics of Bovine Papillomavirus Type 1. *Annu. Rev. Genet.* 22, 235–258 (1988).

376. National Research Council of Canada. HEK 293 expression platform (L-10894 / 11266 / 11565).

377. Yves Durocher. Expression vectors for enhanced transient gene expression and mammalian cells expressing them. (2009).

378. Poeschla, E., Corbeau, P. & Wong-Staal, F. Development of HIV vectors for anti-HIV gene therapy. *Proc. Natl. Acad. Sci. U. S. A.* 93, 11395–11399 (1996).

379. Corbeau, P., Kraus, G. & Wong-Staal, F. Efficient gene transfer by a human immunodeficiency virus type 1 (HIV-1)-derived vector utilizing a stable HIV packaging cell line. *Proc. Natl. Acad. Sci. U. S. A.* 93, 14070–14075 (1996).

380. Srinivasakumar, N. *et al.* The effect of viral regulatory protein expression on gene delivery by human immunodeficiency virus type 1 vectors produced in stable packaging cell lines. *J. Virol.* 71, 5841–5848 (1997).

381. Sanber, K. S. *et al.* Construction of stable packaging cell lines for clinical lentiviral vector production. *Sci. Rep.* 5, 9021 (2015).

382. Marin, V. *et al.* RD-MolPack technology for the constitutive production of self-inactivating lentiviral vectors pseudotyped with the nontoxic RD114-TR envelope. *Mol. Ther. — Methods Clin. Dev.* 3, 16033 (2016).

383. Humbert, O. *et al.* Development of Third-generation Cocal Envelope Producer Cell Lines for Robust Lentiviral Gene Transfer into Hematopoietic Stem Cells and T-cells. *Mol. Ther.* (2016). doi:10.1038/mt.2016.70

384. Sandrin, V. *et al.* Lentiviral vectors pseudotyped with a modified RD114 envelope glycoprotein show increased stability in sera and augmented transduction of primary lymphocytes and CD34+ cells derived from human and nonhuman primates. *Blood* 100, 823–32 (2002).

385. Relander, T. *et al.* 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 (2005).

386. von Kalle, C. *et al.* Increased gene transfer into human hematopoietic progenitor cells by extended in vitro exposure to a pseudotyped retroviral vector. *Blood* 84, 2890–7 (1994).

387. Bunnell, B. A., Muul, L. M., Donahue, R. E., Blaese, R. M. & Morgan, R. A. High-efficiency retroviral-mediated gene transfer into human and nonhuman primate peripheral blood lymphocytes. *Proc. Natl. Acad. Sci. U. S. A.* 92, 7739–43 (1995).

388. Mukherjee, S. & Thrasher, A. J. Gene therapy for PIDs: Progress, pitfalls and prospects. *Gene* 525, 174–181 (2013).

389. Ikeda, Y. *et al.* Continuous high-titer HIV-1 vector production. *Nat. Biotechnol.* 21, 569–572 (2003).

390. Carrondo, M. *et al.* Integrated strategy for the production of therapeutic retroviral vectors. *Hum. Gene Ther.* 22, 370–379 (2011).

391. Schucht, R. *et al.* A New Generation of Retroviral Producer Cells: Predictable and Stable Virus Production by Flp-Mediated Site-Specific Integration of Retroviral Vectors. *Mol. Ther.* 14, 285–292 (2006).

392. Coroadinha, A. S. *et al.* The use of recombinase mediated cassette exchange in retroviral vector producer cell lines: Predictability and efficiency by transgene exchange. *J. Biotechnol.* 124, 457–468 (2006).

393. Stornaiuolo, A. *et al.* RD2-MolPack-Chim3, a Packaging Cell Line for Stable Production of Lentiviral Vectors for Anti-HIV Gene Therapy. *Hum. Gene Ther. Methods* 24, 228–40 (2013).

394. Manilla, P. *et al.* Regulatory considerations for novel gene therapy products: a review of the process leading to the first clinical lentiviral vector. *Hum. Gene Ther.* 16, 17–25 (2005).

395. Rogel, M. E., Wu, L. I. & Emerman, M. The human immunodeficiency virus type 1 vpr gene prevents cell proliferation during chronic infection. *J. Virol.* 69, 882–888 (1995).

396. Jowett, J. B. *et al.* The human immunodeficiency virus type 1 vpr gene arrests infected T cells in the G2 + M phase of the cell cycle. *J. Virol.* 69, 6304–13 (1995).

397. Kaplan, a H. & Swanstrom, R. Human immunodeficiency virus type 1 Gag proteins are

processed in two cellular compartments. *Proc. Natl. Acad. Sci. U. S. A.* 88, 4528–4532 (1991).

398. Kafri, T., van Praag, H., Ouyang, L., Gage, F. H. & Verma, I. M. A packaging cell line for lentivirus vectors. *J. Virol.* 73, 576–584 (1999).

399. Klages, N., Zufferey, R. & Trono, D. A stable system for the high-titer production of multiply attenuated lentiviral vectors. *Mol. Ther.* 2, 170–176 (2000).

400. Farson, D. *et al.* A new-generation stable inducible packaging cell line for lentiviral vectors. *Hum. Gene Ther.* 12, 981–997 (2001).

401. Cockrell, A. S., Ma, H., Fu, K., McCown, T. J. & Kafri, T. A trans-lentiviral packaging cell line for high-titer conditional self-inactivating HIV-1 vectors. *Mol. Ther.* 14, 276–284 (2006).

402. Kaul, M., Yu, H., Ron, Y. & Dougherty, J. P. Regulated lentiviral packaging cell line devoid of most viral cis-acting sequences. *Virology* 249, 167–174 (1998).

403. Ni, Y. *et al.* Generation of a packaging cell line for prolonged large-scale production of high-titer HIV-1-based lentiviral vector. *J. Gene Med.* 7, 818–834 (2005).

404. Xu, K., Ma, H., McCown, T. J., Verma, I. M. & Kafri, T. Generation of a stable cell line producing high-titer self-inactivating lentiviral vectors. *Mol. Ther.* 3, 97–104 (2001).

405. Pacchia, a L., Adelson, M. E., Kaul, M., Ron, Y. & Dougherty, J. P. An inducible packaging cell system for safe, efficient lentiviral vector production in the absence of HIV-1 accessory proteins. *Virology* 282, 77–86 (2001).

406. Sparacio, S., Pfeiffer, T., Schaal, H. & Bosch, V. Generation of a flexible cell line with regulatable, high-level expression of HIV Gag/Pol particles capable of packaging HIV-derived vectors. *Mol. Ther.* 3, 602–612 (2001).

407. Gossen, M. & Bujard, H. Tight control of gene expression in mammalian cells by tetracycline-responsive promoters. *Proc. Natl. Acad. Sci. U. S. A.* 89, 5547–5551 (1992).

408. Yu, H., Rabson, a B., Kaul, M., Ron, Y. & Dougherty, J. P. Inducible human immunodeficiency virus type 1 packaging cell lines. *J. Virol.* 70, 4530–4537 (1996).

409. Kafri, T., Blömer, U., Peterson, D. a, Gage, F. H. & Verma, I. M. Sustained expression of genes delivered directly into liver and muscle by lentiviral vectors. *Nat. Genet.* 17, 314–317 (1997).

410. Westerman, K. a, Ao, Z., Cohen, E. a & Leboulch, P. Design of a trans protease lentiviral packaging system that produces high titer virus. *Retrovirology* 4, 96 (2007).

411. Bestor, T. H. Gene silencing as a threat to the success of gene therapy. *J. Clin. Invest.* 105, 409–411 (2000).

412. Garrick, D., Fiering, S., Martin, D. I. & Whitelaw, E. Repeat-induced gene silencing in mammals. *Nat. Genet.* 18, 56–9 (1998).

413. McBurney, M. W., Mai, T., Yang, X. & Jardine, K. Evidence for repeat-induced gene silencing in cultured Mammalian cells: inactivation of tandem repeats of transfected genes. *Exp. Cell Res.* 274, 1–8 (2002).

414. Ikeda, Y. *et al.* Continuous high-titer HIV-1 vector production. *Nat. Biotechnol.* 21, 569–572 (2003).

415. Stewart, H. J. *et al.* A stable producer cell line for the manufacture of a lentiviral vector for gene therapy of Parkinson's disease. *Hum. Gene Ther.* 22, 357–369 (2011).

416. Hillen, W. & Berens, C. Mechanisms underlying expression of Tn10 encoded tetracycline resistance. *Annu. Rev. Microbiol.* 48, 1–25 (1994).

417. Heinz, N. *et al.* Retroviral and transposon-based tet-regulated all-in-one vectors with reduced background expression and improved dynamic range. *Hum. Gene Ther.* 22, 166–176 (2011).

418. No, D., Yao, T. P. & Evans, R. M. Ecdysone-inducible gene expression in mammalian cells and transgenic mice. *Proc. Natl. Acad. Sci. U. S. A.* 93, 3346–51 (1996).

419. Transfiguracion, J., Jaalouk, D. E., Ghani, K., Galipeau, J. & Kamen, A. Size-exclusion chromatography purification of high-titer vesicular stomatitis virus G glycoprotein-pseudotyped retrovectors for cell and gene therapy applications. *Hum. Gene Ther.* 14, 1139–1153 (2003).

420. Sheridan, P. L. *et al.* Generation of retroviral packaging and producer cell lines for large-scale vector production and clinical application: improved safety and high titer. *Mol. Ther.* 2, 262–75 (2000).

421. Shimizu-Sato, S., Huq, E., Tepperman, J. M. & Quail, P. H. A light-switchable gene promoter system. *Nat. Biotechnol.* 20, 1041–4 (2002).

422. Liu, M., Komiyama, M. & Asanuma, H. Design of light-switchable phage promoter for efficient photo-regulation of gene-expression. *Nucleic Acids Symp. Ser. (Oxf).* 283–284 (2005). doi:10.1093/nass/49.1.283

423. Martínez-García, J. F., Huq, E. & Quail, P. H. Direct targeting of light signals to a promoter element-bound transcription factor. *Science* 288, 859–863 (2000).

424. No Title.

425. Greene, M. R. *et al.* Transduction of human CD34+ repopulating cells with a SIN-lentiviral vector for SCID-X1 produced at clinical scale by a stable cell line. *Hum. Gene Ther. Methods* 308, 121022094731004 (2012).

426. Katz, R. A., Kotler, M. & Skalka, A. M. cis-acting intron mutations that affect the efficiency of avian retroviral RNA splicing: implication for mechanisms of control. *J. Virol.* 62, 2686–2695 (1988).

427. Oppermann, H., Bishop, J. M., Varmus, H. E. & Levintow, L. A joint product of the genes gag and pol of Avian Sarcoma Virus: a possible precursor of reverse transcriptase. *Cell* 12, 993–1005 (1977).

428. Matano, T., Odawara, T., Ohshima, M., Yoshikura, H. & Iwamoto, a. trans-dominant interference with virus infection at two different stages by a mutant envelope protein of Friend murine leukemia virus. *J. Virol.* 67, 2026–33 (1993).

429. Odawara, T., Oshima, M., Doi, K., Iwamoto, A. & Yoshikura, H. Threshold number of provirus copies required per cell for efficient virus production and interference in moloney murine leukemia virus-infected NIH 3T3 cells. *J. Virol.* 72, 5414–24 (1998).

430. Lei, P. & Andreadis, S. T. Stoichiometric limitations in assembly of active recombinant retrovirus. *Biotechnol. Bioeng.* 90, 781–792 (2005).

431. Carrondo, M. J. T., Merten, O.-W., Haury, M., Alves, P. M. & Coroadinha, A. S. Impact of retroviral vector components stoichiometry on packaging cell lines: effects on productivity and vector quality. *Hum. Gene Ther.* 19, 199–210 (2008).

432. O'Brien, J. A. & Lummis, S. C. R. Biolistic transfection of neuronal cultures using a hand-held gene gun. *Nat. Protoc.* 1, 977–81 (2006).

433. Shirahata, Y., Ohkohchi, N., Itagak, H. & Satomi, S. New technique for gene transfection using laser irradiation. *J. Investig. Med.* 49, 184–90 (2001).

434. Plank, C. *et al.* The magnetofection method: Using magnetic force to enhance gene delivery. *Biological Chemistry* 384, 737–747 (2003).

435. Neumann, E., Schaefer-Ridder, M., Wang, Y. & Hofschneider, P. H. Gene transfer into mouse lyoma cells by electroporation in high electric fields. *EMBO J.* 1, 841–5 (1982).

436. Rols, M. P., Coulet, D. & Teissié, J. Highly efficient transfection of mammalian cells by electric field pulses. Application to large volumes of cell culture by using a flow system. *Eur. J. Biochem.* 206, 115–21 (1992).

437. Li, L.-H. *et al.* Highly efficient, large volume flow electroporation. *Technol. Cancer Res. Treat.* 1, 341–350 (2002).

438. Coleman, J. E. *et al.* Efficient large-scale production and concentration of HIV-1-based lentiviral vectors for use in vivo. *Physiol. Genomics* 12, 221–8 (2003).

439. Reed, S. E., Staley, E. M., Mayginnes, J. P., Pintel, D. J. & Tullis, G. E. Transfection of mammalian cells using linear polyethylenimine is a simple and effective means of producing recombinant adeno-associated virus vectors. *J. Virol. Methods* 138, 85–98 (2006).

440. Vandermeulen, G., Marie, C., Scherman, D. & Préat, V. New generation of plasmid backbones devoid of antibiotic resistance marker for gene therapy trials. *Mol. Ther.* 19, 1942–9 (2011).

441. Krieg, A. M. CpG motifs in bacterial DNA and their immune effects. *Annu Rev Immunol* 20, 709–760 (2002).

442. Garnier, A., Côté, J., Nadeau, I., Kamen, A. & Massie, B. Scale-up of the adenovirus expression system for the production of recombinant protein in human 293S cells. *Cytotechnology* 15, 145–55 (1994).

443. Kartenbeck, J., Schmid, E., Franke, W. W. & Geiger, B. Different modes of internalization of proteins associated with adhaerens junctions and desmosomes: experimental separation of lateral contacts induces endocytosis of desmosomal plaque material. *EMBO J.* 1, 725–32 (1982).

444. Merten, O.-W. *et al.* Large-scale manufacture and characterization of a lentiviral vector

produced for clinical ex vivo gene therapy application. *Hum. Gene Ther.* 22, 343–356 (2011).

445. Kutner, R. H., Puthli, S., Marino, M. P. & Reiser, J. Simplified production and concentration of HIV-1-based lentiviral vectors using HYPERFlask vessels and anion exchange membrane chromatography. *BMC Biotechnol.* 9, 10 (2009).

446. Wu, S. C., Huang, G. Y. L. & Liu, J. H. Production of Retrovirus and Adenovirus Vectors for Gene Therapy: A Comparative Study Using Microcarrier and Stationary Cell Culture. *Biotechnol. Prog.* 18, 617–622 (2002).

447. Segura, M. M., Garnier, A., Durocher, Y., Coelho, H. & Kamen, A. Production of lentiviral vectors by large-scale transient transfection of suspension cultures and affinity chromatography purification. *Biotechnol. Bioeng.* 98, 789–799 (2007).

448. Witting, S. R. *et al.* Efficient Large Volume Lentiviral Vector Production Using Flow Electroporation. *Hum. Gene Ther.* 23, 243–249 (2012).

449. Olsen, J. C. & Sechelski, J. Use of sodium butyrate to enhance production of retroviral vectors expressing CFTR cDNA. *Hum. Gene Ther.* 6, 1195–202 (1995).

450. Wade, P. a., Pruss, D. & Wolffe, A. P. Histone acetylation: Chromatin in action. *Trends Biochem. Sci.* 22, 128–132 (1997).

451. Sena-Esteves, M., Tebbets, J. C., Steffens, S., Crombleholme, T. & Flake, A. W. Optimized large-scale production of high titer lentivirus vector pseudotypes. *J. Virol. Methods* 122, 131–139 (2004).

452. Luthman, H. & Magnusson, G. High efficiency polyoma DNA transfection of chloroquine treated cells. *Nucleic Acids Res.* 11, 1295–1308 (1983).

453. Ellis, B. L., Potts, P. R. & Porteus, M. H. Creating higher titer lentivirus with caffeine. *Hum. Gene Ther.* 22, 93–100 (2011).

454. Kuroda, H., Kutner, R. H., Bazan, N. G. & Reiser, J. Simplified lentivirus vector production in protein-free media using polyethylenimine-mediated transfection. *J. Virol. Methods* 157, 113–121 (2009).

455. Vogt, B. *et al.* Lack of superinfection interference in retroviral vector producer cells. *Hum. Gene Ther.* 12, 359–365 (2001).

456. Brandtner, E. M. *et al.* Quantification and characterization of autotransduction in retroviral vector producer cells. *Hum. Gene Ther.* 19, 97–102 (2008).

457. Carroll, R. *et al.* A human immunodeficiency virus type 1 (HIV-1)-based retroviral vector system utilizing stable HIV-1 packaging cell lines. *J. Virol.* 68, 6047–6051 (1994).

458. Haselhorst, D., Kaye, J. F. & Lever, a M. Development of cell lines stably expressing human immunodeficiency virus type 1 proteins for studies in encapsidation and gene transfer. *J. Gen. Virol.* 79 (Pt 2), 231–7 (1998).

459. Kuate, S., Wagner, R. & Überla, K. Development and characterization of a minimal inducible packaging cell line for simian immunodeficiency virus-based lentiviral vectors. *J. Gene Med.* 4, 347–355 (2002).

460. Strang, B. L. *et al.* Human Immunodeficiency Virus Type 1 Vectors with Alphavirus Envelope Glycoproteins Produced from Stable Packaging Cells Human Immunodeficiency Virus Type 1 Vectors with Alphavirus Envelope Glycoproteins Produced from Stable Packaging Cells. 79, 1765–1771 (2005).

461. Muratori, C. *et al.* Generation and characterization of a stable cell population releasing fluorescent HIV-1-based Virus Like Particles in an inducible way. *BMC Biotechnol.* 6, 52 (2006).

462. Lee, C. L., Chou, M., Dai, B., Xiao, L. & Wang, P. Construction of stable producer cells to make high-titer lentiviral vectors for dendritic cell-based vaccination. *Biotechnol. Bioeng.* 109, 1551–1560 (2012).

463. Hu, P., Li, Y., Sands, M. S., McCown, T. & Kafri, T. Generation of a stable packaging cell line producing high-titer PPT-deleted integration-deficient lentiviral vectors. *Mol. Ther. — Methods Clin. Dev.* 2, 15025 (2015).

464. Wurm, F. M. Production of recombinant protein therapeutics in cultivated mammalian cells. *Nat. Biotechnol.* 22, 1393–8 (2004).

465. Ringold, G., Dieckmann, B. & Lee, F. Co-expression and amplification of dihydrofolate reductase cDNA and the Escherichia coli XGPRT gene in Chinese hamster ovary cells. *J. Mol. Appl. Genet.* 1, 165–75 (1981).

466. Le Hir, H., Nott, A. & Moore, M. J. How introns influence and enhance eukaryotic gene

expression. *Trends Biochem. Sci.* 28, 215–20 (2003).

467. Makrides, S. C. Components of vectors for gene transfer and expression in mammalian cells. *Protein Expr. Purif.* 17, 183–202 (1999).

468. Kalwy, S., Rance, J. & Young, R. Toward more efficient protein expression: keep the message simple. *Mol. Biotechnol.* 34, 151–156 (2006).

469. Balland, A. *et al.* Characterisation of two differently processed forms of human recombinant factor IX synthesised in CHO cells transformed with a polycistronic vector. 572, 565–572 (1988).

470. Ryan, M. D., King, A. M. & Thomas, G. P. Cleavage of foot-and-mouth disease virus polyprotein is mediated by residues located within a 19 amino acid sequence. *J. Gen. Virol.* 72 (Pt 11, 2727–32 (1991).

471. Bestor, T. H. The DNA methyltransferases of mammals. *Hum. Mol. Genet.* 9, 2395–2402 (2000).

472. Cosset, F. L., Takeuchi, Y., Battini, J. L., Weiss, R. a & Collins, M. K. High-titer packaging cells producing recombinant retroviruses resistant to human serum. *J. Virol.* 69, 7430–7436 (1995).

473. Girod, P. A., Zahn-Zabal, M. & Mermod, N. Use of the chicken lysozyme 5??? matrix attachment region to generate high producer CHO cell lines. *Biotechnol. Bioeng.* 91, 1–11 (2005).

474. Kim, J.-M. *et al.* Improved recombinant gene expression in CHO cells using matrix attachment regions. *J. Biotechnol.* 107, 95–105 (2004).

475. Zahn-Zabal, M. *et al.* Development of stable cell lines for production or regulated expression using matrix attachment regions. *J. Biotechnol.* 87, 29–42 (2001).

476. Barnes, L. M. & Dickson, A. J. Mammalian cell factories for efficient and stable protein expression. *Curr. Opin. Biotechnol.* 17, 381–6 (2006).

477. Mirkovitch, J., Mirault, M. E. & Laemmli, U. K. Organization of the higher-order chromatin loop: specific DNA attachment sites on nuclear scaffold. *Cell* 39, 223–232 (1984).

478. Bell, A. C., West, A. G. & Felsenfeld, G. The protein CTCF is required for the enhancer blocking activity of vertebrate insulators. *Cell* 98, 387–396 (1999).

479. Girod, P.-A. *et al.* Genome-wide prediction of matrix attachment regions that increase gene expression in mammalian cells. *Nat. Methods* 4, 747–753 (2007).

480. Antoniou, M. *et al.* Transgenes encompassing dual-promoter CpG islands from the human TBP and HNRPA2B1 loci are resistant to heterochromatin-mediated silencing. *Genomics* 82, 269–279 (2003).

481. Benton, T. *et al.* The use of UCOE vectors in combination with a preadapted serum free, suspension cell line allows for rapid production of large quantities of protein. *Cytotechnology* 38, 43–6 (2002).

482. Zhang, F. *et al.* Lentiviral vectors containing an enhancer-less ubiquitously acting chromatin opening element (UCOE) provide highly reproducible and stable transgene expression in hematopoietic cells. *Blood* 110, 1448–1457 (2007).

483. Ye, J. *et al.* Rapid protein production using CHO stable transfection pools. *Biotechnol. Prog.* 26, 1431–1437 (2010).

484. Mallik A, Pinkus G, S. S. Biopharma's capacity crunch. *McKinsey Q. 2002 Spec. Ed. Risk Resilience. McKinsey Co.* 9–11 (2002).

485. Adams, J. M. & Cory, S. Life-or-death decisions by the Bcl-2 protein family. *Trends in Biochemical Sciences* 26, 61–66 (2001).

486. Kim, Y. G., Kim, J. Y., Mohan, C. & Lee, G. M. Effect of Bcl-xL overexpression on apoptosis and autophagy in recombinant Chinese hamster ovary cells under nutrient-deprived condition. *Biotechnol. Bioeng.* 103, 757–766 (2009).

487. Hwang, S. O. & Lee, G. M. Effect of Akt overexpression on programmed cell death in antibody-producing Chinese hamster ovary cells. *J. Biotechnol.* 139, 89–94 (2009).

488. Sunley, K. & Butler, M. Strategies for the enhancement of recombinant protein production from mammalian cells by growth arrest. *Biotechnol. Adv.* 28, 385–394 (2010).

489. Lao, M. S. & Toth, D. Effects of ammonium and lactate on growth and metabolism of a recombinant Chinese hamster ovary cell culture. *Biotechnol. Prog.* 13, 688–691 (1997).

490. Zhang, F., Sun, X., Yi, X. & Zhang, Y. Metabolic characteristics of recombinant Chinese hamster ovary cells expressing glutamine synthetase in presence and absence of glutamine. *Cytotechnology* 51, 21–28 (2006).

491. Kim, S. H. & Lee, G. M. Functional expression of human pyruvate carboxylase for reduced lactic acid formation of Chinese hamster ovary cells (DG44). *Appl. Microbiol. Biotechnol.* 76, 659–665 (2007).

492. Zhou, M. *et al.* Decreasing lactate level and increasing antibody production in Chinese Hamster Ovary cells (CHO) by reducing the expression of lactate dehydrogenase and pyruvate dehydrogenase kinases. *J. Biotechnol.* 153, 27–34 (2011).

493. Lin, C. Y. *et al.* Enhancing protein expression in HEK-293 cells by lowering culture temperature. *PLoS One* 10, (2015).

494. Garcia-Ruiz, E., Gonzalez-Perez, D., Ruiz-Dueñas, F. J., Martínez, A. T. & Alcalde, M. Directed evolution of a temperature-, peroxide- and alkaline pH-tolerant versatile peroxidase. *Biochem. J.* 441, 487–98 (2012).

495. Gorman, C., Padmanabhan, R. & Howard, B. H. High efficiency DNA-mediated transformation of primate cells. *Science* 221, 551–553 (1983).

496. Orrantia, E. & Chang, P. L. Intracellular distribution of DNA internalized through calcium phosphate precipitation. *Exp. Cell Res.* 190, 170–174 (1990).

497. Vaughan, E. E. & Dean, D. A. Intracellular trafficking of plasmids during transfection is mediated by microtubules. *Mol. Ther.* 13, 422–428 (2006).

498. Potter, H. & Heller, R. Transfection by Electroporation. *Curr. Protoc. Neurosci.* (2011). doi:10.1002/0471142301.nsa01es57

499. Grosjean, F., Batard, P., Jordan, M. & Wurm, F. M. S-phase synchronized CHO cells show elevated transfection efficiency and expression using CaPi. in *Cytotechnology* 38, 57–62 (2002).

500. Weber, M., Möller, K., Welzeck, M. & Schorr, J. Short technical reports. Effects of lipopolysaccharide on transfection efficiency in eukaryotic cells. *Biotechniques* 19, 930–40 (1995).

501. Stuchbury, G. & Münch, G. Optimizing the generation of stable neuronal cell lines via pre-transfection restriction enzyme digestion of plasmid DNA. *Cytotechnology* 62, 189–194 (2010).

502. Ivics, Z., Hackett, P. B., Plasterk, R. H. & Izsvák, Z. Molecular Reconstruction of Sleeping Beauty, a Tc1-like Transposon from Fish, and Its Transposition in Human Cells. *Cell* 91, 501–510 (1997).

503. Kawakami, K. & Shima, A. Identification of the Tol2 transposase of the medaka fish Oryzias latipes that catalyzes excision of a nonautonomous Tol2 element in zebrafish Danio rerio. *Gene* 240, 239–44 (1999).

504. Fraser, M. J., Ciszczon, T., Elick, T. & Bauser, C. Precise excision of TTAA-specific lepidopteran transposons piggyBac (IFP2) and tagalong (TFP3) from the baculovirus genome in cell lines from two species of Lepidoptera. *Insect Mol. Biol.* 5, 141–51 (1996).

505. Vink, C. A Hybrid Lentivirus-Transposon Vector for Safer Gene Therapy. (2009).

506. Matasci, M., Baldi, L., Hacker, D. L. & Wurm, F. M. The PiggyBac transposon enhances the frequency of CHO stable cell line generation and yields recombinant lines with superior productivity and stability. *Biotechnol. Bioeng.* 108, 2141–2150 (2011).

507. Balasubramanian, S. *et al.* Rapid recombinant protein production from piggyBac transposon-mediated stable CHO cell pools. *J. Biotechnol.* 200, 61–69 (2015).

508. Li, Z., Michael, I. P., Zhou, D., Nagy, A. & Rini, J. M. Simple piggyBac transposon-based mammalian cell expression system for inducible protein production. *Proc. Natl. Acad. Sci. U. S. A.* 110, 5004–9 (2013).

509. Ley, D. *et al.* MAR Elements and Transposons for Improved Transgene Integration and Expression. *PLoS One* 8, (2013).

510. Howe, S. J. *et al.* Insertional mutagenesis combined with acquired somatic mutations causes leukemogenesis following gene therapy of SCID-X1 patients. 118, (2008).

511. Smith, K. Theoretical mechanisms in targeted and random integration of transgene DNA. *Reprod. Nutr. Dev.* 41, 465–485 (2002).

512. Wong, E. A. & Capecchi, M. R. Analysis of homologous recombination in cultured mammalian cells in transient expression and stable transformation assays. *Somat. Cell Mol. Genet.* 12, 63–72 (1986).

513. Calos, M. P., Lebkowski, J. S. & Botchan, M. R. High mutation frequency in DNA transfected into mammalian cells. *Proc. Natl. Acad. Sci. U. S. A.* 80, 3015–9 (1983).

514. Burdon, T. G. & Wall, R. J. Fate of microinjected genes in preimplantation mouse embryos.

Mol. Reprod. Dev. 33, 436–442 (1992).

515. Chen, C. & Chasin, L. A. Cointegration of DNA molecules introduced into mammalian cells by electroporation. *Somat. Cell Mol. Genet.* 24, 249–256 (1998).

516. Covarrubias, L., Nishida, Y., Terao, M., D'Eustachio, P. & Mintz, B. Cellular DNA rearrangements and early developmental arrest caused by DNA insertion in transgenic mouse embryos. *Mol. Cell. Biol.* 7, 2243–7 (1987).

517. Covarrubias, L., Nishida, Y. & Mintz, B. Early postimplantation embryo lethality due to DNA rearrangements in a transgenic mouse strain. *Proc.Natl.Acad.Sci.U.S.A.* 83, 6020–6024 (1986).

518. Milot E, Belmaaza A, Wallenburg JC, Gusew N, Bradley WE and Chartrand, P. Chromosomal illegitimate recombination in mammalian cells is associated with intrinsically bent DNA elements. *EMBO J.* 11, 5063–5070 (1992).

519. Robins, D. M., Ripley, S., Henderson, A. S. & Axel, R. Transforming DNA integrates into the host chromosome. *Cell* 23, 29–39 (1981).

520. Murnane, J. P. & Yu, L. C. Acquisition of telomere repeat sequences by transfected DNA integrated at the site of a chromosome break. *Mol. Cell. Biol.* 13, 977–983 (1993).

521. Bestor, T. H. The host defence function of genomic methylation patterns. *Novartis Found. Symp.* 214, 187-189-232 (1998).

522. Heartlein, M. W., Knoll, J. H. & Latt, S. a. Chromosome instability associated with human alphoid DNA transfected into the Chinese hamster genome. *Mol. Cell. Biol.* 8, 3611–8 (1988).

523. Murnane, J. P. & Young, B. R. Nucleotide sequence analysis of novel junctions near an unstable integrated plasmid in human cells. *Gene* 84, 201–205 (1989).

524. Merrihew, R. V, Marburger, K., Pennington, S. L., Roth, D. B. & Wilson, J. H. High-frequency illegitimate integration of transfected DNA at preintegrated target sites in a mammalian genome. *Mol. Cell. Biol.* 16, 10–18 (1996).

525. Perez, E. E. *et al.* Establishment of HIV-1 resistance in CD4+ T cells by genome editing using zinc-finger nucleases. *Nat. Biotechnol.* 26, 808–16 (2008).

526. Festenstein, R. *et al.* Locus control region function and heterochromatin-induced position effect variegation. *Science* 271, 1123–5 (1996).

527. Komor, A. C., Kim, Y. B., Packer, M. S., Zuris, J. A. & Liu, D. R. Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. *Nature* 61, 5985–91 (2016).

528. He, X. *et al.* Knock-in of large reporter genes in human cells via CRISPR/Cas9-induced homology-dependent and independent DNA repair. *Nucleic Acids Res.* gkw064- (2016). doi:10.1093/nar/gkw064

529. Goncz, K. K., Prokopishyn, N. L., Chow, B. L., Davis, B. R. & Gruenert, D. C. Application of SFHR to gene therapy of monogenic disorders. *Gene Ther* 9, 691–694 (2002).

530. ACG, P. Designer Genomes. *Techniques* 53–65 (1989).

531. Porteus, M. Using homologous recombination to manipulate the genome of human somatic cells. *Biotechnol. Genet. Eng. Rev.* 24, 195–212 (2007).

532. O'Driscoll, M. & Jeggo, P. a. The role of double-strand break repair - insights from human genetics. *Nat Rev Genet.* 7, 45–54 (2006).

533. Jackson, S. P. & Bartek, J. The DNA-damage response in human biology and disease. *Nature* 461, 1071–8 (2009).

534. Doyon, J. B. *et al.* Rapid and efficient clathrin-mediated endocytosis revealed in genome-edited mammalian cells. *Nat. Cell Biol.* 13, 331–337 (2011).

535. Hockemeyer, D. *et al.* Genetic engineering of human pluripotent cells using TALE nucleases. *Nat. Biotechnol.* 29, 731–4 (2011).

536. Li, H. *et al.* In vivo genome editing restores haemostasis in a mouse model of haemophilia. *Nature* 475, 217–221 (2011).

537. Maresca, M., Lin, V. G., Guo, N. & Yang, Y. Obligate ligation-gated recombination (ObLiGaRe): Custom-designed nuclease-mediated targeted integration through nonhomologous end joining. *Genome Res.* 23, 539–546 (2013).

538. Ran, F. A. *et al.* Double nicking by RNA-guided CRISPR cas9 for enhanced genome editing specificity. *Cell* 154, 1380–1389 (2013).

539. Chen, F. *et al.* High-frequency genome editing using ssDNA oligonucleotides with zinc-finger nucleases. *Nat Methods* 8, 753–755 (2011).

540. Choi, P. S. & Meyerson, M. Targeted genomic rearrangements using CRISPR/Cas

technology. *Nat. Commun.* 5, 3728 (2014).

541. Lee, H. J., Kweon, J., Kim, E., Kim, S. & Kim, J. S. Targeted chromosomal duplications and inversions in the human genome using zinc finger nucleases. *Genome Res.* 22, 539–548 (2012).

542. Khanna, K. K. & Jackson, S. P. DNA double-strand breaks: signaling, repair and the cancer connection. *Nat. Genet.* 27, 247–54 (2001).

543. Chappel, S. C. Method for the modification of the expression characteristics of an endogenous gene of a given cell line. (1990).

544. Yutaka, K. *et al.* Comparison of cell lines for stable production of fucose-negative antibodies with enhanced ADCCYutaka Kanda and Naoko Yamane-Ohnuki contributed equally to this work. *Biotechnol. Bioeng.* 94, 680–688 (2006).

545. Grandjean, M. *et al.* High-level transgene expression by homologous recombination-mediated gene transfer. *Nucleic Acids Res.* 39, (2011).

546. Capecchi, M. R. Altering the genome by homologous recombination. *Science* 244, 1288–92 (1989).

547. Chapman, J. R., Sossick, A. J., Boulton, S. J. & Jackson, S. P. BRCA1-associated exclusion of 53BP1 from DNA damage sites underlies temporal control of DNA repair. *J. Cell Sci.* 125, 3529–34 (2012).

548. Bunting, S. F. *et al.* 53BP1 inhibits homologous recombination in brca1-deficient cells by blocking resection of DNA breaks. *Cell* 141, 243–254 (2010).

549. Helmink, B. A. *et al.* H2AX prevents CtIP-mediated DNA end resection and aberrant repair in G1-phase lymphocytes. *Nature* 469, 245–249 (2011).

550. Rouet, P., Smih, F. & Jasin, M. Introduction of double-strand breaks into the genome of mouse cells by expression of a rare-cutting endonuclease. *Mol. Cell. Biol.* 14, 8096–106 (1994).

551. Urnov, F. D., Rebar, E. J., Holmes, M. C., Zhang, H. S. & Gregory, P. D. Genome editing with engineered zinc finger nucleases. *Nat. Rev. Genet.* 11, 636–46 (2010).

552. Wolfe, S. A., Nekludova, L. & Pabo, C. O. DNA recognition by Cys2His2 zinc finger proteins. *Annu Rev Biophys Biomol Struct* 29, 183–212 (2000).

553. Silva, G. *et al.* Meganucleases and other tools for targeted genome engineering: perspectives and challenges for gene therapy. *Curr. Gene Ther.* 11, 11–27 (2011).

554. Cabaniols, J.-P. *et al.* Meganuclease-driven targeted integration in CHO-K1 cells for the fast generation of HTS-compatible cell-based assays. *J. Biomol. Screen.* 15, 956–67 (2010).

555. Sun, N. & Zhao, H. Transcription activator-like effector nucleases (TALENs): A highly efficient and versatile tool for genome editing. *Biotechnol. Bioeng.* 110, 1811–1821 (2013).

556. Santiago, Y. *et al.* Targeted gene knockout in mammalian cells by using engineered zinc-finger nucleases. *Proc. Natl. Acad. Sci. U. S. A.* 105, 5809–5814 (2008).

557. Liu, P. *et al.* Generation of a triple-gene knockout mammalian cell line using engineered zinc-finger nucleases. *Biotechnol. Bioeng.* 106, 97–105 (2010).

558. Cost, G. J. *et al.* BAK and BAX deletion using zinc-finger nucleases yields apoptosis-resistant CHO cells. *Biotechnol. Bioeng.* 105, 330–40 (2010).

559. Li, J. *et al.* Multiplexed, targeted gene editing in Nicotiana benthamiana for glyco-engineering and monoclonal antibody production. *Plant Biotechnol J* (2015). doi:10.1111/pbi.12403

560. Kim, H. & Kim, J.-S. A guide to genome engineering with programmable nucleases. *Nat. Rev. Genet.* 15, 321–34 (2014).

561. Carroll, D. Genome engineering with targetable nucleases. *Annu. Rev. Biochem.* 83, 409–39 (2014).

562. Marraffini, L. A. & Sontheimer, E. J. CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. *Nat. Rev. Genet.* 11, 181–90 (2010).

563. Koonin, E. V & Makarova, K. S. CRISPR-Cas: evolution of an RNA-based adaptive immunity system in prokaryotes. *RNA Biol.* 10, 679–86 (2013).

564. Barrangou, R. & Marraffini, L. A. CRISPR-cas systems: Prokaryotes upgrade to adaptive immunity. *Molecular Cell* 54, 234–244 (2014).

565. Horvath, P. & Barrangou, R. CRISPR/Cas, the immune system of bacteria and archaea. *Science* 327, 167–170 (2010).

566. Wiedenheft, B., Sternberg, S. H. & Doudna, J. a. RNA-guided genetic silencing systems in bacteria and archaea. *Nature* 482, 331–338 (2012).

567. Bhaya, D., Davison, M. & Barrangou, R. CRISPR-Cas Systems in Bacteria and Archaea: Versatile Small RNAs for Adaptive Defense and Regulation. *Annu. Rev. Genet.* 45, 273–297 (2011).

568. Terns, M. P. & Terns, R. M. CRISPR-based adaptive immune systems. *Current Opinion in Microbiology* 14, 321–327 (2011).

569. Haurwitz, R. E., Jinek, M., Wiedenheft, B., Zhou, K. & Doudna, J. A. Sequence- and structure-specific RNA processing by a CRISPR endonuclease. *Science* 329, 1355–8 (2010).

570. Deltcheva, E. *et al.* CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature* 471, 602–607 (2011).

571. Gesner, E. M., Schellenberg, M. J., Garside, E. L., George, M. M. & Macmillan, A. M. Recognition and maturation of effector RNAs in a CRISPR interference pathway. *Nat. Struct. Mol. Biol.* 18, 688–692 (2011).

572. Sashital, D. G., Jinek, M. & Doudna, J. a. An RNA-induced conformational change required for CRISPR RNA cleavage by the endoribonuclease Cse3. *Nat. Struct. Mol. Biol.* 18, 680–7 (2011).

573. Lintner, N. G. *et al.* The structure of the CRISPR-associated protein csa3 provides insight into the regulation of the CRISPR/Cas system. *J. Mol. Biol.* 405, 939–955 (2011).

574. Makarova, K. S., Wolf, Y. I. & Koonin, E. V. Comparative genomics of defense systems in archaea and bacteria. *Nucleic Acids Res.* 41, 4360–4377 (2013).

575. Makarova, K. S. *et al.* An updated evolutionary classification of CRISPR-Cas systems. *Nat. Rev. Microbiol.* 13, 722–736 (2015).

576. Makarova, K. S. & Koonin, E. V. Annotation and classification of CRISPR-Cas systems. *Methods Mol. Biol.* 1311, 47–75 (2015).

577. Marraffini, L. A. & Sontheimer, E. J. CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. *Science (80-.).* 322, 1843–1845 (2008).

578. Sapranauskas, R. *et al.* The Streptococcus thermophilus CRISPR/Cas system provides immunity in Escherichia coli. *Nucleic Acids Res.* 39, 9275–9282 (2011).

579. Jinek, M. *et al.* A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* 337, 816–21 (2012).

580. Makarova, K. S., Grishin, N. V, Shabalina, S. A., Wolf, Y. I. & Koonin, E. V. A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. *Biol. Direct* 1, 7 (2006).

581. Hsu, P. D. *et al.* DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat. Biotechnol.* 31, 827–32 (2013).

582. Zhou, Y. *et al.* High-throughput screening of a CRISPR/Cas9 library for functional genomics in human cells. *Nature* 509, 487–91 (2014).

583. Ma, H. *et al.* Multiplexed labeling of genomic loci with dCas9 and engineered sgRNAs using CRISPRainbow. *Nat. Biotechnol.* 34, 528–530 (2016).

584. Kimura, Y. *et al.* CRISPR/Cas9-mediated reporter knock-in in mouse haploid embryonic stem cells. *Sci. Rep.* 5, 10710 (2015).

585. Krentz, N. a. J., Nian, C. & Lynn, F. C. TALEN/CRISPR-Mediated eGFP Knock-In Add-On at the OCT4 Locus Does Not Impact Differentiation of Human Embryonic Stem Cells towards Endoderm. *PLoS One* 9, e114275 (2014).

586. Wang, L. *et al.* Large genomic fragment deletion and functional gene cassette knock-in via Cas9 protein mediated genome editing in one-cell rodent embryos. *Sci. Rep.* 5, 17517 (2015).

587. Chu, V. T. *et al.* Efficient generation of Rosa26 knock-in mice using CRISPR/Cas9 in C57BL/6 zygotes. *BMC Biotechnol.* 16, 4 (2016).

588. Kimura, Y., Hisano, Y., Kawahara, A. & Higashijima, S. Transgenic Zebrafish Carrying Reporter /. 1–7 (2014). doi:10.1038/srep06545

589. Lee, J. S., Kallehauge, T. B., Pedersen, L. E. & Kildegaard, H. F. Site-specific integration in CHO cells mediated by CRISPR/Cas9 and homology-directed DNA repair pathway. *Sci. Rep.* 5, 8572 (2015).

590. Wang, Z. *et al.* CRISPR/Cas9-Derived Mutations Both Inhibit HIV-1 Replication and Accelerate Viral Escape. *Cell Rep.* 15, 481–489 (2016).

591. Reardon, S. First CRISPR clinical trial gets green light from US panel. *Nature* (2016). doi:10.1038/nature.2016.20137

592. WuXi Biologics. Case Study: FUT8 KO using CRISPR/Cas9 Technology. Available at: http://www.wuxibiologics.com/client-support-resources-2/case-studies-for-development-services/.

593. Grindley, N. D. F., Whiteson, K. L. & Rice, P. a. Mechanisms of site-specific recombination. *Annu. Rev. Biochem.* 75, 567–605 (2006).

594. Turan, S. & Bode, J. Site-specific recombinases: from tag-and-target- to tag-and-exchange-based genomic modifications. *FASEB J.* 25, 4088–4107 (2011).

595. Holliday, R. A mechanism for gene conversion in fungi. *Genet. Res.* 5, 282–304 (1964).

596. Landy, A. Dynamic, Structural, and Regulatory Aspects of lambda Site-Specific Recombination. *Annu. Rev. Biochem.* 58, 913–941 (1989).

597. Kuhstoss, S. & Rao, R. N. Analysis of the integration function of the streptomycete bacteriophage phi C31. *Journal of molecular biology* 222, 897–908 (1991).

598. Groth, a C., Olivares, E. C., Thyagarajan, B. & Calos, M. P. A phage integrase directs efficient site-specific integration in human cells. *Proc. Natl. Acad. Sci. U. S. A.* 97, 5995–6000 (2000).

599. Ma, Q. wen *et al.* Identification of pseudo attP sites for phage ??{symbol}C31 integrase in bovine genome. *Biochem. Biophys. Res. Commun.* 345, 984–988 (2006).

600. Bi, Y. *et al.* Pseudo attP sites in favor of transgene integration and expression in cultured porcine cells identified by streptomyces phage phiC31 integrase. *BMC Mol. Biol.* 14, 20 (2013).

601. Sternberg, N. & Hamilton, D. Bacteriophage P1 site-specific recombination. I. Recombination between loxP sites. *J. Mol. Biol.* 150, 467–486 (1981).

602. Hoess, R. H., Ziese, M. & Sternberg, N. P1 site-specific recombination: nucleotide sequence of the recombining sites. *Proc. Natl. Acad. Sci. U. S. A.* 79, 3398–3402 (1982).

603. Seibler, J., Sch??beler, D., Fiering, S., Groudine, M. & Bode, J. DNA cassette exchange in ES cells mediated by FLF recombinase: An efficient strategy for repeated modification of tagged loci by marker-free constructs. *Biochemistry* 37, 6229–6234 (1998).

604. Verhoeyen, E., Hauser, H. & Wirth, D. Evaluation of retroviral vector design in defined chromosomal loci by Flp-mediated cassette replacement. *Hum. Gene Ther.* 12, 933–944 (2001).

605. Loonstra, A. *et al.* Growth inhibition and DNA damage induced by Cre recombinase in mammalian cells. *Proc. Natl. Acad. Sci. U. S. A.* 98, 9209–14 (2001).

606. Schmidt-Supprian, M. & Rajewsky, K. Vagaries of conditional gene targeting. *Nat. Immunol.* 8, 665–668 (2007).

607. Lee, G. & Saito, I. Role of nucleotide sequences of loxP spacer region in Cre-mediated recombination. *Gene* 216, 55–65 (1998).

608. Lee, J. & Jayaram, M. Role of partner homology in DNA recombination: Complementary base pairing orients the 5???-hydroxyl for strand joining during Flp site-specific recombination. *J. Biol. Chem.* 270, 4042–4052 (1995).

609. Umlauf, S. W. & Cox, M. M. The functional significance of DNA sequence structure in a site-specific genetic recombination reaction. *EMBO J.* 7, 1845–52 (1988).

610. Albert, H., Dale, E. C., Lee, E. & Ow, D. W. Site-specific integration of DNA into wild-type and mutant lox sites placed in the plant genome. *Plant J.* 7, 649–659 (1995).

611. Araki, K., Araki, M. & Yamamura, K. Targeted integration of DNA using mutant lox sites in embryonic stem cells. *Nucleic Acids Res.* 25, 868–72 (1997).

612. Hoess, R. H., Wierzbicki, A. & Abremski, K. The role of the loxP spacer region in P1 site-specific recombination. *Nucleic Acids Res.* 14, 2287–300 (1986).

613. Branda, C. S. & Dymecki, S. M. Talking about a revolution: The impact of site-specific recombinases on genetic analyses in mice. *Developmental Cell* 6, 7–28 (2004).

614. Wiberg, F. C. *et al.* Production of target-specific recombinant human polyclonal antibodies in mammalian cells. *Biotechnol. Bioeng.* 94, 396–405 (2006).

615. McLeod, M., Craft, S. & Broach, J. R. Identification of the crossover site during FLP-mediated recombination in the Saccharomyces cerevisiae plasmid 2 microns circle. *Mol. Cell. Biol.* 6, 3357–67 (1986).

616. Buchholz, F., Angrand, P. O. & Stewart, a F. Improved properties of FLP recombinase evolved by cycling mutagenesis. *Nat. Biotechnol.* 16, 657–662 (1998).

617. Schlake, T. & Bode, J. Use of mutated FLP recognition target (FRT) sites for the exchange of expression cassettes at defined chromosomal loci. *Biochemistry* 33, 12746–12751 (1994).

618. Gaj, T. & Barbas, C. F. Genome engineering with custom recombinases. *Methods Enzymol.*

546, 79–91 (2014).

619. Gaj, T., Mercer, A. C., Sirk, S. J., Smith, H. L. & Barbas, C. F. A comprehensive approach to zinc-finger recombinase customization enables genomic targeting in human cells. *Nucleic Acids Res.* 41, 3937–3946 (2013).

620. Maeder, M. L., Thibodeau-Beganny, S., Sander, J. D., Voytas, D. F. & Joung, J. K. Oligomerized pool engineering (OPEN): an 'open-source' protocol for making customized zinc-finger arrays. *Nat. Protoc.* 4, 1471–501 (2009).

621. Sander, J. D. *et al.* Selection-free zinc-finger-nuclease engineering by context-dependent assembly (CoDA). *Nat. Methods* 8, 67–9 (2011).

622. Kaern, M., Elston, T. C., Blake, W. J. & Collins, J. J. Stochasticity in gene expression: from theories to phenotypes. *Nat. Rev. Genet.* 6, 451–464 (2005).

623. Lai, T., Yang, Y. & Ng, S. K. Advances in Mammalian Cell Line Development Technologies. 579–603 (2013). doi:10.3390/ph6050579

624. Ng, S. K. Generation of high-expressing cells by methotrexate amplification of destabilized dihydrofolate reductase selection marker. *Methods Mol. Biol.* 801, 161–172 (2012).

625. Westwood, A. D., Rowe, D. A. & Clarke, H. R. G. Improved recombinant protein yield using a codon deoptimized DHFR selectable marker in a CHEF1 expression plasmid. *Biotechnol. Prog.* 26, 1558–1566 (2010).

626. Ho, S. C. L. *et al.* IRES-mediated Tricistronic vectors for enhancing generation of high monoclonal antibody expressing CHO cell lines. *J. Biotechnol.* 157, 130–139 (2012).

627. Yenofsky, R. L., Fine, M. & Pellow, J. W. A mutant neomycin phosphotransferase II gene reduces the resistance of transformants to antibiotic selection pressure. *Proc. Natl. Acad. Sci. U. S. A.* 87, 3435–9 (1990).

628. Chin, C. L. S. H. C. L., Chin, H. K., Chin, C. L. S. H. C. L., Lai, E. T. & Ng, S. K. Engineering selection stringency on expression vector for the production of recombinant human alpha1-antitrypsin using Chinese Hamster ovary cells. *BMC Biotechnol.* 15, 44 (2015).

629. Alt, F. W., Kellems, R. E. & Schimke, R. T. Synthesis and degradation of folate reductase in sensitive and methotrexate-resistant lines of S-180 cells. *J Biol Chem* 251, 3063–3074 (1976).

630. Gandor, C., Leist, C., Fiechter, A. & Asselbergs, F. A. M. Amplification and expression of recombinant genes in serum-independent Chinese hamster ovary cells. *FEBS Lett.* 377, 290–294 (1995).

631. Pallavicini, M. G., DeTeresa, P. S., Rosette, C., Gray, J. W. & Wurm, F. M. Effects of methotrexate on transfected DNA stability in mammalian cells. *Mol. Cell. Biol.* 10, 401–404 (1990).

632. Simonsen, C. C. & Levinson, a D. Isolation and expression of an altered mouse dihydrofolate reductase cDNA. *Proc. Natl. Acad. Sci. U. S. A.* 80, 2495–2499 (1983).

633. Wirth, M., Bode, J., Zettlmeissl, G. & Hauser, H. Isolation of overproducing recombinant mammalian cell lines by a fast and simple selection procedure. *Gene* 73, 419–426 (1988).

634. Kaufman, R. J., Murtha, P., Ingolia, D. E., Yeung, C. Y. & Kellems, R. E. Selection and amplification of heterologous genes encoding adenosine deaminase in mammalian cells. *Proc. Natl. Acad. Sci. U. S. A.* 83, 3136–3140 (1986).

635. Fan, L. *et al.* Improving the efficiency of CHO cell line generation using glutamine synthetase gene knockout cells. *Biotechnol. Bioeng.* 109, 1007–15 (2012).

636. Browne, S. M. & Al-Rubeai, M. Selection methods for high-producing mammalian cell lines. *Trends Biotechnol.* 25, 425–32 (2007).

637. Barnes, L. M., Bentley, C. M. & Dickson, A. J. Molecular Definition of Predictive Indicators of Stable Protein Expression in Recombinant NSO Myeloma Cells. *Biotechnol. Bioeng.* 85, 115–121 (2004).

638. Gorman, C. M., Howard, B. H. & Reeves, R. Expression of recombinant plasmids in mammalian cells is enhanced by sodium butyrate. *Nucleic Acids Res* 11, 7631–7648 (1983).

639. Mazur, X., Fussenegger, M., Renner, W. a & Bailey, J. E. Higher productivity of growth-arrested Chinese hamster ovary cells expressing the cyclin-dependent kinase inhibitor p27. *Biotechnol. Prog.* 14, 705–13 (1998).

640. Puck, T. T. & Marcus, P. I. A RAPID METHOD FOR VIABLE CELL TITRATION AND CLONE PRODUCTION WITH HELA CELLS IN TISSUE CULTURE: THE USE OF X-IRRADIATED CELLS TO SUPPLY CONDITIONING FACTORS. *Proc. Natl. Acad. Sci. U. S. A.* 41, 432–7 (1955).

641. Anne Underwood, P. & Bean, P. A. Hazards of the limiting-dilution method of cloning

hybridomas. *J. Immunol. Methods* 107, 119–128 (1988).

642. Yang, G. & Withers, S. G. Ultrahigh-throughput FACS-based screening for directed enzyme evolution. *ChemBioChem* 10, 2704–2715 (2009).

643. Marder, P., Maciak, R. S., Fouts, R. L., Baker, R. S. & Starling, J. J. Selective cloning of hybridoma cells for enhanced immunoglobulin production using flow cytometric cell sorting and automated laser nephelometry. *Cytometry* 11, 498–505 (1990).

644. DeMaria, C. T. *et al.* Accelerated clone selection for recombinant CHO CELLS using a FACS-based high-throughput screen. *Biotechnol. Prog.* 23, 465–472 (2007).

645. Meng, Y. G., Liang, J., Lee, W. & Chisholm, V. Green fluorescent protein as a second selectable marker for selection of high producing clones from transfected CHO cells. 242, 201–207 (2000).

646. Mancia, F. *et al.* Optimization of protein production in mammalian cells with a coexpressed fluorescent marker. *Structure* 12, 1355–60 (2004).

647. Yoshikawa, T. *et al.* Flow cytometry: an improved method for the selection of highly productive gene-amplified CHO cells using flow cytometry. *Biotechnol. Bioeng.* 74, 435–442 (2001).

648. Powell, K. T. & Weaver, J. C. Gel microdroplets and flow cytometry: rapid determination of antibody secretion by individual cells within a cell population. *Biotechnol. (Nature Publ. Company)* 8, 333–7 (1990).

649. Weaver, J. C., McGrath, P. & Adams, S. Gel microdrop technology for rapid isolation of rare and high producer cells. *Nat. Med.* 3, 583–585 (1997).

650. Borth, N., Zeyda, M., Kunert, R. & Katinger, H. Efficient selection of high-producing subclones during gene amplification of recombinant Chinese hamster ovary cells by flow cytometry and cell sorting. *Biotechnol. Bioeng.* 71, 266–273 (2001).

651. Koller, M. R. *et al.* High-throughput laser-mediated in situ cell purification with high purity and yield. *Cytom. Part A* 61, 153–161 (2004).

652. Hanania, E. G. *et al.* Automated in situ measurement of cell-specific antibody secretion and laser-mediated purification for rapid cloning of highly-secreting producers. *Biotechnol. Bioeng.* 91, 872–876 (2005).

653. Lee, C., Ly, C. & Sauerwald, T. High-throughput screening of cell lines expressing monoclonal antibodies. *Bioprocess Int.* 32–35 (2006).

654. Xu, X. *et al.* The genomic sequence of the Chinese hamster ovary (CHO)-K1 cell line. *Nat. Biotechnol.* 29, 735–741 (2011).

655. Baik, J. Y. *et al.* Initial transcriptome and proteome analyses of low culture temperature-induced expression in CHO cells producing erythropoietin. *Biotechnol. Bioeng.* 93, 361–371 (2006).

656. Yee, J. C., Gerdtzen, Z. P. & Hu, W. S. Comparative transcriptome analysis to unveil genes affecting recombinant protein productivity in mammalian cells. *Biotechnol. Bioeng.* 102, 246–263 (2009).

657. Schepers, K. *et al.* Dissecting T cell lineage relationships by cellular barcoding. *J. Exp. Med.* 205, 2309–18 (2008).

658. van Heijst, J. W. J. *et al.* Recruitment of antigen-specific CD8+ T cells in response to infection is markedly efficient. *Science* 325, 1265–1269 (2009).

659. Gerrits, A. *et al.* Cellular barcoding tool for clonal analysis in the hematopoietic system. *Blood* 115, 2610–8 (2010).

660. Cheung, A. M. S. *et al.* Analysis of the clonal growth and differentiation dynamics of primitive barcoded human cord blood cells in NSG mice. *Blood* 122, 3129–37 (2013).

661. Lu, R., Neff, N. F., Quake, S. R. & Weissman, I. L. Tracking single hematopoietic stem cells in vivo using high-throughput sequencing in conjunction with viral genetic barcoding. *Nat. Biotechnol.* 29, 928–933 (2011).

662. Gerlach, C. *et al.* Heterogeneous differentiation patterns of individual CD8+ T cells. *Science* 340, 635–9 (2013).

663. Livet, J. *et al.* Transgenic strategies for combinatorial expression of fluorescent proteins in the nervous system. *Nature* 450, 56–62 (2007).

664. Naik, S. H., Schumacher, T. N. & Perié, L. Cellular barcoding: A technical appraisal. *Exp. Hematol.* 42, 598–608 (2014).

665. Peikon, I. D., Gizatullina, D. I. & Zador, A. M. In vivo generation of DNA sequence diversity for cellular barcoding. *Nucleic Acids Res.* 42, 1–10 (2014).

666. Maeda, N. *et al.* Development of a DNA barcode tagging method for monitoring dynamic changes in gene expression by using an ultra high-throughput sequencer. *Biotechniques* 45, 95–97 (2008).

667. Hoffmann, C. *et al.* DNA bar coding and pyrosequencing to identify rare HIV drug resistance mutations. *Nucleic Acids Res.* 35, e91–e91 (2007).

668. Hillenmeyer, M. E. *et al.* The chemical genomic portrait of yeast: uncovering a phenotype for all genes. *Science (80-.).* 320, 362–365 (2008).

669. Smith, A. M. *et al.* Quantitative phenotyping via deep barcode sequencing. *Genome Res.* 19, 1836–1842 (2009).

670. Robinson, D. G., Chen, W., Storey, J. D. & Gresham, D. Design and analysis of Bar-seq experiments. *G3 (Bethesda).* 4, 11–8 (2014).

671. Gresham, D. *et al.* System-level analysis of genes and functions affecting survival during nutrient starvation in Saccharomyces cerevisiae. *Genetics* 187, 299–317 (2011).

672. Naik, S. H., Schumacher, T. N. & Perié, L. Cellular barcoding: A technical appraisal. *Exp. Hematol.* 42, 598–608 (2014).

673. Filion, G. J. The TRiP technology. Available at: http://www.genomearchitecture.com/research-lines.

674. Cellecta. CellTracker™ Lentiviral Barcode Library. Available at: https://www.cellecta.com/products-services/cellecta-pooled-lentiviral-libraries/celltracker-lentiviral-barcode-library/.

675. Bhang, H. C. *et al.* Studying clonal dynamics in response to cancer therapy using high-complexity barcoding. *Nat. Med.* 21, 440–8 (2015).

676. Sancho, P. *et al.* MYC/PGC-1?? balance determines the metabolic phenotype and plasticity of pancreatic cancer stem cells. *Cell Metab.* 22, 590–605 (2015).

677. Zufferey, R. *et al.* Self-Inactivating Lentivirus Vector for Safe and Efficient In Vivo Gene Delivery Self-Inactivating Lentivirus Vector for Safe and Efficient In Vivo Gene Delivery. 72, (1998).

678. ATCC. *293T (ATCC® CRL-3216™)* Available at: https://www.lgcstandards-atcc.org/Products/All/CRL-3216.aspx?geo_country=gb.

679. Lund, A. H., Duch, M. & Pedersen, F. S. Increased cloning efficiency by temperature-cycle ligation. *Nucleic Acid Res.* 24, 1996–1997 (1996).

680. Inoue, H., Nojima, H., 0kayama, H. High efficiency transformation of Escherichia coli with plasmids. *Gene* 96, 23–28 (1990).

681. Naldini, L., Blömer, U., Gage, F. H., Trono, D. & Verma, I. M. Efficient transfer, integration, and sustained long-term expression of the transgene in adult rat brains injected with a lentiviral vector. *Proc. Natl. Acad. Sci. U. S. A.* 93, 11382–8 (1996).

682. Dull, T. *et al.* A Third-Generation Lentivirus Vector with a Conditional Packaging System A Third-Generation Lentivirus Vector with a Conditional Packaging System. 72, (1998).

683. White, S. M. *et al.* Lentivirus vectors using human and simian immunodeficiency virus elements. *J. Virol.* 73, 2832–40 (1999).

684. Shimada, H., Obayashi, T., Takahashi, N., Matsui, M. & Sakamoto, A. Normalization using ploidy and genomic DNA copy number allows absolute quantification of transcripts, proteins and metabolites in cells. *Plant Methods* 6, 9 (2010).

685. Rand, K. N. *et al.* Headloop suppression PCR and its application to selective amplification of methylated DNA sequences. *Nucleic Acids Res.* 33, 1–11 (2005).

686. Blankenberg, D. *et al.* Galaxy: A web-based genome analysis tool for experimentalists. *Curr. Protoc. Mol. Biol.* 1–21 (2010). doi:10.1002/0471142727.mb1910s89

687. Afgan, E. *et al.* The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Res.* gkw343 (2016). doi:10.1093/nar/gkw343

688. Goecks, J., Nekrutenko, A. & Taylor, J. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.* 11, R86 (2010).

689. Blankenberg, D. *et al.* Manipulation of FASTQ data with galaxy. *Bioinformatics* 26, 1783–1785 (2010).

690. Andrews, S. FastQC: a quality control tool for high throughput sequence data. Available at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

691. Zorita, E., Cusco, P. & Filion, G. J. Sequence analysis Starcode : sequence clustering based

on all-pairs search. 31, 1913–1919 (2015).

692. Kent, W. J. BLAT - The BLAST-like alignment tool. *Genome Res.* 12, 656–664 (2002).

693. Kent, W. J. *et al.* The Human Genome Browser at UCSC. *Genome Res.* 12, 996–1006 (2002).

694. Quinlan, A. R. & Hall, I. M. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841–842 (2010).

695. Jurka, J. Repbase Update: A database and an electronic journal of repetitive elements. *Trends in Genetics* 16, 418–420 (2000).

696. Hocum, J. D. *et al.* VISA--Vector Integration Site Analysis server: a web-based server to rapidly identify retroviral integration sites from next-generation sequencing. *BMC Bioinformatics* 16, 212 (2015).

697. Sinici, I., Zarghooni, M., Tropak, M. B., Mahuran, D. J. & Özkara, H. A. Comparison of HCMV IE and EF-1α promoters for the stable expression of β-subunit of hexosaminidase in CHO cell lines. *Biochem. Genet.* 44, 173–180 (2006).

698. Schambach, A. *et al.* Equal potency of gammaretroviral and lentiviral SIN vectors for expression of O6-methylguanine-DNA methyltransferase in hematopoietic cells. *Mol. Ther.* 13, 391–400 (2006).

699. Zychlinski, D. *et al.* Physiological Promoters Reduce the Genotoxic Risk of Integrating Gene Vectors. *Mol. Ther.* 16, 718–725 (2008).

700. Montiel-Equihua, C. a *et al.* The β-globin locus control region in combination with the EF1α short promoter allows enhanced lentiviral vector-mediated erythroid gene expression with conserved multilineage activity. *Mol. Ther.* 20, 1400–9 (2012).

701. Avedillo Diez, I. *et al.* Development of novel efficient SIN vectors with improved safety features for Wiskott-Aldrich syndrome stem cell based gene therapy. *Mol. Pharm.* 8, 1525–1537 (2011).

702. Qin, J. Y. *et al.* Systematic comparison of constitutive promoters and the doxycycline-inducible promoter. *PLoS One* 5, (2010).

703. Somers, A., , A. Omari, C.C. Ford, J.A. Mills, L. Ying, A., Sommer Gianotti, J.M. Jean, B.W. Smith, R. Lafyatis, M.F. Demierre, D.J. Weiss, D. L. & French, P. Gadue, G.J. Murphy, G. M. and D. N. K. Generation of transgene-free lung disease-specific human iPS cells using a single excisable lentiviral stem cell cassette. *Stem Cells* 28, 1728–1740 (2010).

704. Crooks, G. E., Hon, G., Chandonia, J. M. & Brenner, S. E. WebLogo: A sequence logo generator. *Genome Res.* 14, 1188–1190 (2004).

705. Hayashi, K., Nakazawa, M., Ishizaki, Y., Hiraoka, N. & Obayashi, A. Regulation of inter- and intramolecular ligation with T4 DNA ligase in the presence of polyethylene glycol. *Nucleic Acids Res.* 14, 7617–7631 (1986).

706. Bystrykh, L. V. Generalized DNA barcode design based on Hamming codes. *PLoS One* 7, e36852 (2012).

707. Buschmann, T. & Bystrykh, L. V. Levenshtein error-correcting barcodes for multiplexed DNA sequencing. *BMC Bioinformatics* 14, 272 (2013).

708. Levenshtein, V. I. Binary codes capable of correcting deletions, insertions, and reversals. *Sov. Phys. Dokl.* 10, 707–710 (1966).

709. Chao, A., Pan, H. Y. & Chiang, S. C. The Petersen - Lincoln estimator and its extension to estimate the size of a shared population. *Biometrical J.* 50, 957–970 (2008).

710. Chapman D.G. Some properties of the hypergeometric distribution with applications to zoological sample censuses. *Publ. Stat.* 131–160 (1951).

711. Krebs, C. J. *Ecological Methodology*. *Harper and Row* (1989). doi:10.1007/s007690000247

712. Seber, G. a F. The estimation of animal abundance and related parameters. *New York* 2, 654 (1982).

713. Feller, W. *An Introduction to Probability Theory and Its Applications*. *Wiley* 2, (1968).

714. McInerney, P., Adams, P. & Hadi, M. Z. Error Rate Comparison during Polymerase Chain Reaction by DNA Polymerase. *Mol. Biol. Int.* 2014, 287430 (2014).

715. Fehse, B., Kustikova, O. S., Bubenheim, M. & Baum, C. Pois(s)on – It's a Question of Dose…. *Gene Ther.* 11, 879–881 (2004).

716. Kustikova, O. S. *et al.* Dose finding with retroviral vectors: correlation of retroviral vector copy numbers in single cells with gene transfer efficiency in a cell population. *Blood* 102, 3934–7 (2003).

717. Nolan-Stevaux, O. *et al.* Measurement of Cancer Cell Growth Heterogeneity through Lentiviral Barcoding Identifies Clonal Dominance as a Characteristic of In Vivo Tumor

Engraftment. *PLoS One* 8, (2013).

718. Boettcher, M. *et al.* Decoding pooled RNAi screens by means of barcode tiling arrays. *BMC Genomics* 11, 7 (2010).

719. Zhang, B. *et al.* The significance of controlled conditions in lentiviral vector titration and in the use of multiplicity of infection (MOI) for predicting gene transfer events. *Genet. Vaccines Ther.* 2, 6 (2004).

720. Ellis, J. Silencing and variegation of gammaretrovirus and lentivirus vectors. *Hum. Gene Ther.* 16, 1241–6 (2005).

721. Längle-Rouault, F. *et al.* Up to 100-fold increase of apparent gene expression in the presence of Epstein-Barr virus oriP sequences and EBNA1: implications of the nuclear import of plasmids. *J. Virol.* 72, 6181–5 (1998).

722. Jager, V. *et al.* High level transient production of recombinant antibodies and antibody fusion proteins in HEK293 cells. *BMC Biotechnol* 13, (2013).

723. Tsai, Y.-C. *et al.* Linear correlation between average fluorescence intensity of green fluorescent protein and the multiplicity of infection of recombinant adenovirus. *J. Biomed. Sci.* 22, 31 (2015).

724. Charrier, S. *et al.* Quantification of lentiviral vector copy numbers in individual hematopoietic colony-forming cells shows vector dose-dependent effects on the frequency and level of transduction. *Gene Ther.* 18, 479–487 (2011).

725. Bystrykh, L. V., de Haan, G. & Verovskaya, E. in *Methods in Molecular Biology* (ed. Qu, K. D. B. and C.-K.) 345–360 (2014). doi:10.1007/978-1-4939-1133-2_23

726. Worthington, M. T., Pelo, J. & Luo, R. Q. Cloning of random oligonucleotides to create single-insert plasmid libraries. *Anal. Biochem.* 294, 169–175 (2001).

727. Watson, R. J., Schildraut, I., Qiang, B. Q., Martin, S. M. & Visentin, L. P. NdeI: a restriction endonuclease from Neisseria denitrificans which cleaves DNA at 5'-CATATG-3' sequences. *FEBS Lett.* 150, 114–6 (1982).

728. Wang, A. H. J., Hakoshima, T., van der Marel, G., van Boom, J. H. & Rich, A. AT base pairs are less stable than GC base pairs in Z-DNA: The crystal structure of d(m5CGTAm5CG). *Cell* 37, 321–331 (1984).

729. Gilles, A. *et al.* Accuracy and quality assessment of 454 GS-FLX Titanium pyrosequencing. *BMC Genomics* 12, 245 (2011).

730. Cheung, A. M. S. *et al.* Analysis of the clonal growth and differentiation dynamics of primitive barcoded human cord blood cells in NSG mice. *Blood* 122, 3129–3137 (2013).

731. Grosselin, J. *et al.* Arrayed lentiviral barcoding for quantification analysis of hematopoietic dynamics. *Stem Cells* 31, 2162–2171 (2013).

732. Verovskaya, E. *et al.* Heterogeneity of young and aged murine hematopoietic stem cells revealed by quantitative clonal analysis using cellular barcoding. *Blood* 122, 523–32 (2013).

733. Cornils, K. *et al.* Multiplexing clonality: Combining RGB marking and genetic barcoding. *Nucleic Acids Res.* 42, (2014).

734. Lu Rong, Neff Norma, Quake Stephen, W. I. Tracking single hematopoietic stem cells in vivo using high-throughput sequencing in conjunction with viral genetic barcoding. *Nat. Biotechnol.* 29, 928–933 (2012).

735. Brugman, M. H. *et al.* Development of a diverse human T-cell repertoire despite stringent restriction of hematopoietic clonality in the thymus. *Proc. Natl. Acad. Sci. U. S. A.* 112, E6020-7 (2015).

736. Porter, S. N., Baker, L. C., Mittelman, D. & Porteus, M. H. Lentiviral and targeted cellular barcoding reveals ongoing clonal dynamics of cell lines in vitro and in vivo. *Genome Biol.* 15, R75 (2014).

737. Peshwa, M. V., Kyung, Y. S., McClure, D. B. & Hu, W. S. Cultivation of mammalian cells as aggregates in bioreactors: Effect of calcium concentration on spatial distribution of viability. *Biotechnol. Bioeng.* 41, 179–187 (1993).

738. Wu, J., Rostami, M. R., Cadavid Olaya, D. P. & Tzanakakis, E. S. Oxygen transport and stem cell aggregation in stirred-suspension bioreactor cultures. *PLoS One* 9, (2014).

739. Renner, W. A., Jordan, M., Eppenberger, H. M. & Leist, C. Cell-cell adhesion and aggregation: Influence on the growth behavior of CHO cells. *Biotechnol. Bioeng.* 41, 188–193 (1993).

740. Grell, M. *et al.* Induction of cell death by tumour necrosis factor (TNF) receptor 2, CD40 and CD30: A role for TNF-R1 activation by endogenous membrane-anchored TNF. *EMBO J.*

18, 3034–3043 (1999).

741. Houtz B, Trotter J, S. Tips on Cell Preparation for Flow Cytometric Analysis and Sorting. *BD FACService Technotes* Vol.4. p 3 (2004). Available at: https://www.bdbiosciences.com/documents/BD_Research_Sorting_TechBulletin.pdf.

742. Covault, J., Liu, Q. yang & El-Deeb, S. Calcium-activated proteolysis of intracellular domains in the cell adhesion molecules NCAM and N-cadherin. *Mol. Brain Res.* 11, 11–16 (1991).

743. Oppenheimer-Marks, N. & Grinnell, F. Calcium ions protect cell-substratum adhesion receptors against proteolysis. Evidence from immunoabsorption and electroblotting studies. *Exp. Cell Res.* 152, 467–475 (1984).

744. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437, 376–80 (2005).

745. Dohm, J. C., Lottaz, C., Borodina, T. & Himmelbauer, H. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res.* 36, e105–e105 (2008).

746. Eid, J. *et al.* Real-time DNA sequencing from single polymerase molecules. *Science* 323, 133–138 (2009).

747. Liu, H. S., Jan, M. S., Chou, C. K., Chen, P. H. & Ke, N. J. Is green fluorescent protein toxic to the living cells? *Biochem. Biophys. Res. Commun.* 260, 712–7 (1999).

748. Bartz, S. R. & Vodicka, M. a. Production of high-titer human immunodeficiency virus type 1 pseudotyped with vesicular stomatitis virus glycoprotein. *Methods* 12, 337–342 (1997).

749. Appelt, J.-U. *et al.* QuickMap: a public tool for large-scale gene therapy vector insertion site mapping and analysis. *Gene Ther.* 16, 885–893 (2009).

750. Huston, M. W. *et al.* Comprehensive investigation of parameter choice in viral integration site analysis and its effects on the gene annotations produced. *Hum. Gene Ther.* 23, 1209–19 (2012).

751. Calabria, A. *et al.* VISPA: a computational pipeline for the identification and analysis of genomic vector integration sites. *Genome Med.* 6, 67 (2014).

752. Saiki, R. *et al.* Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science (80-.).* 239, 487–491 (1988).

753. Ramsköld, D. *et al.* Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat. Biotechnol.* 30, 777–82 (2012).

754. Corney, D. C. RNA-seq Using Next Generation Sequencing. *Mater. Methods* 3, 203 (2013).

755. Wu, X., Li, Y., Crise, B., Burgess, S. M. & Munroe, D. J. Weak palindromic consensus sequences are a common feature found at the integration target sites of many retroviruses. *J Virol* 79, 5211–5214 (2005).

756. Bernard, P. & Allshire, R. C. Centromeres become unstuck without heterochromatin. *Trends in Cell Biology* 12, 419–424 (2002).

757. Biffi, A. *et al.* Lentiviral vector common integration sites in preclinical models and a clinical trial reflect a benign integration bias and not oncogenic selection. *Blood* 117, 5332–9 (2011).

758. Mitchell, R. S. *et al.* Retroviral DNA integration: ASLV, HIV, and MLV show distinct target site preferences. *PLoS Biol.* 2, (2004).

759. Wang, G. P., Ciuffi, A., Leipzig, J., Berry, C. C. & Bushman, F. D. HIV integration site selection : Analysis by massively parallel pyrosequencing reveals association with epigenetic modifications HIV integration site selection : Analysis by massively parallel pyrosequencing reveals association with epigenetic modificatio. 1186–1194 (2007). doi:10.1101/gr.6286907

760. Moiani, A. *et al.* Genome-wide analysis of alpharetroviral integration in human hematopoietic stem/progenitor cells. *Genes (Basel).* 5, 415–429 (2014).

761. Zheng, D. *et al.* Pseudogenes in the ENCODE regions: Consensus annotation, analysis of transcription, and evolution. *Genome Res.* 17, 839–851 (2007).

762. Tutar, Y. Pseudogenes. *Comparative and Functional Genomics* 2012, (2012).

763. Petroski, M. D. & Deshaies, R. J. Function and regulation of cullin-RING ubiquitin ligases. *Nat. Rev. Mol. Cell Biol.* 6, 9–20 (2005).

764. Feng, L., Allen, N. S., Simo, S. & Cooper, J. A. Cullin 5 regulates Dab1 protein levels and neuron positioning during cortical development. *Genes Dev.* 21, 2717–2730 (2007).

765. Burnatowska-Hledin, M. *et al.* VACM-1 receptor is specifically expressed in rabbit vascular endothelium and renal collecting tubule. *Am J Physiol* 276, F199–209. (1999).

766. Kondoh, S. *et al.* A novel gene is disrupted at a 14q13 breakpoint of t(2;14) in a patient

with mirror-image polydactyly of hands and feet. *J. Hum. Genet.* 47, 136–9 (2002).

767. Hacein-bey-abina, S. & Schmidt, M. correspondence A Serious Adverse Event after Successful Gene Therapy for X-Linked Severe Combined Immunodeficiency. 255–266 (2003).

768. Kivioja, T. *et al.* Counting absolute numbers of molecules using unique molecular identifiers. *Nat. Methods* 9, 72–74 (2011).

769. Malausa, T. *et al.* High-throughput microsatellite isolation through 454 GS-FLX Titanium pyrosequencing of enriched DNA libraries. *Mol. Ecol. Resour.* 11, 638–644 (2011).

770. Hodges, E. *et al.* Genome-wide in situ exon capture for selective resequencing. *Nat. Genet.* 39, 1522–7 (2007).

771. Gnirke, A. *et al.* Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat. Biotechnol.* 27, 182–9 (2009).

772. Gilmartin, G. M., Fleming, E. S. & Oetjen, J. Activation of HIV-1 pre-mRNA 3' processing in vitro requires both an upstream element and TAR. *EMBO J.* 11, 4419–28 (1992).

773. Zaiss, A.-K., Son, S. & Chang, L.-J. RNA 3' readthrough of oncoretrovirus and lentivirus: implications for vector safety and efficacy. *J. Virol.* 76, 7209–7219 (2002).

774. Robasky, K., Lewis, N. E. & Church, G. M. The role of replicates for error mitigation in next-generation sequencing. *Nat. Rev. Genet.* 15, 56–62 (2014).

775. Wilson, C., Bellen, H. J. & Gehring, W. J. Position effects on eukaryotic gene expression. *Annu. Rev. Cell Biol.* 6, 679–714 (1990).

776. Yap, M. W., Kingsman, S. M. & Kingsman, a. J. Effects of stoichiometry of retroviral components on virus production. *J. Gen. Virol.* 81, 2195–2202 (2000).

777. Cong, L., Ran, F., Cox, D., Lin, S. & Barretto, R. Multiplex Genome Engineering Using CRISPR / Cas Systems. *Science (80-.).* 819, (2013).

778. Lanza, A. M., Kim, D. S. & Alper, H. S. Evaluating the influence of selection markers on obtaining selected pools and stable cell lines in human cells. *Biotechnol. J.* 8, 811–821 (2013).

779. Ma, H. *et al.* Pol III Promoters to Express Small RNAs: Delineation of Transcription Initiation. *Mol. Ther. Nucleic Acids* 3, e161 (2014).

780. Merzlyak, E. M. *et al.* Bright monomeric red fluorescent protein with an extended fluorescence lifetime. *Nat Methods* 4, 555–557 (2007).

781. ECACC. ECACC Catalogue Entry for HEK 293. *hpacultures.org.uk*

782. Kuscu, C., Arslan, S., Singh, R., Thorpe, J. & Adli, M. Genome-wide analysis reveals characteristics of off-target sites bound by the Cas9 endonuclease. *Nat Biotechnol* 32, 677–683 (2014).

783. Esvelt, K. M. *et al.* Orthogonal Cas9 proteins for RNA-guided gene regulation and editing. *Nat Methods* 10, 1116–1121 (2013).

784. Ran, F. A. *et al.* Genome engineering using the CRISPR-Cas9 system. *Nat. Protoc.* 8, 2281–2308 (2013).

785. von Groll, A., Levin, Y., Barbosa, M. C. & Ravazzolo, A. P. Linear DNA low efficiency transfection by liposome can be improved by the use of cationic lipid as charge neutralizer. *Biotechnol. Prog.* 22, 1220–4

786. Barnes, L. M., Bentley, C. M. & Dickson, A. J. Characterization of the stability of recombinant protein production in the GS-NS0 expression system. *Biotechnol. Bioeng.* 73, 261–70 (2001).

787. Barnes, L. M., Moy, N. & Dickson, A. J. Phenotypic variation during cloning procedures: analysis of the growth behavior of clonal cell lines. *Biotechnol. Bioeng.* 94, 530–537 (2006).

788. Pilbrough, W., Munro, T. P. & Gray, P. Intraclonal protein expression heterogeneity in recombinant CHO cells. *PLoS One* 4, (2009).

789. Cai, L., Friedman, N. & Xie, X. S. Stochastic protein expression in individual cells at the single molecule level. *Nature* 440, 358–362 (2006).

790. Raj, A., Peskin, C. S., Tranchina, D., Vargas, D. Y. & Tyagi, S. Stochastic mRNA synthesis in mammalian cells. *PLoS Biol.* 4, 1707–1719 (2006).

791. Kim, N. S., Kim, S. J. & Lee, G. M. Clonal variability within dihydrofolate reductase-mediated gene amplified Chinese hamster ovary cells: stability in the absence of selective pressure. *Biotechnol. Bioeng.* 60, 679–688 (1998).

792. Sung, K. Y., Sun, O. H. & Gyun, M. L. Enhancing effect of low culture temperature on specific antibody productivity of recombinant Chinese hamster ovary cells: Clonal variation.

Biotechnol. Prog. 20, 1683–1688 (2004).

793. Kim, S. H. & Lee, G. M. Differences in optimal pH and temperature for cell growth and antibody production between two Chinese hamster ovary clones derived from the same parental clone. *J. Microbiol. Biotechnol.* 17, 712–720 (2007).

794. Yang, Y. & Seed, B. Site-specific gene targeting in mouse embryonic stem cells with intact bacterial artificial chromosomes. *Nat. Biotechnol.* 21, 447–51 (2003).

795. Hendel, A. *et al.* Quantifying genome-editing outcomes at endogenous loci with SMRT sequencing. *Cell Rep.* 7, 293–305 (2014).

796. Wang, T., Wei, J. J., Sabatini, D. M. & Lander, E. S. Genetic screens in human cells using the CRISPR-Cas9 system. *Science* 343, 80–4 (2014).

797. Merkle, F. T. *et al.* Efficient CRISPR-Cas9-Mediated Generation of Knockin Human Pluripotent Stem Cells Lacking Undesired Mutations at the Targeted Locus. *Cell Rep.* 11, 875–883 (2015).

798. Marshall, H. M. *et al.* Role of PSIP 1/LEDGF/p75 in lentiviral infectivity and integration targeting. *PLoS One* 2, (2007).

799. Wu, X., Li, Y., Crise, B., Burgess, S. M. & Munroe, D. J. Weak Palindromic Consensus Sequences Are a Common Feature Found at the Integration Target Sites of Many Retroviruses Weak Palindromic Consensus Sequences Are a Common Feature Found at the Integration Target Sites of Many Retroviruses. 79, 5211–5214 (2005).

800. Birol, I. *et al.* Assembling the 20 Gb white spruce (Picea glauca) genome from whole-genome shotgun sequencing data. *Bioinformatics* 29, 1492–1497 (2013).

801. Roberts, R. J., Carneiro, M. O. & Schatz, M. C. The advantages of SMRT sequencing. *Genome Biol.* 14, 405 (2013).

802. Trachtenberg, E. A. & Holcomb, C. L. Next-generation HLA sequencing using the 454 GS FLX system. *Methods Mol. Biol.* 1034, 197–219 (2013).

803. Hengen, P. N. *Shearing DNA for genomic library construction*. *Trends Biochem Sci* 22, 273–274 (1997).

804. Thorstenson, Y. R., Hunicke-Smith, S. P., Oefner, P. J. & Davis, R. W. An automated hydrodynamic process for controlled, unbiased DNA shearing. *Genome Res.* 8, 848–855 (1998).

805. Knierim, E., Lucke, B., Schwarz, J. M., Schuelke, M. & Seelow, D. Systematic comparison of three methods for fragmentation of long-range PCR products for next generation sequencing. *PLoS One* 6, (2011).

806. Brady, T. *et al.* A method to sequence and quantify DNA integration for monitoring outcome in gene therapy. *Nucleic Acids Res.* 39, 1–8 (2011).

807. Paruzynski, A. *et al.* Genome-wide high-throughput integrome analyses by nrLAM-PCR and next-generation sequencing. *Nat. Protoc.* 5, 1379–1395 (2010).

808. Barbosa, C., Peixeiro, I. & Romão, L. Gene Expression Regulation by Upstream Open Reading Frames and Human Disease. *PLoS Genetics* 9, (2013).

809. Ng, S. K., Wang, D. I. C. & Yap, M. G. S. Application of destabilizing sequences on selection marker for improved recombinant protein productivity in CHO-DG44. *Metab. Eng.* 9, 304–316 (2007).

810. Ikeda, Y. *et al.* Continuous high-titer HIV-1 vector production. *Nat. Biotechnol.* 21, 569–572 (2003).

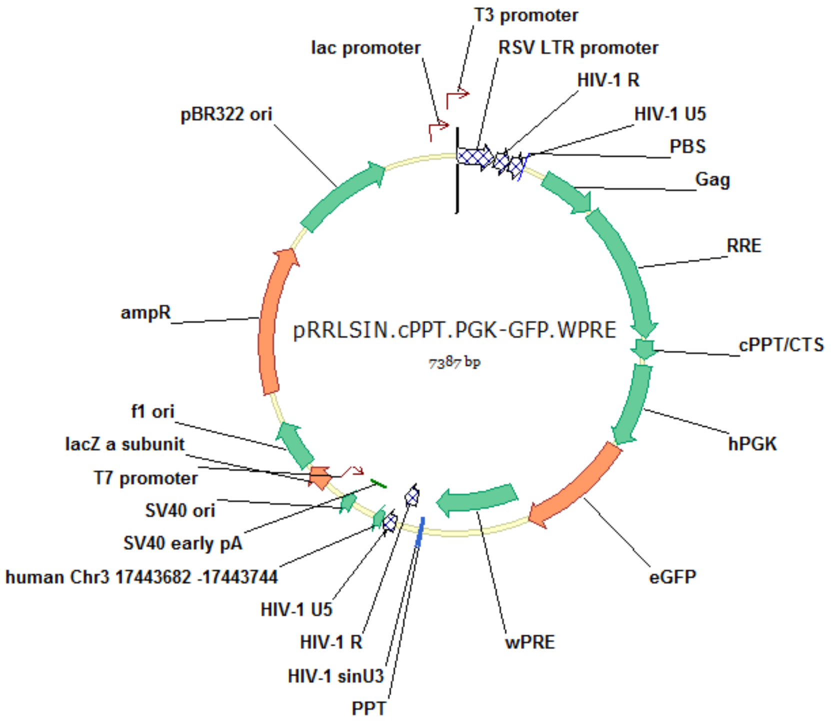
811. Bukovsky, a a, Song, J. P. & Naldini, L. Interaction of human immunodeficiency virus-derived vectors with wild-type virus in transduced cells. *J. Virol.* 73, 7087–7092 (1999).

812. Lucke, S., Grunwald, T. & Uberla, K. Reduced mobilization of Rev-responsive element-deficient lentiviral vectors. *J. Virol.* 79, 9359–62 (2005).

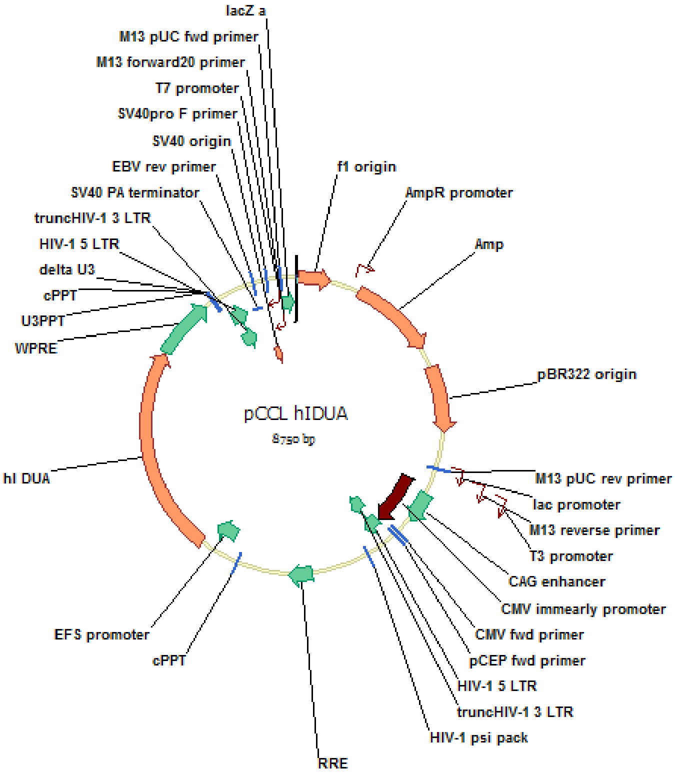
813. Donnelly, M. L. L. *et al.* Analysis of the aphthovirus 2A/2B polyprotein 'cleavage' mechanism indicates not a proteolytic reaction, but a novel translational effect: A putative ribosomal 'skip'. *J. Gen. Virol.* 82, 1013–1025 (2001).

Appendix

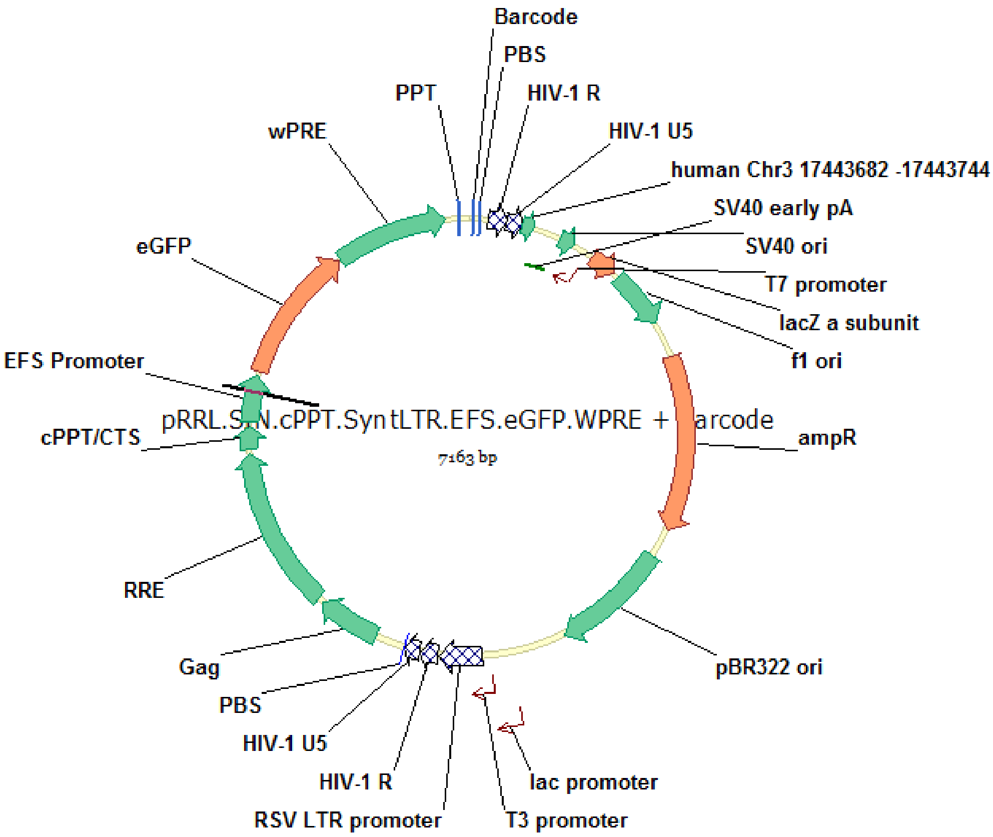
A. Plasmids



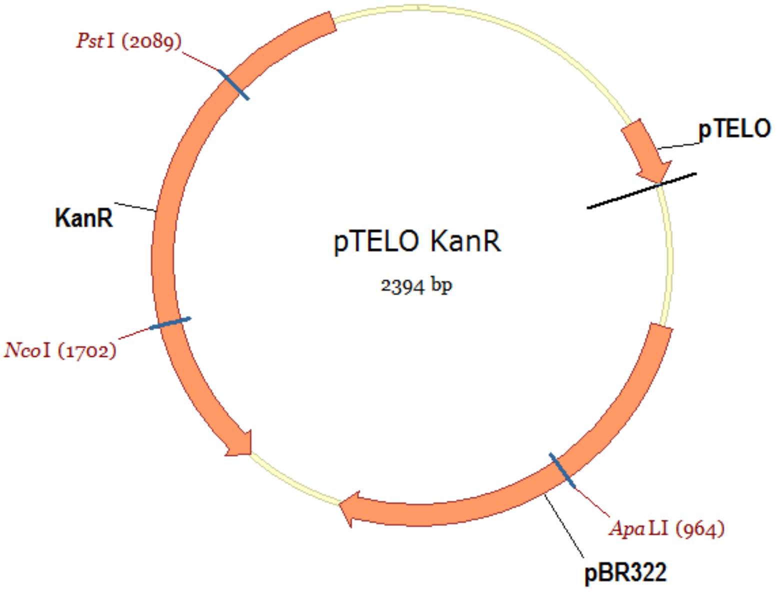
Plasmid map of pRRL SIN cPPT PGK eGFP WPRE



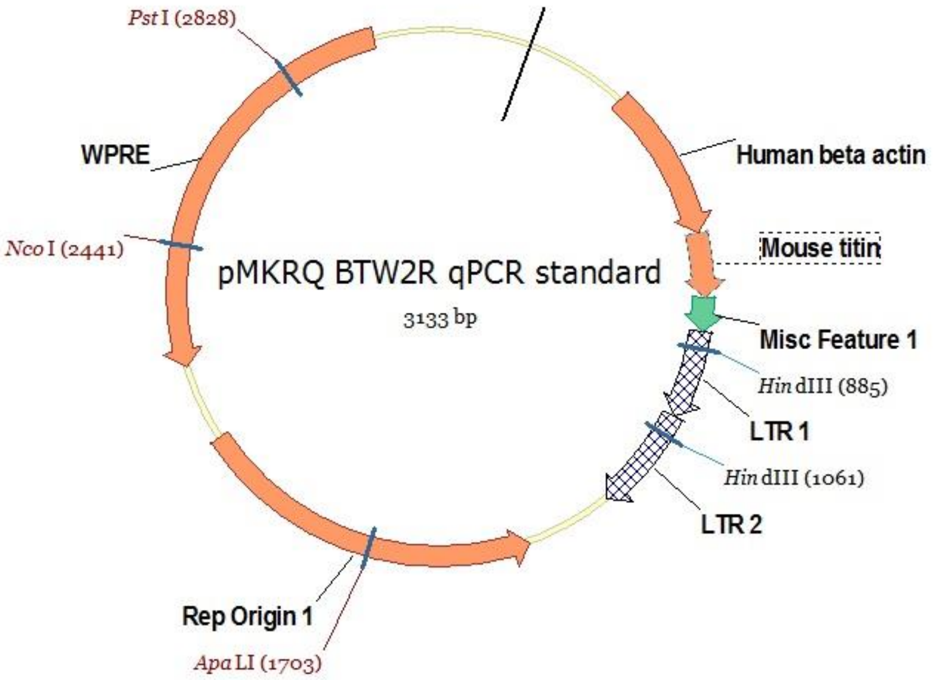
Plasmid map of pCCLSIN hIDUA



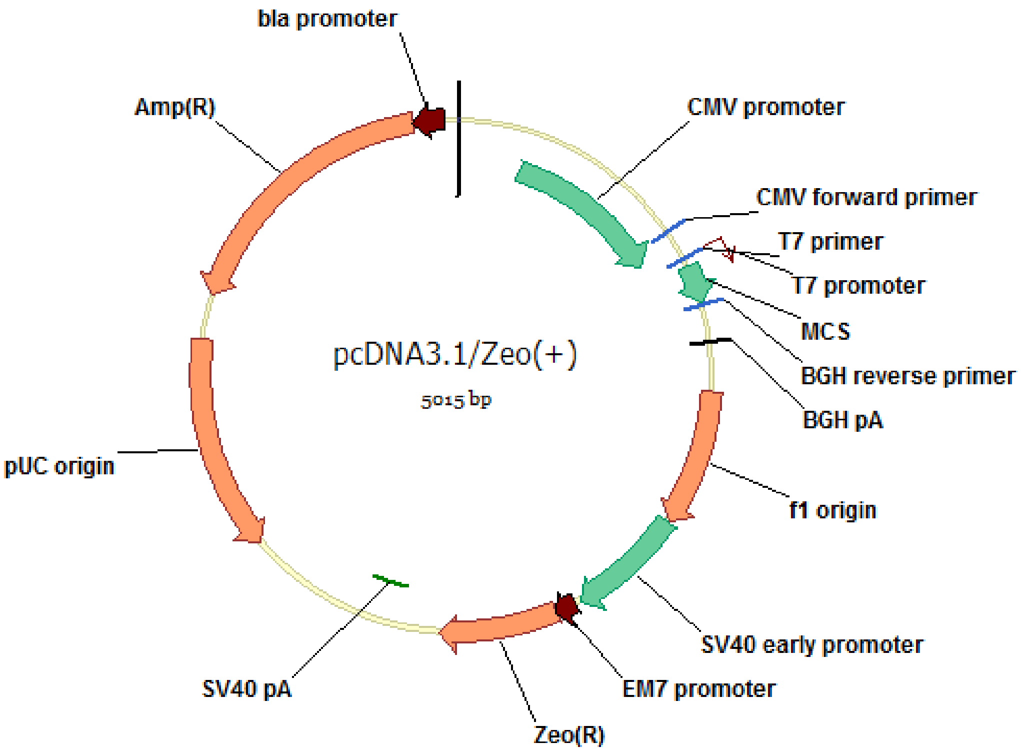
Plasmid map of pRRL SIN Synthetic LTR cPPT EFS eGFP WPRE + barcoded library (pSYNT)



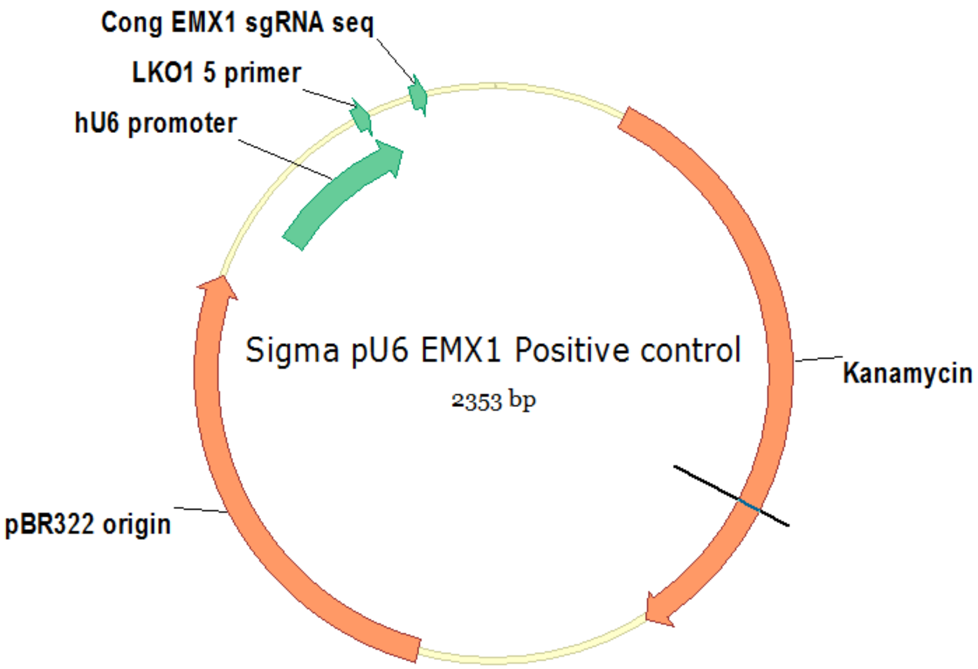
Plasmid map of pTELO



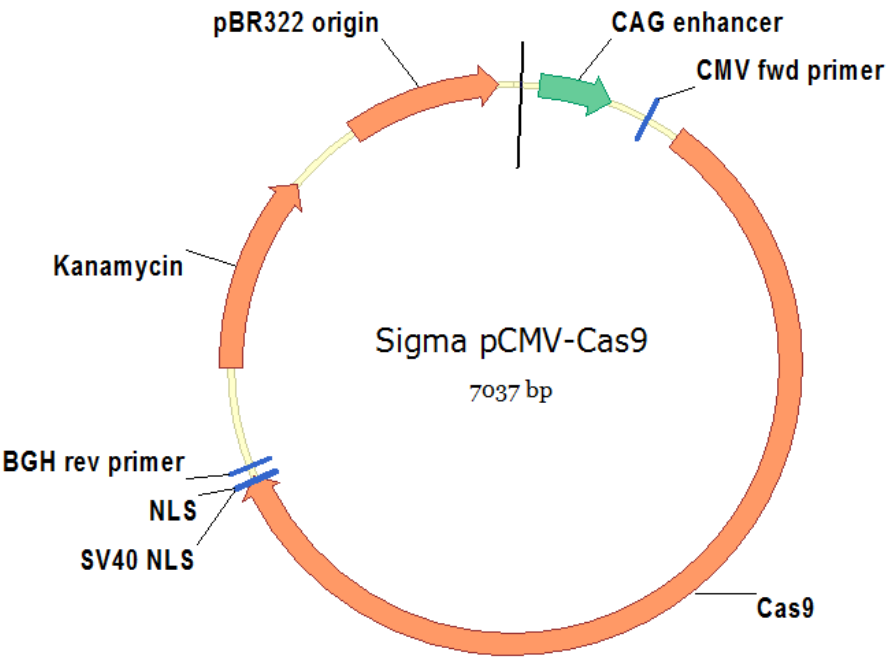
Plasmid map of pMKRQ BTW2R (positive control for qPCR)



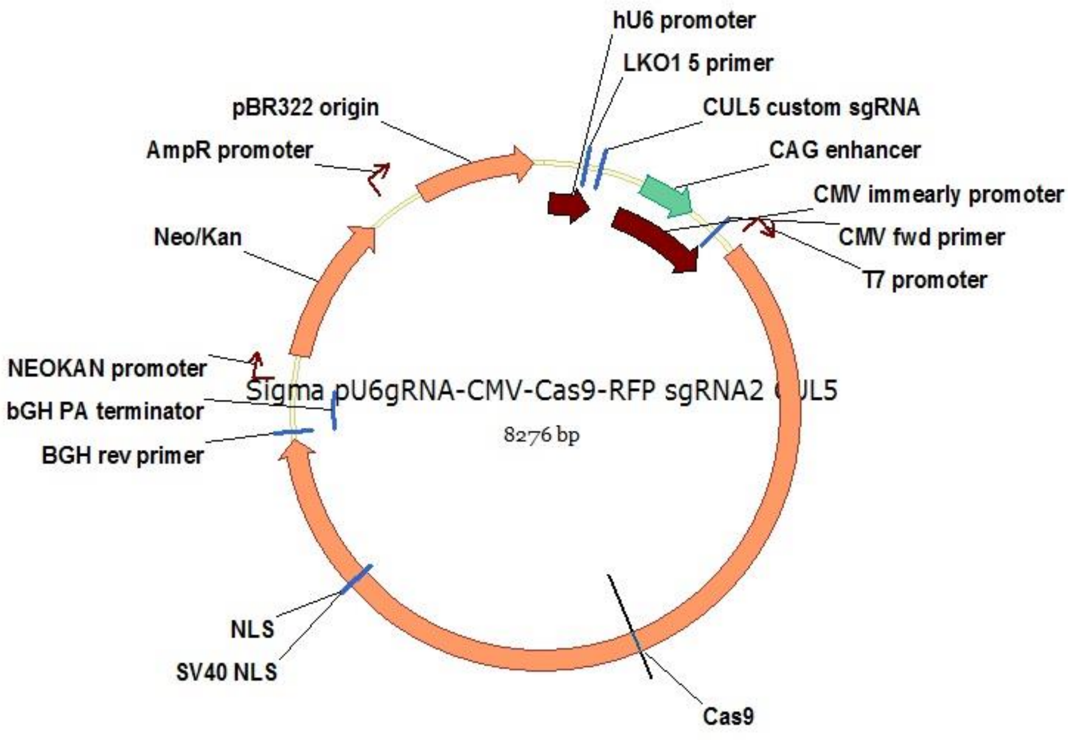
Plasmid map of pcDNA 3.1/Zeo from Thermo Fisher Scientific (Cat no. V86020)



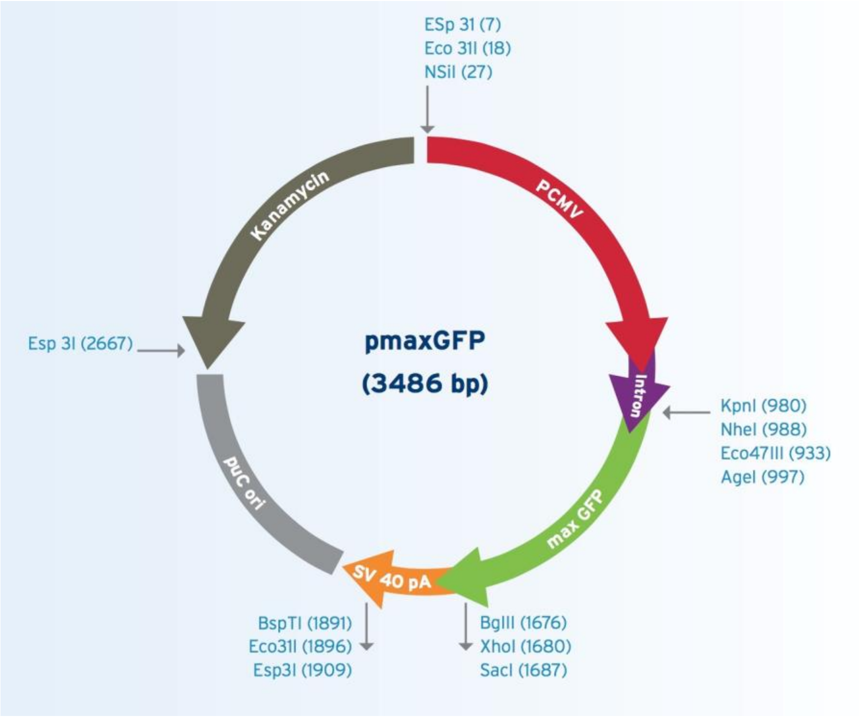
Plasmid map of pU6-sgRNA EMX1 20nt position from Cong *et al.*, from Sigma (Cat no.CRISPR01).



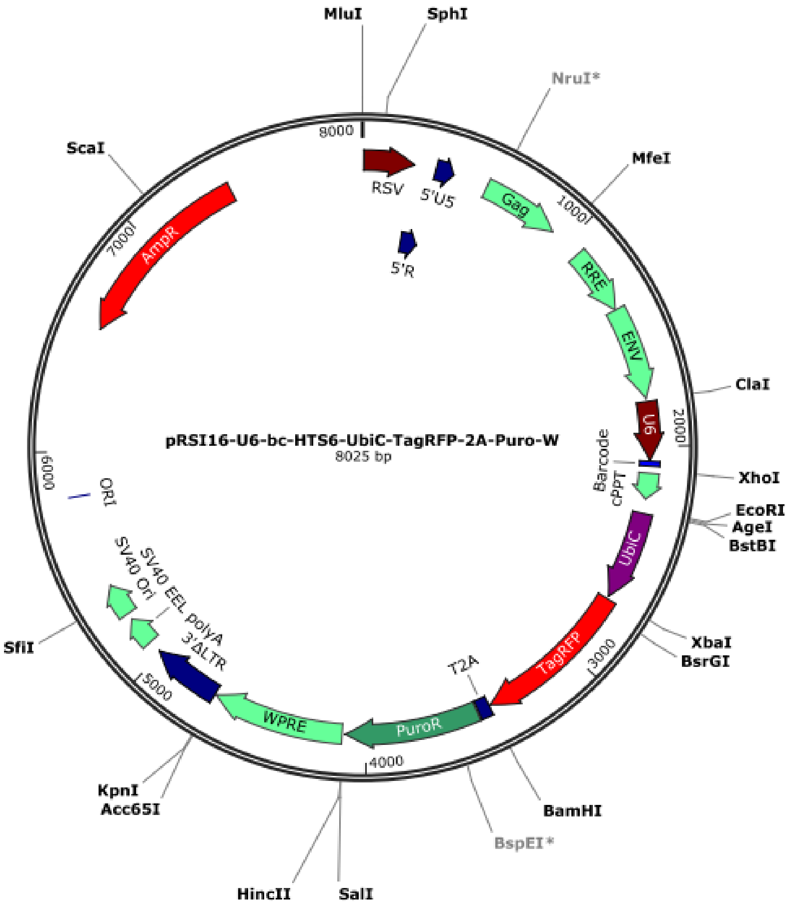
Plasmid map of pCMV-Cas9 from Sigma (CAS9P)



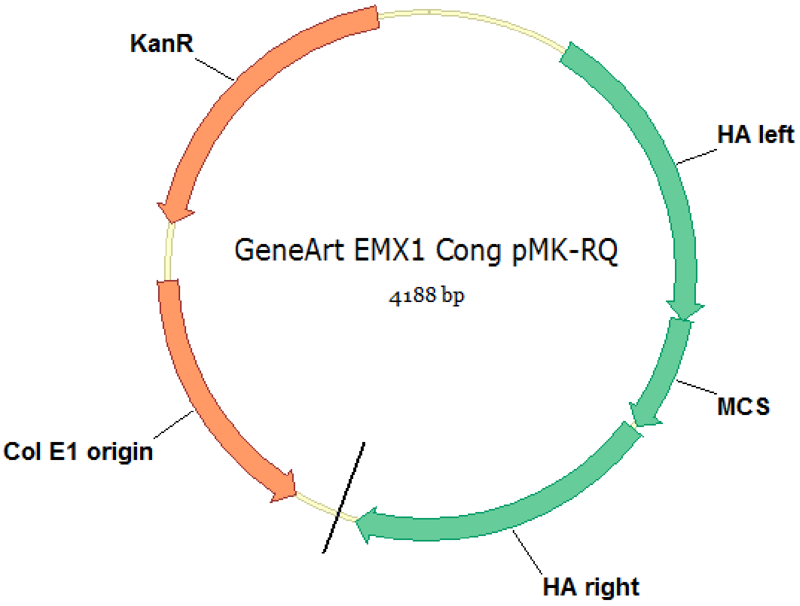
Plasmid map of pCMV-Cas9 U6-sgRNA from Sigma (CUL 5 sgRNA)



Plasmid map of pmax GFP (Nucleofection Kit V. Lonza VCA-1003)



Plasmid Map of CellTracker® Lentiviral Barcode Library Vector (Cellecta)



Plasmid map of GeneArt HA1-MCS-HA2 (Cong *et al.*, EMX1)

Bioinformatic scripts

Extract_barcode_library

```
use warnings;

$five_prime = "GACAAGATCCATATGAGTAA";
$N = qr/[ACGT]/;
$W = qr/[AT]/;
$S = qr/[GC]/;
$barcode = qr/($N$N$N)ATC($N$S)GAT($N$N)AAA($N$N)GGT($N$W)AAC($N$N)TGA($N$N$N)/;
$three_prime  = "TGGTAACACCGACTAGGATC";

$matched = 0;
while (<>) {
    /^@/ or die;
    chomp;
    /^\S+/;
    $read_name = $&;
    chomp($read = <>);
    <>;
    <>;

    if (@barcode = ($read =~ /$five_prime$barcode$three_prime/)) {
        print "$read_name\t" . join("-", @barcode) . "\n";
        $matched++;
    }
}

print STDERR "Found $matched read (pairs) with a barcode.\n";
```

Extract_viral_insertion_barcodes

```
use warnings;
$input_file = shift;

$MIN_LINKER_SCORE = 40;
$MIN_LTR_SCORE = 60;
$MIN_LTR2_SCORE = 30;
$N = qr/[ACGT]/;
$W = qr/[AT]/;
$S = qr/[GC]/;
$barcode_pattern                                                         =
qr/($N$N$N)ATC($N$S)GAT($N$N)AAA($N$N)GGT($N$W)AAC($N$N)TGA($N$N$N)/;

if (-e "$input_file.vs_linker") {
    print STDERR "Skipping alignment to linker sequence as file '$input_file.vs_linker'
already exists.\n";
}
else {
    print STDERR "Aligning merged sequences against Linker...\n";
    system "./ssearch36 -3 -m 8 $input_file linker.fa > $input_file.vs_linker";
}

if (-e "$input_file.vs_ltr") {
    print  STDERR  "Skipping  alignment  to  ltr  sequence  as  file  '$input_file.vs_ltr'
already exists.\n";
}
else {
    print STDERR "Aligning merged sequences against LTR1...\n";
    system "./ssearch36 -3 -m 8 $input_file ltr.fa > $input_file.vs_ltr";
}
```

```
if (-e "$input_file.vs_ltr2") {
    print STDERR "Skipping alignment to ltr2 sequence as file '$input_file.vs_ltr2'
already exists.\n";
}
else {
    print STDERR "Aligning merged sequences against LTR2...\n";
    system "./ssearch36 -3 -m 8 $input_file ltr2.fa > $input_file.vs_ltr2";
}

print STDERR "Collecting alignment coordinates...\n";
open LINKER, "$input_file.vs_linker" or die $!;
open LTR,    "$input_file.vs_ltr" or die $!;
open LTR2,   "$input_file.vs_ltr2" or die $!;
open SEQ,    $input_file or die $!;

$matched = $total = 0;
while ($linker = <LINKER>) {
    $ltr = <LTR>;
    $ltr2 = <LTR2>;
    <SEQ>;
    $seq = <SEQ>;
    chomp $seq;
    $total++;

    @L1 = split /\t/, $linker;
    @L2 = split /\t/, $ltr;
    @L3 = split /\t/, $ltr2;

    # Sanity check
    if ($L1[0] ne $L2[0] || $L1[0] ne $L3[0]) {
        die;
    }

    # A read that passes must align to linker and both LTRs.
    # In this case extract barcode and host sequence.
    # Finally make sure that the barcode is of right length and pattern.
    if ($L1[11] < $MIN_LINKER_SCORE) { next; }
    if ($L2[11] < $MIN_LTR_SCORE) { next; }
    if ($L3[11] < $MIN_LTR2_SCORE) { next; }

    # Extract the host sequence and the potential barcode sequence
    if ($L2[6] <= $L1[7]) {
        $host_seq = "";
    }
    else {
        $host_seq = substr($seq, $L1[7], $L2[6] - $L1[7] - 1);
    }
    if ($L3[6] <= $L2[7]) {
        $barcode = "";
    }
    else {
        $barcode = substr($seq, $L2[7], $L3[6] - $L2[7] - 1);
    }

    # Barcode must match the expected pattern
    if (@barcode = ($barcode =~ $barcode_pattern)) {
        print join(
            "\t",
            $L1[0],
            join("-", @barcode),
            $host_seq,
        ) . "\n";

        $matched++;
    }
}
```

```
print STDERR "Found $matched/$total read pairs with a barcode and host sequence.\n";
```

fastq_to_fasta

```
        use warnings;

        while ($name = <>) {
            $name =~ s/^\@/>/ or die;
            $seq = <>;
            <>;
            <>;

            print $name;
            print $seq;
}
```

extract_rt-pcr_barcodes.pl

```
        use warnings;
        $input_file = shift;

        $MIN_LTR_SCORE = 60;
        $MIN_LTR2_SCORE = 30;
        $N = qr/[ACGT]/;
        $W = qr/[AT]/;
        $S = qr/[GC]/;
        $barcode_pattern                                                    =
        qr/($N$N$N)ATC($N$S)GAT($N$N)AAA($N$N)GGT($N$W)AAC($N$N)TGA($N$N$N)/;

        if (-e "$input_file.vs_ltr") {
            print   STDERR   "Skipping   alignment   to   ltr   sequence   as   file
        '$input_file.vs_ltr' already exists.\n";
        }
        else {
            print STDERR "Aligning merged sequences against LTR1...\n";
            system "./ssearch36 -3 -m 8 $input_file ltr.fa > $input_file.vs_ltr";
        }

        if (-e "$input_file.vs_ltr2") {
            print   STDERR   "Skipping   alignment   to   ltr2   sequence   as   file
        '$input_file.vs_ltr2' already exists.\n";
        }
        else {
            print STDERR "Aligning merged sequences against LTR2...\n";
            system "./ssearch36 -3 -m 8 $input_file ltr2.fa > $input_file.vs_ltr2";
        }

        print STDERR "Collecting alignment coordinates...\n";
        open LTR,    "$input_file.vs_ltr" or die $!;
        open LTR2,   "$input_file.vs_ltr2" or die $!;
        open SEQ,    $input_file or die $!;

        $matched = $total = 0;
        while ($ltr = <LTR>) {
            $ltr2 = <LTR2>;
            <SEQ>;
            $seq = <SEQ>;
            chomp $seq;
            $total++;

            @L1 = split /\t/, $ltr;
            @L2 = split /\t/, $ltr2;
            # Sanity check
            if ($L1[0] ne $L2[0]) {
                die;
            }

            # A read that passes must align both LTRs.
            # Extract barcode and host sequence.
            # Finally make sure that the barcode is of right length and pattern.
            if ($L1[11] < $MIN_LTR_SCORE) { next; }
            if ($L2[11] < $MIN_LTR2_SCORE) { next; }

            # Extract the barcode sequence
            if ($L2[6] <= $L1[7]) {
                $barcode = "";
            }
            else {
                $barcode = substr($seq, $L1[7], $L2[6] - $L1[7] - 1);
            }

            # Barcode must match the expected pattern
            if (@barcode = ($barcode =~ $barcode_pattern)) {
                print join(
                    "\t",
                    $L1[0],
                    join("-", @barcode),
                ) . "\n";

                $matched++;
            }
        }

        print STDERR "Found $matched/$total read pairs with a barcode and host
        sequence.\n";
```

get_best_hit_from_psl

```
use warnings;

$MIN_ID_THRESHOLD = 0.999;

# Parse first line
chomp($_ = <>);
@F = split /\t/;
if (
    $F[10] != $F[12] ||  # Last base of the host sequence was aligned?
    !last_base_is_a_match($F[8], $F[21], $F[22])  # The last base of the host sequence
was a match?
) {
    $last_read = "";
}
else {
    $last_read = $F[9];
    $last_score = get_identity($F[10], $F[1], $F[5], $F[7]);
    $last_strand = $F[8];
    $last_chr = $F[13];
    $last_start = $F[15];
    $last_end = $F[16];
    $last_length = $F[10];
    $multi_best_hit = 1;
}

while (<>) {
    chomp;
    @F = split /\t/;

    # As above, only accept host sequence alignments where the final base
    # was part of the alignment and aligned to the reference as a match.
    if (
        $F[10] != $F[12] ||
        !last_base_is_a_match($F[8], $F[21], $F[22])
    ) {
        next;
    }

    $cur_score = get_identity($F[10], $F[1], $F[5], $F[7]);

    # If we encountered a new host sequence, then print the best alignment
    # of the previous host sequence (if identity threshold is exceeded).
    # If we are still traversing the alignments of the current host sequence,
    # then just keep colleting alignments.
    if ($F[9] ne $last_read) {
        # Print only if score/length >= identity
        if (
            $last_read ne "" &&
            $last_score > $MIN_ID_THRESHOLD
        ) {
            print join(
                "\t",
                $last_chr,
                $last_start-1,  # BED format start is 0-based
                $last_end,
                $last_read,
                $last_score,
                $last_strand,
                $multi_best_hit
            ) . "\n";
        }

        $last_read = $F[9];
        $last_score = $cur_score;
        $last_strand = $F[8];
        $last_chr = $F[13];
        $last_start = $F[15];
        $last_end = $F[16];
        $last_length = $F[10];
        $multi_best_hit = 1;
    }
    else {
        if ($cur_score > $last_score) {
            $last_read = $F[9];
            $last_score = $cur_score;
            $last_strand = $F[8];
            $last_chr = $F[13];
            $last_start = $F[15];
            $last_end = $F[16];
            $last_length = $F[10];
            $multi_best_hit = 1;
        }
        elsif ($cur_score == $last_score) {
            $multi_best_hit++;
        }
    }
}

if (
    $last_read ne "" &&
    $last_score > $MIN_ID_THRESHOLD
) {
    print join(
        "\t",
        $last_chr,
        $last_start-1,  # BED format start is 0-based
        $last_end,
        $last_read,
        $last_score,
        $last_strand,
        $multi_best_hit
    ) . "\n";
}


sub last_base_is_a_match {
    if ($_[0] eq "+") {
        return substr($_[1], -2, 1) eq substr($_[2], -2, 1);
    }
    else {
        return substr($_[1], 0, 1) eq substr($_[2], 0, 1);
    }
}

sub get_identity {
    ($length, $mismatches, $insertions, $deletions) = @_;
    return(1 - ($mismatches + $insertions + $deletions) / ($length + $insertions +
$deletions));
}
```

plot_plasmid_library_distributions.R

```
library(stringdist)

files = c(
    # "PlasmidPCR_11.barcodes.txt",
    # "PlasmidPCR_49.barcodes.txt"
    "PlasmidPCR_11.starcode_barcodes.txt",
    "PlasmidPCR_49.starcode_barcodes.txt"
)
minimum_frequency = 2

cat("Running plot_plasmid_library_distributions.R...\n", file = stderr())
cat("Parameters:\n", file = stderr())
cat(paste("Input files: ", paste(files, collapse = " "), "\n", sep = ""), file
= stderr())
cat(paste("Minimum frequency: ", minimum_frequency, "\n", sep = ""), file =
stderr())

for (f in files) {
    cat(paste("Analysing file ", f, "...\n", sep = ""), file = stderr())
    sample = sub("\\..+", "", f)
    d = read.table(f, header = F, sep = "\t", stringsAsFactors = F)
    c = table(d[,2])
    c.f = c[c>1]

    pdf(paste(sample, ".clustering.pdf", sep = ""), w = 20, h = 8)
    par(mfrow = c(2, 1), mar = c(0, 4, 4, 2) + .1)

    hclust_res = hclust(as.dist(stringdistmatrix(names(c.f), names(c.f), method
= "hamming")))
    plot(hclust_res, ylab = "Number of differences", xlab = "", sub =
paste("Minimum frequency: ", minimum_frequency, sep = ""), labels = F, main =
sample, lwd = 0.5)

    par(mar = c(5, 4, 0, 2) + .1)
```

```
            plot(hclust_res$order, c.f, type = "h", bty = "n", ylim = c(3000, 0), xlab
        = "Barcodes", ylab = "Frequency", lwd = 0.5, xaxt = "n")
            dev.off()

            pdf(paste(sample, ".distribution.pdf", sep = ""))
            plot(sort(c.f),  (1:length(c.f))/length(c.f),  type  =  "l",  main  =  sample,
        xlab = "Frequency of barcode", ylab = "Cumulative density")
            dev.off()
        }
```

Barcode_error_correction.R

```
#Determine dissimilarity between barcodes
#create a results set, which is as long as the number of barcodes.
results <- matrix(data=NA,nrow=dim(vals)[1],ncol=dim(vals)[1])

# Split the barcode string and compare the barcodes one by one, this is embarrasingly
parallel
# and could be done much faster (in a little bit more complicated way) than shown here.
system.time(
        for (i in 1:dim(dat)[1]) {
                for (j in 1:dim(dat)[1]) {
                        results[i,j] <- sum(unlist(strsplit(rownames(dat)[i], split=""))
!=  unlist(strsplit(rownames(dat)[j], split="")))
                }
        }
)

#The results matrix now contains the number of bases that differ between each barcode
#A histogram of the dissimilarity between the codes.
pdf("Dissimillarity histogram.pdf")
hist(results, main="Histogram of barcode dissimilarities")
dev.off()

# now threshold on the allowed number of mismatches (here 2)
results2 <- results
results2[results2 > 2 ] <- 0

# load the igraph library and generate an graph based on the adjacency matrix we made
library(igraph)
g1 <- graph.adjacency(results2)


# now can we use the cluster membership from the graph to determine which barcodes are
# similar and sum all rows that belong to the same cluster.
processed.dat <- rowsum(dat, clusters(g1)$membership)


# Normalize the counts by the sum of the column to get a matrix of error-corrected,
normalized
# counts.
normvals <- processed.dat/colSums(processed.dat)[col(processed.dat)]

# Sort the matrix from highest to lowest number of normalized counts
sums <- apply(normvals, 1, sum)
sorter <- order(sums, decreasing=TRUE)
normvals <- normvals[sorter,]

# plot the data

library(plotrix)
plot.colors<-
c("#004586FF","#FF420eFF","#FFD320FF","#579D1CFF","#7e0021ff","#83caffff","#314004ff"
,"#aecf00ff","#4b1f6fff","#ff950eff","#c5000bff","#0084d1ff")
stackpoly(t(normvals), stack=TRUE, col=plot.colors, xaxlab=colnames(normvals))
```

MHB08-059_check_and_assign_pSYNT

```
#! /usr/bin/perl

# Creative Commons Attribution-Share Alike 3.0 Netherlands License
# 21 March 2013, .
# use lib '/home/mhb/Desktop/ensembl/Ensembl/ensembl/modules';

use warnings;
use strict;
use Cwd;
use Bio::SeqIO;



if ($#ARGV+1 == 0) {die "Please call with directory name\n"};
print "Called with ",$#ARGV+1," parameters which ",
        @ARGV == 1 ? "was" : "were" ,"\n";

print ("$_\n") foreach (@ARGV) ;

#The OUTPUT FILE goes here
my $outfile = "$ARGV[0]" . ".txt";
open OUT, ">$outfile" || die "cannot open $outfile\n";

sleep 1;

my $dir = $ARGV[0];
opendir DIR, $dir || die "Cannot open $dir\n";


#Put files in an array
my @files = grep { !/^\./ && -f "$dir/$_" }readdir DIR;
print "FILES: @files\n";
sleep 1;

my %samplehash;
my %filehash;
my %barcodestore;

# using hot pipe
$|=1;

foreach my $testfile (sort @files) {
        my $infile = "$dir/$testfile";
        #$testfile =~ /(PTGZ_\d{3}.)/;
        #$testfile =~ /(V11/;
        my $filename_for_hash = $testfile;
        $filehash{$filename_for_hash}++;
        my $seqio_object = Bio::SeqIO->new(-file => "<$infile");
        #hash to store barcodes in

        my $missed=0;
        my $counter=0;

        print "Performing Sample Barcode lookup... $infile\n";
        while (my $seq_object = $seqio_object->next_seq()) {
                $counter++;
                my $test = $seq_object -> seq();
                my $id = $seq_object -> display_id() ;
                my $seq = $seq_object -> seq();
                my $length = $seq_object-> length();

                my $samplecode=0;
                # match either seed code in front or after the barcode
                #if ($test =~ /ACAAGTAAGG(.{33})/){  #MATCH 33 bp after the key sequence
                #       $samplecode=$1;
                #}
                #elsif ($test =~ /(.{33})GACGGCCAGTG/){
                #       $samplecode=$1;#put barcode in hash
                #}

                #or match the barcode
                #if($test =~ /(GG.{3}AC.{3}GT.{3}CG.{3}TA.{3}CA.{3}TG.{3}GA)/) {  #PTGZ
                if($test   =~   /(.{3}ATC.{2}GAT.{2}AAA.{2}GGT.{2}AAC.{2}TGA.{3})/)   {
#PSYNT
```

```
                        $samplecode=$1;
                        #print "$samplecode\n";
                }
                else {
                        $missed++;
                        #print "Not matched, $missed\n";
                        next;
                }
                # store barcode, then filename, then value
                $samplehash{$filename_for_hash}{$samplecode}++;

                # also keep a record of all barcodes we've encountered
                $barcodestore{$samplecode}++;
        }
}


print "Reporting:\n";

# print a tab for good alignment in R: IS THIS NEEDED?
#print OUT "\t";
foreach my $barcode (keys %barcodestore) { #code
                print OUT "$barcode\t";
}
print OUT "\n";

foreach my $sample (sort keys %samplehash) { #25mar added sort here
        print OUT "$sample\t";
        foreach my $barcode2 (keys %barcodestore) { #code
                if    (exists    $samplehash{$sample}->{$barcode2})    {    print    OUT
$samplehash{$sample}->{$barcode2} . "\t" } else {print OUT "0\t"};

        }
        print OUT "\n";
}

exit;
```

make_ucsc_gene_txs.sh.

```
tail -n +2 ensembl_hg19_protein_coding_genes_with_ccds_id.txt | cut -f1-6 | sort -u |
perl -aF/\\t/ -ne '$F[3]--; print join("\t", @F[0,3,4,1,2], ($F[5] == -1 ? "-" : "+"))
. "\n"' > hg19_genes.bed
tail -n +2 ensembl_hg19_protein_coding_genes_with_ccds_id.txt | cut -f1-2,6-9 | perl -
aF/\\t/ -ne 'chomp @F; $F[4]--; print join("\t", @F[0,4,5,1,3], ($F[2] == -1 ? "-" :
"+")) . "\n"' > hg19_txs.bed
perl -aF/\\t/ -ne 'chomp @F; if ($F[5] eq "+") { print join("\t", $F[0], $F[1], $F[1]+1,
$F[3], $F[4], $F[5]) . "\n" } else { print join("\t", $F[0], $F[2]-1, $F[2], @F[3..5])
. "\n" }' hg19_txs.bed > hg19_tss.bed

gunzip -c hg19_cpg_islands.bed.gz | perl -pe 's/^chr//' | gzip -c > temp.gz
mv temp.gz hg19_cpg_islands.bed.gz
```

Columbus Script: Off-target integration (% of BFP out of GFP+ cells)

Input Image	**Stack Processing :** Individual Planes **Flatfield Correction :** None		
Calculate Image (2)		**Method :** By Formula Formula : A-B Channel A : FITC Channel B : DAPI Negative Values : Set to Zero Undefined Values : Set to Local Average	Output Image : Green only
Find Cells	**Channel :** Green only **ROI :** None	**Method :** M Diameter : 40 µm Splitting Coefficient : 0.4 Common Threshold : 0.4	Output Population : Green only cells
Calculate Image		**Method :** By Formula Formula : A+B Channel A : DAPI Channel B : FITC Negative Values : Set to Zero Undefined Values : Set to Local Average	Output Image : Blue and Green
Find Cells (3)	**Channel :** Blue and Green **ROI :** None	**Method :** M Diameter : 40 µm Splitting Coefficient : 0.4 Common Threshold : 0.4	Output Population : Total cells
Define Results	**Method :** List of Outputs **Population : Green only cells** Number of Objects Apply to All : None **Population : Total cells** Number of Objects Apply to All : None **Method :** Formula Output Formula : a/b*100 Population Type : Objects Variable A : Green only cells - Number of Objects Variable B : Total cells - Number of Objects Output Name : % green only **Population : Green only cells :** None Population : Total cells : None		

Thermo Fisher Scientific, 1kb plus DNA ladder (Cat no. 10787018)