Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

FACULTY
OF MATHEMATICS, INFORMATICS
AND NATURAL SCIENCES

University of Hamburg

Faculty of Mathematics, Informatics and Natural Sciences

Tracking Clonal Composition of Stem Cell Transplants by Genetic Barcoding and Implementation of a Targeted Barcoding Procedure

---

Dissertation submitted to the University of Hamburg by

Frauke Gotzhein

In partial fulfillment of the requirements for the degree of

Doctor of Natural Sciences (Dr. rer. nat.)

January 2021

Submitted on January 20, 2021

Supervisor: Prof. Dr. Ingo Müller
Co-Supervisor: Prof. Dr. Wolfram Brune

Thesis defense on March 19, 2021

In the presence of the committee members

Prof. Dr. Hartmut Schlüter
Prof. Dr. Ingo Müller
Prof. Dr. Michael Kolbe

The practical work of this thesis was conducted from 06/2017—12/2020 in the Research Institute Children's Cancer Center Hamburg in the group of Prof. Dr. Ingo Müller (UKE, Department of Pediatric Hematology and Oncology, Division of Pediatric Stem Cell Transplantation and Immunology).

List of figures

Figure 1: Common model of discrete hematopoietic hierarchy. 6
Figure 2: Schematic representation of the lentiviral particle structure. 11
Figure 3: Insertional mutagenesis and safety strategies. 14
Figure 4: BC32 system for cellular barcoding and tracking. 16
Figure 5: CRISPR/Cas9 gene editing process. 20
Figure 6: Experimental procedure and transplant composition. 24
Figure 7: Methods of detection of BC32 vectors. 25
Figure 8: Engraftment and hematopoietic reconstitution after stem cell transplantation..... 27
Figure 9: Contribution of barcoded HSPCs-derived cells to BM reconstitution after HSCT. 29
Figure 10: Contribution of barcoded HSPCs-derived cells to splenic reconstitution after HSCT. 30
Figure 11: Quantification of the barcode backbone by NGS. 32
Figure 12: Short-term spatial dynamics of hematopoietic reconstitution. 34
Figure 13: Long-term spatial dynamics of hematopoietic reconstitution. 35
Figure 14: Quantification of the barcode backbone by ddPCR. 37
Figure 15: HSPC contribution in mice experiments with higher initial transduction rates. 39
Figure 16: Schematic overview of the CRISPR/Cas9-mediated knock-in strategy. 41
Figure 17: Comparison of ILVs and IDLVs. 43
Figure 18: Total efficiency and specificity of Cas9 to induce DSBs at different loci. 45
Figure 19: Knock-in efficiency in bulk and single cells based on the delivery method. 46
Figure 20: Cloning procedure for ID-containing BC32 constructs. 80

List of tables

Table 1: Predicted efficiencies of gRNAs. 42
Table 2: Ready-made commercial kits utilized. 64
Table 3: Antibodies utilized for sort and staining of murine cells. 64
Table 4: Enzymes used for amplification and cloning. 65
Table 5: Media, buffer and additives used for their preparation. 66
Table 6: Primer and oligonucleotides sorted by their specific purposes. 67
Table 7: Original plasmids used for cloning and lentivirus packaging. 70
Table 8: Software and online tools used for data handling. 70
Table 9: Instruments used for analysis and experimental setup. 71
Table 10: Consumables used for ddPCR and mouse experimental procedures. 71
Table 11: Assignment of different barcode backbones to fluorescent markers in the LV constructs. 75
Table 12: Master mix for ddPCR reaction. 77
Table 13: Program for ddPCR amplification. 77
Table 14: Plasmid ID sequences for DNA discrimination. 79
Table 15: CRISPR/Cas9 sgRNA sequences of target loci. 81
Table 16: Master mix for Q5 PCR amplification of CRISPR/Cas9 donor homology arms. 82
Table 17: PCR programm for Q5 PCR amplification of CRISPR/Cas9 donor homology arms. 82
Table 18: Master mix for Q5 site-directed mutagenesis of lentiviral gag/pol plasmid. 83
Table 19: PCR program for Q5 site-directed mutagenesis. 83
Table 20: Initial and generated plasmids used for barcoding and CRISPR/Cas9 experiments. 84
Table 21: Master mix for barcode library generation. 85
Table 22: Program for barcode generation. 85
Table 23: Master mix for Illumina PCR library generation. 86
Table 24: PCR program for Illumina library PCR. 86

Table of contents

List of abbreviations 1
1 Abstract 3
2 German abstract 4
3 Introduction 5
3.1 The hematopoietic system 5
3.2 Hematopoietic stem cell transplantation 8
3.3 Lentiviral vectors in gene therapy 11
3.4 Genetic barcoding 14
3.5 CRISPR/Cas9 system 18
4 Aims of the thesis 22
5 Results 23
5.1 In vivo reconstitution dynamics 23
5.1.1 Experimental design 23
5.1.2 HSPC cell counts for transplantation 25
5.1.3 Hematopoietic reconstitution and engraftment 26
5.1.4 Tracking of marked HSPC progeny 28
5.1.5 Clonal analysis of HSPC reconstitution 31
5.1.6 Influence of promoter strength on BC32 data interpretation 38
5.2 CRISPR/Cas9 barcoding 41
5.2.1 Experimental design 41
5.2.2 Target efficiency prediction of gRNAs 42
5.2.3 IDLV delivery of CRISPR/Cas9 components 42
5.2.4 SpCas9 target specificity 43
5.2.5 CRISPR/Cas9 barcoding efficiency 45
6 Discussion 48
6.1 In vivo reconstitution dynamics 48
6.1.1 HSPC frequencies in the stem cell grafts 48
6.1.2 Drawbacks in achieving experimental baseline conditions 49
6.1.3 Lentiviral transduction of HSPC populations 51
6.1.4 Technical issues of cell recovery and barcoding procedure 52
6.1.5 Clonal contribution to hematopoietic recovery 53
6.1.6 Concluding remarks 56
6.2 CRISPR/Cas9 barcoding 57
6.2.1 Choice of appropriate loci for targeted barcoding 57
6.2.2 IDLVs as delivery vehicle 58
6.2.3 CRISPR/Cas9 barcoding evaluation 60
6.2.4 Delivery and knock-in in murine HSPCs 61
6.2.5 Future directions 62
7 Materials 64
7.1 Kits 64
7.2 Antibodies 64

7.3 Enzymes 65
7.4 Cell culture media, buffers, and compounds 65
7.4.1 Cell culture media 66
7.4.2 Transient transfection buffer 66
7.4.3 Buffers used for cell preparation and analysis 66
7.4.4 Media used for bacterial culture 67
7.4.5 Buffers used for DNA purification and analysis 67
7.5 Primer 67
7.6 Plasmids 69
7.7 Software and webtools 70
7.8 Instruments 70
7.9 Consumables 71
8 Methods 71
8.1 DNA Isolation 71
8.2 Gel electrophoresis 72
8.3 Cell culture 73
8.4 Animal maintenance, husbandry and care 73
8.5 HSPC isolation 73
8.6 Stem cell transplantation 75
8.7 Blood sampling 75
8.8 Final sampling 75
8.9 Droplet digital PCR 76
8.10 Lentiviral vector production 77
8.11 Viral titer estimation 78
8.12 Cloning 79
8.13 Generation of barcode libraries 84
8.14 Next-generation sequencing and bioinformatics 86
8.15 Validation of SpCas9 editing efficiency 87
8.16 Lentiviral transduction 88
8.17 Transient transfection 88
8.18 CRISPR/Cas9-mediated knock-in validation 88
9 Bibliography 90
10 Appendix I
10.1 Appendix A: Barcode content from NGS analyses I
10.2 Appendix B: Barcode overlap between compartments II
10.3 Appendix C: Mean titers of LV production III
10.4 Appendix C: Hazardous substances III
10.5 Appendix D: Plasmid maps V
11 Acknowledgment VIII
12 Declaration on oath X

List of abbreviations

ADA	Adenosine deaminase
ALD	Adrenoleukodystrophy
Amp	Ampicillin
B-ALL	B cell acute lymphoblastic leukemia
BFP	Blue fluorescent protein
BM	Bone marrow
BMSC	Bone marrow stem cells
Bu	Busulfan
Cas	CRISPR-associated system
CFD	Cutting frequency determination
CGD	Chronic granulomatous disease
CLOUD	Continuum of low-primed undifferentiated HSPCs
CLP	Common lymphoid progenitors
CMP	Common myeloid progenitors
CMRP	Common myeloid repopulating progenitors
CMV	Cytomegalovirus
CRISPR	Clustered regularly interspersed short palindromic repeats
crRNA	CRISPR RNA
Cs	Cesium
Cy	Cyclophosphamide
ddPCR	Droplet digital PCR
DMEM	Dulbecco's Modified Eagle Medium
DPBS	Dulbecco's Phosphate-buffered Saline
ds	Double-stranded
DSB	Double-strand breaks
EBMT	European Society for Blood and Marrow Transplantation
EFS	Elongation factor 1α short
EU	European Union
FACS	Fluorescence-activated cell sorting
FBS	Fetal bovine serum
Flu	Fludarabine
FP	Fluorescent protein
G-CSF	Granulocyte colony-stimulating factor
GFP	Green fluorescent protein
gRNA	Guide RNA
GSH	Genomic safe harbor
GVHD	Graft-versus-host disease
HA	Homology arm
HD	Hamming distance
HDR	Homology-directed repair
HEBS	HEPES-buffered saline
HIV-1	Human immunodeficiency virus type 1
HPRT	Hypoxanthine-guanine phosphoribosyl transferase
HSC	Hematopoietic stem cells
HSCT	Hematopoietic stem cell transplantation
HSPC	Hematopoietic stem and progenitor cell
HSV	Herpes simplex virus
IDLV	Integrase-deficient lentiviral vector
ILV	Integrating lentiviral vector
indels	Insertions or deletions
IT-HSC	Intermediate-term hematopoietic stem cell
IVC	Individually ventilated cages
LAM PCR	Linear amplification-mediated PCR
LB	Lysogeny broth
LHA	Left homology arm
LM PCR	Ligation-mediated PCR
LSK	Lin-Sca-1+cKit+
LT-HSC	Long-term hematopoietic stem cell
LTR	Long terminal repeat
LV	Lentiviral vector
MACS	Magnetic-activated cell sorting
MERP	Megakaryocyte-erythrocyte repopulating progenitors
MkRP	Megakaryocyte repopulating progenitors
MMEJ	Microhomology-mediated end joining
MOI	Multiplicity of infection
MoMLV	Moloney murine leukemia virus
MPP	Multipotent progenitors
NGS	Next-generation sequencing
NHEJ	Non-homologous end-joining
NK	Natural killer

NLS	Nuclear localization signal
PAM	Protospacer adjacent motif
PB	Peripheral blood
PBS	Here: transfer RNA-binding site
PBSC	Peripheral blood stem cells
PCR	Polymerase chain reation
PIC	Pre-integration complex
rAAV6	Recombinant adeno-associated virus serotype 6
RHA	Right homology arm
RNP	Ribonucleoprotein
RRE	Rev response element
RT	Reverse transcriptase
Sca-1	Stem cell antigen-1
SCID	Severe combined immunodeficiency
SDF-1	Stromal cell-derived factor-1
SFFV	Spleen focus-forming virus
sgRNA	Single guide RNA
SIN	Self-inactivating
SLAM	signaling lymphocytic activation molecule
SOC	superoptimal broth with catabolite repression
ss	Single-stranded
ST-HSC	Short-term hematopoietic stem cell
TALEN	Transcription activator-like effector nucleases
TBE	Tris-borate-EDTA
TBI	Total body irradiation
TIDE	Tracking of indels by decomposition
tracrRNA	Trans-activating CRISPR RNA
tSt	Toxic stuffer
UV	Ultraviolet
VCN	Vector copy number
VIS	Viral integration site
VSV-G	Vesicular stomatitis virus G glycoprotein
WAS	Wiskott-Aldrich syndrome
WPRE	Woodchuck hepatitis virus posttranscriptional regulatory element
WT	Wild type
ZFN	Zinc finger nucleases
6-TG	6-thioguanine

1 Abstract

The transplantation of hematopoietic stem cells (HSCs) to treat diseases of the blood-forming system has nowadays become a standard procedure. To this end, cells expressing the surface protein CD34 are harvested from a potential donor and transferred to the patient to replace the defective cells. Interestingly, the $CD34^+$ population is not exclusively composed of HSCs, but represents a heterogeneous pool with additional downstream progenitors. So far, little is known about the underlying cellular processes and the interplay between hematopoietic stem and progenitor cells (HSPC) during reconstitution in the host. This raised the question, whether certain conditions used during the process (HSPC source or pre-conditioning regimen) might influence cellular behavior. Due to the ubiquitous CD34 expression on human progenitor cells, it was reasoned that unnraveling the disparate mechanisms of clonal contribution could possibly lead to reduction of the aplastic phase after pre-conditioning and thus, improvement of the outcome after hematopoietic stem cell transplantation (HSCT).

Using a previously established lentiviral barcoding system (BC32), different HSPC subpopulations, such as HSCs, multipotent progenitors (MPPs), common myeloid progenitors (CMPs), and common lymphoid progenitors (CLPs) were marked and transplanted into myeloablated recipient animals. Cellular output was tracked over time by fluorescent marker expression dependent flow cytometry analysis, droplet digital PCR (ddPCR), and in-depth next-generation sequencing (NGS) analysis of barcode count and frequency. High donor engraftment was observed and the transplanted cells stably contributed to hematopoietic reconstitution, as shown by analysis of mature cells in spleens. Clonal analyses revealed highly polyclonal reconstitution patterns from all four subpopulations, with decreasing numbers of contributing clones in the long-term phase. Taken together, the data emphasize the short-term purpose of MPP contribution and additionally suggest a possible relevance for the long-term blood reconstitution. Moreover, the results provide empirical support that the BC32 system represents a powerful clonal tracking technique to follow four HSPC subpopulations simultaneously after HSCT.

The second prospect of the study was to establish an *in vitro* protocol for neutral CRISPR/Cas9 barcoding into a genomic safe harbor (GSH), in order to avoid undesired effects due to lentiviral integration. For this purpose, integrase-deficient lentiviral vectors (IDLVs) were designed and used for evaluation in a murine cell line. Concomitant with slightly reduced SpCas9 efficiencies by using IDLVs, off-target effects were expectedly low. Comparing different GSH target sites, it was shown that barcode knock-in efficiencies were unaffected by the selected region. In addition, initial low barcoding efficiencies were substantially increased by the use of a generated Cas9-expressing cell line, hinting to the essential need of more elaborated delivery methods of all components for CRISPR/Cas9 barcoding.

In conclusion, the work reported here represents a comprehensive study on neutral marking and subsequent retrieval of barcodes from cells. Altogether, the findings confirm the possibility to combine the CRISPR/Cas9 system with genetic barcoding and pave the way for future applications of neutral *in vivo* cell tracking.

2 German abstract

Die Transplantation von hämatopoetischen Stammzellen (HSCs) zur Behandlung von Erkrankungen des blutbildenden Systems ist heutzutage zum Standardverfahren geworden. Zu diesem Zweck werden Zellen, die das Oberflächenprotein CD34 exprimieren, von einem potenziellen Spender entnommen und dem Patienten verabreicht, um die defekten Zellen zu ersetzen. Interessanterweise besteht die $CD34^+$ Population nicht ausschließlich aus HSCs, sondern stellt eine heterogene Mischung mit den nachgeschalteten Progenitoren dar. Bislang ist wenig über die zugrundeliegenden zellulären Prozesse und die Interaktion zwischen hämatopoetischen Stamm- und Vorläuferzellen (HSPC) während der Rekonstitution im Wirt bekannt. Dies warf die Frage auf, ob bestimmte Bedingungen während des Prozesses (HSPC Quelle oder Konditionierungsart) das zelluläre Verhalten beeinflussen. Aufgrund der ubiquitären CD34-Expression auf den humanen Vorläuferzellen wurde vermutet, dass die Entschlüsselung der klonalen Mechanismen zu einer Verkürzung der aplastischen Phase nach der Konditionierung und damit zu einer Verbesserung des Therapieergebnisses nach hämatopoetischer Stammzelltransplantation (HSCT) führen könnte.

Unter Verwendung eines zuvor etablierten lentiviralen *Barcoding* Systems (BC32) wurden verschiedene HSPC-Subpopulationen, wie HSCs, multipotente Vorläufer (MPPs), myeloische Vorläufer (CMPs) und lymphoide Vorläufer (CLPs) markiert und in myeloablatierte Empfängertiere transplantiert. Zellanalysen erfolgten durch fluoreszenzmarkerbasierte Durchflusszytometrie, *digital droplet* PCR (ddPCR) und umfassendes *Next Generation Sequencing* (NGS) der Barcode Anzahl und Häufigkeit. Es wurde ein hohes Spender *Engraftment* beobachtet und die transplantierten Zellen trugen stabil zur hämatopoetischen Rekonstitution bei, gezeigt anhand der Analyse reifer Zellen der Milz. Klonalitätsanalysen erwiesen hochgradig polyklonale Rekonstitutionsmuster aller vier Subpopulationen, mit abnehmender Anzahl beteiligter Klone in der Langzeitphase. Zusammengefasst heben die Daten den kurzzeitigen Beitrag der MPPs hervor und deuten des Weiteren auf eine mögliche Rolle der Zellen für die langfristige Blutrekonstitution hin. Darüber hinaus belegen die Ergebnisse, dass das BC32 System eine leistungsfähige Technik zur gleichzeitigen Verfolgung der vier HSPC-Subpopulationen nach HSCT darstellt.

Das zweite Ziel der Studie war die Etablierung eines *in vitro* Protokolls für neutrales CRISPR/Cas9 *Barcoding* in einem sogenannten genomischen *Safe Harbor* (GSH), um Nebeneffekte durch die lentivirale Integration zu vermeiden. Zu diesem Zweck wurden Integrase-defiziente lentivirale Vektoren (IDLVs) entworfen und zur Bewertung in einer murinen Zelllinie eingesetzt. Einhergehend mit leicht reduzierten SpCas9 Effizienzen durch die Verwendung von IDLVs waren die *Off-Target* Effekte erwartungsgemäß gering. Beim Vergleich verschiedener GSHs zeigte sich, dass die *Knock-in* Effizienzen unabhängig von der gewählten Region waren. Darüber hinaus wurden die anfänglich niedrigen Effizienzen durch die Verwendung einer generierten Cas9-exprimierenden Zelllinie erheblich gesteigert, was auf den Bedarf zur Weiterentwicklung der Übertragungsmethoden der Komponenten hinweist.

Zusammenfassend stellt die hier präsentierte Arbeit eine umfassende Studie zur neutralen Markierung und Rückgewinnung von Barcodes aus Zellen dar. Insgesamt bestätigen die Ergebnisse die Kombinierbarkeit des CRISPR/Cas9-Systems mit genetischem *Barcoding* und ebnen den Weg für zukünftige Anwendungen neutraler *in vivo* Zellverfolgung.

3 Introduction

3.1 The hematopoietic system

Blood is a dynamic and highly regenerative tissue. Daily, more than 1×10^{12} cells are replenished, thereby creating a finely balanced system able to react to environmental challenges, such as infections[1]. Most of our current knowledge on blood formation and homeostasis was derived from functional assays or transplantation studies in murine models.

A landmark in unraveling the mechanisms of blood formation, i.e. hematopoiesis, was work of Till and McCulloch in the early sixties. By investigating splenic colonies in irradiated mice into which single-cell progenitors had been transplanted, they showed that the colonies are clones formed through the differentiation of transplanted cells along three lines (erythrocytic, granulocytic, and megakaryocytic)[2].

In the next decades, a multitude of studies provided additional information on the routes of differentiation, resulting in the so-called 'common paradigm of hematopoiesis', according to which hematopoietic stem and progenitor cells (HSPCs) are organized in a hierarchical tree-like diagram with the hematopoietic stem cell (HSC) at the apex[3-6] (Figure 1). Residing in the bone marrow (BM), HSCs are supposed to actively account for the reconstitution of the blood system throughout the entire lifespan of an organism, constantly giving rise to mature hematopoietic cells. They are defined by two essential properties: multipotency, which allows them to differentiate into all downstream lineage branches and unlimited, robust self-renewal capacity, which ensures a steady base of the system[7-11]. Over the years, subgroups of the heterogeneous pool of HSCs have been defined based on their phenotypic characteristics and repopulating capabilities. Cells providing long-term multi-lineage reconstitution (>12 weeks) have been termed long-term HSCs (LT-HSCs)[1,12], and have been shown to divide once per 30–50 days[13,14]. The other two subgroups are intermediate-term (IT-) and short-term (ST-) HCSs. ST-HSCs are closely related to the multipotent progenitors (MPPs) and maintain transient engraftment capability[4,12,15-17]. In contrast, IT-HSCs persist for six to eight months post transplantation before becoming extinct[12]. Upon differentiation, cells lose their stem cell features and divide towards a lymphoid-, myeloid- or megakaryocyte/erythroid-restricted phenotype[18,19]. The first lineage commitment showing up as a branching point in the hierarchy is the fate decision of MPPs to differentiate towards a myeloid- or lymphoid-primed phenotype. The resulting cells are called CMPs (common myeloid progenitors) and CLPs (common lymphoid progenitors), respectively. These 'committed progenitors' then give rise to the corresponding mature lineages of the lymphoid compartment, i.e., B lymphocytes, T lymphocytes, and natural killer (NK) cells, as well as the myeloid compartment, i.e., granulocytes, monocytes, megakaryocytes, and erythrocytes.

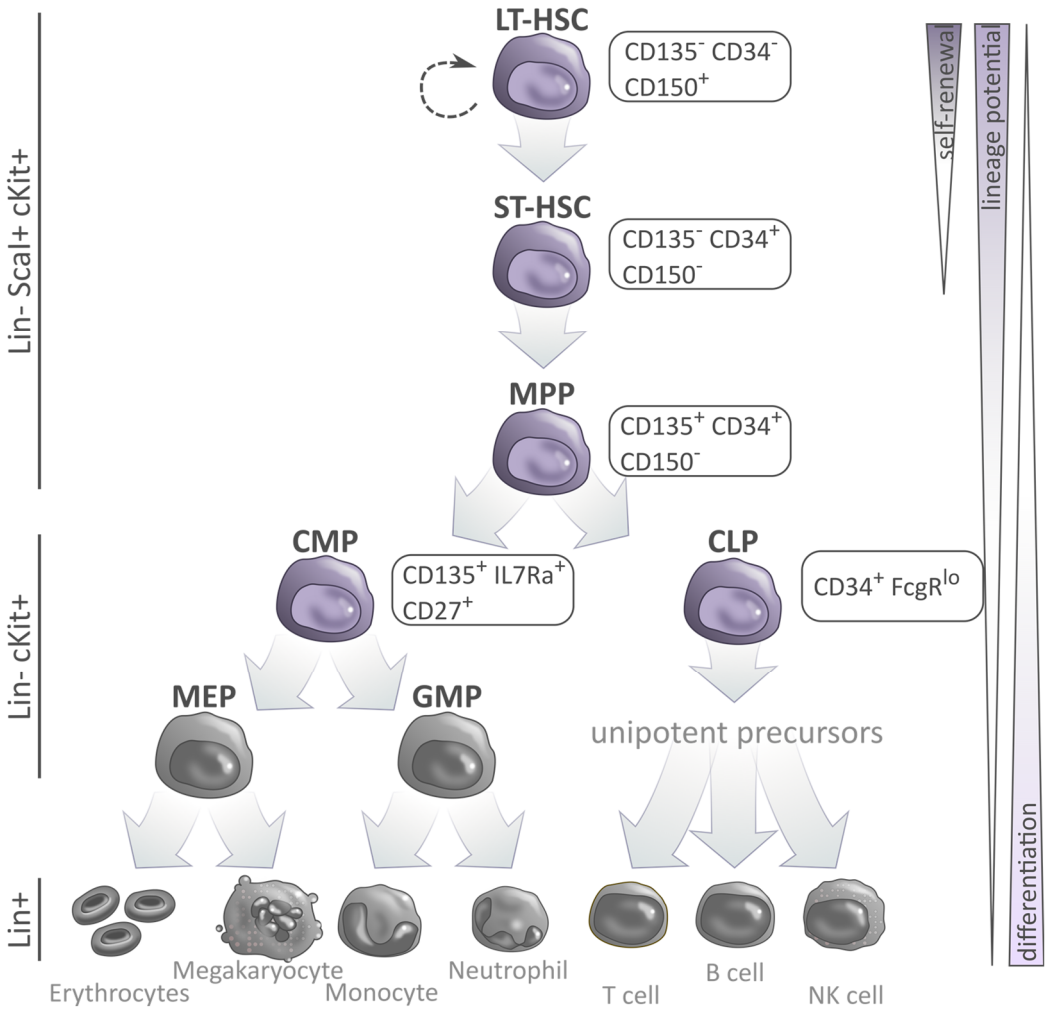


Figure 1: Common model of discrete hematopoietic hierarchy. The hematopoietic differentiation is currently viewed as a hierarchical tree with the LT-HSC at the apex harboring unlimited self-renewing potential and multilineage differentiation ability. Differentiation of HSCs via MPPs to a more lineage-restricted phenotype results in branching to progenitor populations with a myeloid or lymphoid bias (CMPs and CLPs). En route to mature blood cells (erythrocytes, megakaryocytes, monocytes, granulocytes, T- or B lymphocytes, and NK cells) differentiation surpasses further bi- and unipotent progenitor states (i.e. MEP or GMP), all defined by a specific marker profile. Common markers (specific for whole groups) are displayed in the far left (Lin-Sca-1+cKit+, Lin-cKit+ or Lin+ cells) and characteristic markers defining the subpopulations are shown in the boxes next to the cell populations (murine system). Increasing or decreasing abilities of the populations for self-renewal, lineage potential or differentiation is depicted on the right. Cell populations relevant for the current work are highlighted in purple.
CLP: common lymphoid progenitor, CMP: common myeloid progenitor, GMP: granulocyte-macrophage progenitor, LT-HSC: long-term hematopoietic stem cell, MEP: megakaryocyte-erythroid progenitor, MPP: multipotent progenitor, ST-HSC: short-term hematopoietic stem cell

Both the committed progenitors as well as the cell types forming their derived lineages can be isolated based on their distinctive surface marker profiles using single-cell sorting, an advantage that was extensively utilized in the studies that led to the construction of the classical model of hematopoiesis (Figure 1). The Weissman group was the first to produce an enriched HSC population back in 1988[20]. Since then, a myriad of surface marker profiles has been identified and revised, revolutionizing the dissection of HSPC populations in increasing detail[9,21]. The first parameter based on which immature HSPC populations are isolated is the absence (Lin^-) of common lineage markers (B220, CD4, CD8, Gr-1, Mac-1 and Ter-119)

combined with the presence of stem cell antigen-1 (Sca-1) and cKit (CD117). Cells displaying these properties are defined as the $Lin^{-}Sca\text{-}1^{+}cKit^{+}$ (LSK) population of the BM[16,20,22-24]. Since HSCs are among the LSK population, which is enriched in HSCs (approximately 3—10 % of the LSK cells are true HSCs)[21,23,25,26]. Several other strategies have been established to enhance the enrichment of HSCs from BM, some of which do not include the aforementioned antigens. Recently established protocols that include three cell surface glycoproteins belonging to the signaling lymphocytic activation molecule (SLAM) family, namely, CD150, CD244, and CD48 as additional markers give promising results in the enrichment of HSC and progenitor populations. Notably, one such protocol was successful in isolating a highly HSC-enriched (40–50 %) fraction from whole BM[27]. Moreover, CD150 was shown to enable distinguishing between different subsets of HSCs, indicating an early myeloid- or lymphoid-priming[28]. Other utilized markers include endothelial protein C receptor (EPCR or CD201)[29] or distinct patterns of CD49f expression on MPPs[12]. However, it should be mentioned that most of the aforementioned markers have been only studied in the context of the murine hematopoietic system, while far fewer studies have been conducted on human HSPC subpopulations and their surface marker profiles.

Besides assessing the clonal processes that take place during embryonic development, another fundamental drive in hematological research is to investigate cell determination and developmental routes in the adult organism. Thus, it may add new knowledge to the study of developmental defects and also improve the ability to restore defective tissue through clinical applications like hematopoietic stem cell transplantation (HSCT). The hematopoietic system offers the indispensable prerequisites for in-depth analyses, in particular its lack of interconnections and the ability of complete renewal following eradication and transplantation, as these properties facilitate peripheral blood (PB) isolation and separation of distinct cell populations. Furthermore, HSPCs possess the ability to form colonies in a semi-solid medium under specific conditions and are thereby eligible for *in vitro* clonality assays. The most common functional test for HSCs *in vivo* is HSCT between congenic mouse strains, which finally allows for quantitative, functional HSC readouts, by assessing distinct expression of CD45 isoforms (CD45.1 and CD45.2) on cellular surfaces[30-33]. Interestingly, blood reconstitution can be initiated by transplantation of limiting dilutions of cells, as shown by a study aimed to uncover HSC frequency in a cell pool, [34] or even by a single transplanted HSC[15]. However, considering that the definition of a true HSC includes long-term multi-lineage reconstitution, most approaches offer only retrospective validations of the results and serial transplantations are mandatory to validate self-renewal capacity[16]. Up until recently, xenograft rejection post transplantation of human BM cells into mice hampered the use of animal models to assess human blood reconstitution. However, new humanized mouse models with incompetent adaptive immunity emerged to overcome this limitation[35] (reviewed in [36]).

While evaluating the underlying mechanisms of hematopoietic reconstitution in a transplantation setting, the disparity of native and perturbed hematopoiesis needs to be taken into account. Moreover, most current knowledge is based on studies utilizing HSC isolation

and genetic *in vitro* manipulation, which may introduce alterations in cellular physiology. Thus, efforts have been made to introduce non-invasive techniques. Most of them rely on the Cre-loxP site-specific recombination system, in which tamoxifen inducible Cre recombinase is used to excise specific loci, thereby creating unique patterns, enabling *in vivo* fate mapping of cells in mice. Compared to the transplantation approach, this system shows a strong myeloid bias in HSC contribution, a rather diverse (pool) than individual cell contribution, and a slow division rate of 110 days (1 %/ day) (reviewed in [37]). As in the current thesis, the potential of cells under stress brought about by adoptive cell transfer was investigated, it must be mentioned that the findings do not represent physiological *in vivo* processes.

Recent findings have challenged the classical view of hematopoietic differentiation described above. New paradigms were created suggesting a rather continuous than discrete hematopoiesis in which a continuum of low-primed undifferentiated HSPCs ("CLOUD") bypass transition through multi- or bipotent stages[38]. The HSC population itself was divided into myeloid-biased, balanced, and or lymphoid-biased[39-42] cells. Furthermore, it was proposed, that subgroups of megakaryocyte repopulating progenitors (MkRPs), megakaryocyte-erythrocyte repopulating progenitors (MERPs), and common myeloid repopulating progenitors (CMRPs)[43-45] coexist in the HSC population. Finally, understanding of MPPs has also evolved in recent years. Four subpopulations (MPP1-4) have been described in accordance with their lineage restrictions, immunophenotype, or BM abundance[46,47]. These advances reveal the heterogeneity of HSPC populations that had been long considered well-defined and stress the need for revision of old models. This can be achieved by studying the HSPC populations in close detail, for example by investigating the replenishment of blood after HSCT.

3.2 Hematopoietic stem cell transplantation

The transplantation of HSCs to restore the blood and immune systems is a unique clinical practice that was developed after the Second World War. One of the major factors fueling the research interest in this area was the need for a way to restore the damage radiation inflicts on these systems, an effect that became evident by the symptoms faced by the survivors of the atom bombs dropped near the end of the war. Two independent groups focused on mouse models of radiation-induced BM aplasia and were eventually able to demonstrate that not only BM could be restored by syngeneic marrow graft[48], but also leukemia relapse could be prevented[49]. The first HSCT in humans was performed by E. Donall Thomas in 1957, but since basic knowledge on histocompatibility was lacking and no matching of recipient and donors was performed, the six treated patients died within a few months[50]. Soon after the discovery of the HLA system in humans a series of human transplantation procedures were initiated by Thomas and others[51], with increasing success rates by the years in acute leukemia. Due to his pioneer work and his preserving efforts, he received the Nobel Prize in 1990.

Nowadays, HSCT is used as the standard of care for a multitude of congenital or acquired hematologic malignancies[52], while recently indications like autoimmune diseases and

metabolic disorders were also included[53,54]. It is classified as allogeneic or autologous, depending on the donor stem cell source. Allogenic transplantation, in which cells from a voluntary donor (related or unrelated) are used, is the standard treatment for leukemia, some lymphomas, and myelodysplastic syndromes, multiple myeloma, and homozygous sickle cell disease. On the other hand, autologous transplantation using HSCs collected before treatment is used to rescue the immune system and prevent prolonged BM aplasia in patients receiving high-dose chemotherapy. This is of critical importance in cases where myeloablative chemotherapy is inevitable, such as in lymphomas, leukemia, germline tumors, and soft-tissue sarcomas. As mentioned above, the initial application of HSC transplantation back in the '50s was unsuccessful due to a lack of histocompatibility between the donors and recipients. Today, even in cases where donor and recipient are partially compatible, the latter undergoes a preceding treatment (pre-conditioning) with the aim of (i) immunosuppression, (ii) myeloablation and (iii) eradication of residual tumor cells (in cancer patients). Different regimens (chemo-, radio-, or combined chemoradiotherapy) for pre-conditioning may be chosen, based on the level of myeloablation, intensity, and toxicity indicated for the patient, which in turn depend on the diagnosis, disease status, donor availability, as well as patient characteristics such as age, overall fitness, and comorbidities (reviewed in [55]). In clinical practice, the three most common myeloablative regimens for adult patients are combinations of cyclophosphamide/ total body irradiation (CyTBI), busulfan/ cyclophosphamide (BuCy) and fludarabine/ busulfan (FluBu). Adverse reactions concomitant with the actual scope of these drugs and treatment regimens, i.e. the eradication of cells, have been reported for all of them. To mention only a few, long-term side effects of TBI are infertility and cataract formation[56], hyperthyroidism, and thyroiditis[57]. On the other hand, busulfan is an alkylating agent with unpredictable bioavailability and tolerability, which has been linked to hepatic veno-occlusive disease[58]. In a retrospective analysis, it was shown that, while FluBu showed similar relapse rates and overall survival to CyTBI, non-relapse mortality and acute graft-versus-host disease (GVHD) were significantly impaired following CyTBI treatment[58]. A murine model studying acute GVHD in mice also highlighted the differences of the pre-conditioning regimens: BuCy resulted in a more severe form of aGVHD, while FluBu was associated to massive BM damage[59]. However, comprehensive comparative trials between the regimens regarding reconstitution rates and the impact of changes in BM niche on repopulating cell populations are still missing.

As mentioned above, human and murine HSCs are well characterized by a subset of surface markers, which offer an effective means for sorting and analysis (for murine markers see Figure 1). In the clinical setting, the number of $CD34^+$ cells in a graft is indicative of the HSC content and is thus considered the best predictive parameter of transplant quality and the kinetics of platelet and neutrophil engraftment[60,61]. According to the European Society for Blood and Marrow Transplantation (EBMT), transplants are recommended to contain a minimal number of 2-4 × 10^6 $CD34^+$ cells/kg body weight[62-66]. The cells to be transplanted were commonly harvested by repeated BM aspirations in the pelvic crest, yielding at least 3 × 10^8 nucleated cells/kg[67]. Nowadays this method is only applied in a minority of cases, in particular when harvesting a sufficient amount of HSCs from pediatric donors is a limiting factor, or

specific reasons that necessitate a more careful evaluation of the risk for GVHD[68]. In the early ’90s it was found that peripheral blood stem cells (PBSCs) mobilized by cytokines could be used in an equal manner as bone marrow stem cells (BMSCs)[69-71]. Today, PBSC transplants account for 70—95 % of all HSCTs[72,73]. PBSCs are preferred because they engraft faster compared to BMSCs, possibly improving patient outcomes, and their harvesting is undeniably connected to a less stressful procedure for the donors[74]. First, granulocyte colony-stimulating factor (G-CSF) is injected into the donor to induce myeloid hyperplasia and excessive release of $CD34^+$ cells to the periphery. Although currently the underlying mechanism is not completely resolved, the dominant view considers the process to be mediated by the downregulation of the chemokine stromal cell-derived factor-1 (SDF-1 or CXCL12) and the proteolytic cleavage of adhesion molecules, such as cKit ligand, and VCAM, by the metalloproteinases neutrophil elastase and cathepsin G, which contributes to HSC mobilization into the circulation[75,76]. Interestingly, increased HSC counts can be achieved by the use of plerixafor (AMD3100), a small molecule CXCR4 antagonist that inhibits the CXCR4-SDF1 axis and interaction. In preclinical mouse models a threefold increase in LSK cells was achieved by the combinatory use of G-CSF and plerixafor compared to G-CSF alone[15,77,78]. After the release of PBSCs has been induced, they are collected by leukapheresis.

Irrespective of the source of the transplants, their exact composition is unknown, since the only available clinical parameter – CD34 surface expression – does not uniquely describe *bona fide* LT-HSC. The question for possible influences of different HSPC subpopulations on hematopoietic reconstitution, which are undoubtfully present in the heterogeneous cell pool remains to be answered. So far, in-detail characterization of HSPC subsets in grafts is not routinely performed but some reports have already begun to address this issue[66,79].

Even though HSCT is used in daily practice and characterized by promising curative potential for life-threatening diseases, it does have drawbacks. Despite the adoption of precautionary measures such as histocompatibility checks and pre-conditioning, transplant-related mortality (TRM) remains a major concern. A main determinant of HSCT outcomes are processes that take place in the short time period after the intervention preceding donor cell engraftment and the actual transplantation. This critical phase of BM aplasia, anemia, and thrombocytopenia bears the responsibility for many post-transplant complications. With GVHD and infections being the most common causes of morbidity and mortality[80], approximately 25 % of patients die within one year after HSCT[81]. Hence, a better understanding of cellular processes taking place during the period of aplasia might eventually reduce the risk of TRM.

One of the recent developments in autologous HSCT is the use of *ex vivo* gene therapy for adding or editing defective genes in HSCs, which are then reintroduced into the host organism. This allows the use of corrected HSCs for the therapy of congenital diseases, eliminating major immunological complications and the need for a matched donor. By their accessibility, multipotency and life-long blood supply HSCs represent an optimal target for gene therapy, which, once corrected, lead to permanent gene function in all downstream lineages. In order to incorporate the lacking genes into the genome of the host cells, integrating lentiviral

vectors (LVs) are nowadays the gene transfer vehicles of choice in gene therapy and ongoing clinical trials.

3.3 Lentiviral vectors in gene therapy

Starting to become the focus of scientific attention as a means to achieve permanent genetic modification of host cells and their progeny in the early eighties[82], nowadays LVs are one of the most widely used gene-transfer vehicles in research and gene therapy. As indicated by their name, they derive from lentiviruses, a genus of viruses belonging to the family of complex *Retroviridae*. Sharing the same features regarding structure and genome organization, they are characterized by the presence of an envelope and two copies of a single-stranded (ss) (+)-sense RNA that serves as their genome. The RNA genome, together with viral enzymes, is enclosed inside a protein structure called nucleocapsid. In lentiviruses, the nucleocapsid is surrounded by an outer protein layer comprised of the matrix protein, which is in turn encompassed by the envelope glycoprotein (Env)-studded, host cell membrane-derived lipidic membrane[83] (Figure 2).

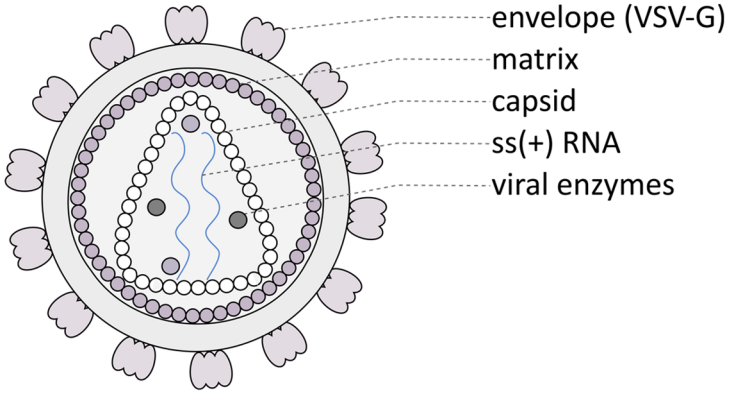


Figure 2: Schematic representation of the lentiviral particle structure. The outer structure of the lentiviral particles forms the envelope, which can be adapted and be utilized for pseudotyping in order to infect a certain range of cells (here: VSV-G). Viral ss (+) RNA genome is located inside the nucleocapsid, which is surrounded by a proteic capsid, an inner matrix layer, and a double lipidic membrane on the outside. For gene transfer purposes, the naturally occurring RNA genome can be exchanged with a transgene of interest. Additionally, viral enzymes (RT, Proteinase, and Integrase) are present in the nucleocapsid.
RT: reverse transcriptase, VSV-G: vesicular stomatitis virus G glycoprotein

After the virus enters the cell, the RNA genome is transcribed into a DNA intermediate by the viral enzyme reverse transcriptase (RT), which is then integrated into the genome of the host organism in a semi-random fashion[84,85] with the help a pre-integration complex (PIC). The PIC is forwarded to the nucleus of the cell in an ATP-dependent manner. Once integrated, the provirus is transcribed by the cellular transcription machinery (RNA Polymerase II), and viral mRNA is released to the cytoplasm. There it guides the production of viral components, which are assembled into new viral particles in lipid rafts close to the cell membrane. Finally, the

new viral particles exit the cell by budding from the cell membrane, completing this unique replication cycle.

The genome of human immunodeficiency virus type 1 (HIV-1), a representative lentivirus, can be separated into coding sequences for regulatory proteins (Tat and Rev), accessory proteins (Vif, Nef, Vpr and Vpu) and structural proteins (Gag and Env), and the RT polymerase. Among them, the accessory proteins and the regulatory protein Tat are not necessary for the viral life cycle. Thus, the corresponding coding sequences can be eliminated or exchanged for the purpose of utilizing lentiviruses as a tool of stable gene transfer into dividing and non-dividing cells. To this end, researchers started tampering with the genetic architecture of LVs in the mid '90s[86]. The aim was to make use of the gene transfer itself while minimizing most of the viral features that are non-essential to this process. For minimizing safety concerns, researchers adopted the strategy of dividing viral genes into several independent expression cassettes or plasmids *in trans* to allow for particle formation but avoid replication, i.e., thereby creating replication-defective LVs. The first generation vector systems comprised: (i) a packaging cassette with a substituted 5' long terminal repeat (LTR) for a cytomegalovirus (CMV) promoter and the addition of a polyA site in exchange for the 3' LTR, (ii) an envelope cassette coding for a viral envelope protein (e.g., vesicular stomatitis virus G glycoprotein, VSV-G), lacking packaging signal (psi) and transfer RNA-binding site (PBS), and (iii) a transgene cassette, harboring 5' LTR, psi, rev response element (RRE), and the transgene under the control of an endogenous promoter[86]. Interchanging viral information for naturally occurring lentiviral Env proteins – a process called pseudotyping – enables to switch the tropism of the viral vector to the organism the glycoproteins originate from. This method offers a wide range of target cell-specific adaptations for LV application. Refinement led to a second generation of LVs, in which the main split packaging of viral genes was maintained, but all non-essential accessory genes were discarded, without any negative influence on viral maturation or infection rates[87-89]. Further optimization was conducted by completely eliminating the *tat* gene and moving *rev* to a fourth expression cassette, a step that reduced the risk of replication-competent LVs arising from homologous recombination events[90].

However, due to the possible risk of activation of neighboring host genes (including oncogenes) by transcriptional interference of the integrated virus (see also Figure 3), the need for a higher safety level led to the development of self-inactivating (SIN) LVs. These vectors contain a partial deletion of the U3 region of the 3'LTR. During reverse transcription, this deletion is transferred to the 5'LTR, resulting in the provirus being devoid of promoter and enhancer elements in the LTRs (Figure 3B). In SIN LVs, transgene expression is Tat-independent and driven by a strong heterologous promoter (specifically, CMV)[91-93] that replaced the 5'LTR. The use of SIN LVs decreased the potential of insertional mutagenesis to a minimum[94].

The first successful application of a retroviral vector in genetic engineering was the use of the Moloney murine leukemia virus (MoMLV) to incorporate a copy of the herpes simplex virus (HSV) thymidine kinase into HSPCs[95]. Shortly after, the first human gene transfer experiment was performed in melanoma patients by Rosenberg et al., proving the applicability, in terms

of both effectiveness and safety, of virally altered cells in humans[96]. This was followed by clinical trials using this technology to treat primary immunodeficiencies, with the first target being T cells introduced with a functional copy of the gene coding for adenosine deaminase (ADA) of ADA-severe combined immunodeficiency (SCID) patients[97]. To ensure a life-long supply of genetically modified cells, the procedure can be performed *ex vivo* in HSPCs (in combination with a non-myeloablative regimen), which give rise to ADA-expressing progeny[98-100]. After promising results of these trials, the therapeutic use of gamma retroviral vectors was also applied to X-linked severe combined immunodeficiency (SCID-X1)[101,102], Wiskott-Aldrich syndrome (WAS)[103,104] and X-linked chronic granulomatous disease (CGD)[105,106]. To date, more than 3000 gene therapy clinical trials including retroviruses have been approved, with increasing numbers each year (according to The Journal of Gene Medicine Gene Therapy Clinical Trials Worldwide website at http://www.wiley.co.uk/genmed/clinical, accessed on 11th November 2020). About 304 (9.9 %) of these made use of LVs. First used, to successfully treat X-linked adrenoleukodystrophy (ALD)[107], thereby replacing defective microglia without the occurrence of any adverse effects, LVs are nowadays the classical gene-delivery vehicle in HSCs[108]. Gene therapy products from retroviral vectors, which are approved in the European Union (EU) are Strimvelis (ADA-SCID)[100], Kymriah/ tisagenlecleucel (B cell acute lymphoblastic leukemia (B-ALL) and B-cell lymphoma)[109], Yescarta/ axicabtagene ciloleucel (B-cell lymphoma)[110], and Zynteglo (beta thalassemia).

Despite the abovementioned success of inserting genes in HSCs or effector cells for treating severe monogenetic diseases, adverse effects such as leukemogenesis and clonal dominance were reported in several cases. Specifically, lymphoproliferative or myelodysplastic disorders occurred in the X-SCID (5 out of 23 patients developed leukaemia)[111,112], WAS (7 out of 10 patients developed leukaemia)[103,104], and X-CGD clinical trials (all three patients developed myelodysplastic disorders)[113,114]. Clonal dominance was also reported despite clinical improvement in a clinical trial of beta thalassemia[115]. The underlying mechanism was found to be insertional mutagenesis, i.e., mutations caused by the integration of the provirus inside or in close proximity to a host gene influencing the transcription machinery of the host cells leading to aberrant expression of the production of aberrant transcripts[84,116]. Elements from the viral vector which are prone to interfere with the surrounding genome are promoter elements, which can then drive gene expression from the 5'LTR, splicing and polyadenylation signals, leading to a premature stop signal, aberrant fusion products, or deletion of 3'UTR-situated mRNA stabilizing motifs[84,116] (Figure 3A).

Fortunately, since the utilization of viral vectors with SIN architecture, no events of clonal dominance or severe accompanying disorders were reported. A further step towards safer clinical application was the development of integrase-deficient lentiviral vectors (IDLVs), as their use greatly reduces the risk of mutagenesis. IDLVs form episomes instead of integrating into the host genome. Transgene expression is transient, as episomes are not passed to the cellular progeny. Clinical applications that may benefit from IDLVs are (i) *in vitro* and *in vivo* reporter gene delivery, (ii) antigen expression for immunotherapy, (iii) transient gene

expression to accommodate genetic engineering tools (like clustered regularly interspaced short palindromic repeats (CRISPR)/Cas9), and (iv) gene replacement therapy.

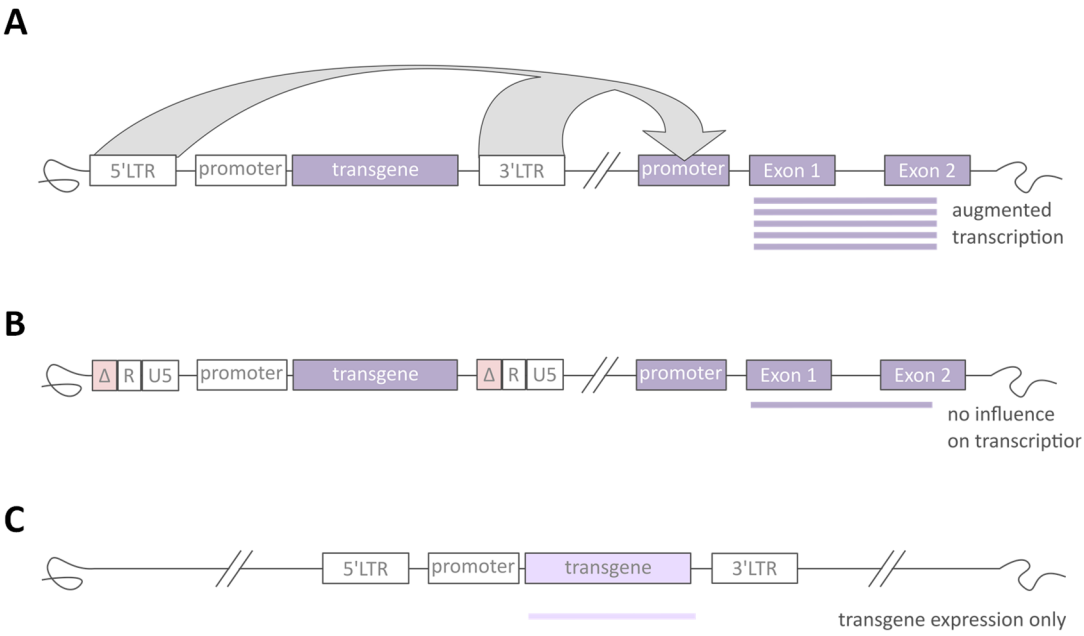


Figure 3: Insertional mutagenesis and safety strategies. A) After viral integration into the host genome, the provirus interacts with endogenous cellular elements in proximity to the integration site, leading to upregulation of neighboring genes through enhancer activation. Enhancer elements present in viral LTRs influence endogenous promoter activity and thereby elevate mRNA transcription. B) Viral vectors with SIN architecture avoid the risk of enhancing surrounding gene transcription. A deletion of enhancer elements in the U3 region of viral vectors (here: Δ) is transferred to the 5'LTR of proviral DNA during reverse transcription. D) Ideally, viral integration is targeted to a genetically inert site (GSH), e.g., a non-coding or otherwise favorable region, thus no changes in endogenous host gene expression are caused. Only the transgene itself is expressed.
GSH: genomic safe harbor, LTR: long terminal repeat, SIN: self-inactivating

LVs do not only represent a powerful tool in gene therapy but are also widely used for studying developmental processes and performing clonal analyses in cell populations. Once integrated, they serve as inheritable marks to track cellular progeny by unique insertion sites, by the expression of a fluorescent marker or the introduction of a sequence-specific label. The latter is termed a genetic barcode.

3.4 Genetic barcoding

The concept of cellular development analysis at the single-cell resolution level is a continuous vision in research. The discovery of retroviral vectors and their capacity to stably integrate semi-randomly into the genome of host cells gave researchers a new tool, as the inserted sequence can be used to track cellular progeny. This simple yet genius concept enables clonal tracking of individual cell fate decisions over time, allowing researchers to address open questions about single-cell commitment to ontogeny and tissue constitution. Such questions are particularly relevant in the study of a tissue as dynamic as the hematopoietic system, which is not only constantly renewed under steady-state conditions but also has a remarkable

reconstitution ability that allows to recuperate after total eradication even from a single-cell, a property on which HSCT therapies are based[15,117].

The first trials to trace the lineage fate of HSCs were performed in 1986 when Lemischka et al. used retrovirus-mediated gene transfer to mark HSCs *in vitro* and then tracked the fate of these cells after their transplantation into lethally irradiated recipients[118]. Subsequent studies extrapolated the results of Southern blot analyses[119-123] or fluorophore expression[124-126] to identify progenitor fates. Viral integration sites (VIS), which serve as unique genetic marks due to their semi-random integration patterns, can be determined via ligation-mediated (LM) or linear amplification-mediated (LAM) polymerase chain reaction (PCR)[127-130] to name a few. However, both methods rely on the fragmentation of genomic DNA and ligation of a linker DNA cassette for PCR-based amplification of the unknown sequence between viral LTR and linker[131]. Hence, the VIS retrieval is based on the choice of the restriction enzyme, the amplicon length, and amplification efficiency, thereby biasing the output[132,133].

Further development came about in the mid-‘90s when Golden et al. introduced a new technique that uniquely marked cells with retroviral DNA libraries[134]. Later on, these genetic labels were named genetic barcodes, as a reference to the well-known, unique binary barcode system used to identify goods. Schumacher[135] and Bystrykh[136] evolved the idea of genetic barcodes – now comprising high numbers of non-coding DNA stretches of known length – by coupling it to high-throughput technologies such as next-generation sequencing (NGS). A comparative analysis recently proved that genetic barcoding is at least 5-fold more efficient than VIS analysis due to sequencing depth[137].

It should be noted that, besides the hematopoietic system, the study of various other fields can be facilitated by unraveling cell lineages. A body of studies employing barcoding tools assessed physiological phenomena beyond HSCs, such as T-cell behavior[135,138-140] and mammary tissue development[39,141]. Other fields that benefitted from this technology were assessment of metastatic behavior[142-144], recurrence of tumor outgrowth after surgical resection[145], phenotypic plasticity[146,147], and analysis of drug response mechanisms[143,144,147-152], for each of which appropriate BC tracking systems were developed. Focusing on aberrant development and pathogenic mechanisms, barcoding was used to study the clonal dynamics of transformation and chemotherapy resistance formation[148,153]. More recently developed barcoding approaches unraveled mechanisms in intratumoral heterogeneity[154].

Based on the work of the Schumacher[135] and Bystrykh[136] groups our team recently developed and refined their own barcode system introducing slight adaptations aimed at overcoming limitations in quantification and barcode retrieval[155,156]. It was termed BC32 due to the characteristic stretch of 32 variable base positions (N) in a defined order of $N_3(M_3N_2)_{16}$, where M_3 is for fixed base triplets of known sequence. The positions of variable bases are filled randomly by one of the four bases (A, C, T or G) during synthesis. Assuming equal probabilities for each base during the process of oligonucleotide generation, the calculated theoretical complexity of the library is 1.8×10^{19} barcodes. Moreover, combinatory systems utilizing both genetic and optical marking (RGB marking[157], see also [158-163]) were reasoned to be of great interest due to its broader scope of application. Thus, the here used BC32 system harbors two

main features: first, a certain assignment of a specific barcode backbone to a fluorescent reporter protein (FP) encoded by the same LV construct and, second, the availability of four different assigned barcode-FP pairs (see Figure 4). The latter is of special interest as it follows for simultaneous analyses of up to four distinct cell populations, and thereby offers the opportunity to examine clonal and spatial interplay between cell populations, for example HSPCs.

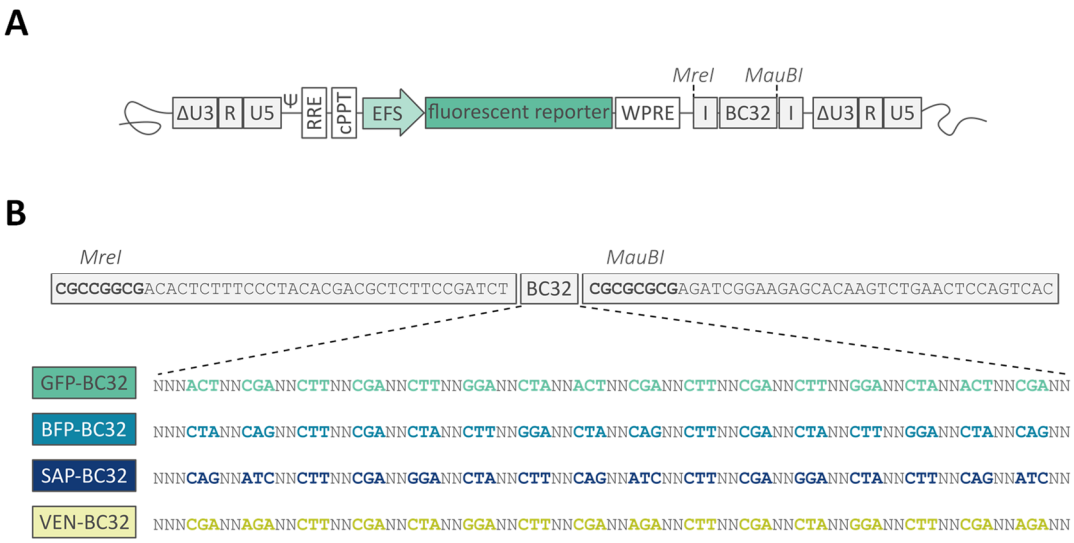


Figure 4: BC32 system for cellular barcoding and tracking. (A) Viral BC32 construct after integration into the genome (provirus). A fluorescent marker is controlled by an EFS promoter and expression is augmented by the presence of WPRE. The barcode is located downstream of the fluorescent marker, flanked by two restriction sites (MreI and MauBI) and truncated Illumina adapter sequences (I) for subsequent NGS analysis. Viral regulatory elements, such as the packaging signal psi (Ψ), RRE, and cPPT are depicted in white color, whereas viral LTRs (ΔU3-R-U5) are shown in light grey. (B) To simplify cloning and NGS procedures, the BC32 is cloned into the construct in conjunction with a 5'-adjacent Illumina adapter between two restriction enzyme recognition sequences. The 3' Illumina adapter sequence is present in the viral construct downstream of the 3' MauBI recognition sequence. Four BC32 backbones (GFP, eBFP, T-Sapphire, and Venus) characterized by stable base positions of known sequence (colored) allow for simultaneous analyses of up to four distinct subpopulations. N positions depict variable positions for random nucleotide insertion during barcode synthesis.
cPPT: central polypurine tract, eBFP: blue fluorescent protein, EFS: elongation factor 1a short , GFP: green fluorescent protein, I: Illumina sequences, LTR: long terminal repeat, NGS: next-generation sequencing, RRE: rev response element, WPRE: woodchuck hepatitis virus posttranscriptional regulatory element

Besides the BC32 system, several alternative lineage-tracing strategies have been suggested. Native hematopoiesis has been intensively studied using an inducible transposase system ('sleeping beauty') and subsequent analysis of transposon insertion sites in HSPCs and mature cells[164,165], a technique similar to the analysis of VIS. In 2007, a promising murine fluorescent reporter system based on Cre-mediated inversion ('Brainbow mouse') opened up new possibilities in the optical tracing of lineage relationships[124,166]. A similar technique was developed in 2017 when Pei and colleagues introduced an endogenous, genetic *in situ* barcoding system ('Polylox barcoding') exploiting Cre recombinase-driven random DNA recombination[167,168]. Specifically, they generated an inducible marking system with >1 × 10^5 retrievable barcodes and showed the applicability by tracking HSC clone fates *in vivo*. Several

other groups focused on genetic barcoding of cells via CRISPR/Cas9 modifications, like the incorporation of a genetic region (barcode) consisting of consecutive CRISPR/Cas9 target sites, leading to SpCas9-mediated editing variants in this defined locus. Others focused on coupling the SpCas9 enzyme to a deaminase/ polymerase variant, thereby inducing genetic variants in a defined locus[169-171].

Comprehensive barcode analysis after sample and data collection relies nowadays mostly on counting sequencing reads obtained from NGS runs. Different from microarrays, the output of deep sequencing offers an exact measurement of relative barcode frequency. To address the lack of standardized protocols for conscious and reliable data handling, our group published in 2020 a user-friendly interface for barcode data normalization and analysis based on a systematic code written in R[172]. As most pipelines used so far, they discuss the two main factors of barcode data handling, i.e., filtering of PCR noise and pooling of related or descendent barcodes, respectively. While the latter is mostly a question of useful data visualization for the respective purpose, the former requires careful evaluation of normalization. If the data cannot be matched to a reference library or correlated between duplicates, the normalization procedure relies on the filtering of low-frequency reads to discard sequencing errors or matching and allocating them to a high frequency read with the lowest sequence similarity. The latter is described by the hamming distance (HD) of the barcodes, which can be estimated by the number of at which the matching nucleotides differ. Accordingly, a high HD corresponds to low sequence similarity, and vice versa. While it was suggested, based on common (binominal) statistics, that at least five observations need to be present for significantly relevant evidence[173], it still remains unclear if cells present in low frequencies should be disregarded from analyses[139]. To avoid false-positive results from PCR errors, the use of unique molecular identifiers (UMIs) has been suggested[174]. These are short sequence tags added to the PCR template right before amplification to allocate amplicons and verify the similarity of the output. Their use can reveal whether a nucleotide mismatch between two sequence was introduced during the course of PCR and DNA amplification or if two *bona fide* distinct barcodes are present. Additionally, data should be filtered for distribution bias in barcode abundance to adjust for the importance given to a cell contributing to 50 % or more of to the clonal progeny compared to a cell contributing to 1 % or less. Therefore, the Shannon information index has been used to adjust for clone sizes and also focus on barcode skewing[156,172,175]. Altogether, by these high-throughput techniques, using adequate data analysis pipelines, an enormous number of individual cell fates can be analyzed and traced back to their origins in parallel. Taking into consideration that the analyzed cell population resembles only a fraction of the entire HSPC pool and its potential capacities, final extrapolation of the data needs to be performed to extract the biological context and validation of an appropriate and equally distributed sample size[176,177].

While others took advantage of lentiviral VIS data from human gene therapy trials[178-183] as unique marks for lineage tracing studies, the continuous application of integrating viruses for barcoding remains a major concern (see section 3.3). This arises from unpredictable alterations of the cellular physiology caused by viral integration (reviewed in [184]) and possibly

hampered or biased quantification of barcodes based on their position in the genome[132,185]. The aforementioned approaches circumvent these issues by implementing a defined genetic location for their alternative barcode systems. However, they also suffer from limitations due to unequally sized barcodes and, therefore, from biased amplification. Thus, it is of great interest to advance the here described barcode system by means of targeting the well-established and validated system to integrate into a 'neutral' genomic site, thereby creating a targeted barcoding approach bypassing semi-random lentiviral integration (Figure 3C). Such neutral genomic sites are termed genomic safe harbors (GSH, reviewed in [186-188]), which are characterized by offering adequate transgene expression without perturbation of endogenous genes and their function and thereby, would allow for precise un-biased barcoding.

3.5 CRISPR/Cas9 system

As one of the cutting-edge technologies in targeted gene-editing, the CRISPR/Cas9 system gained outstanding popularity in nearly all fields of biomedical and translational research since its discovery in 2012[189]. Its robustness and versatility makes the system user-friendly and applicable for generating stable gene modifications in a comparatively short time span. Unlike other tools used for gene editing, such as zinc finger nucleases (ZFNs), transcription activator-like effector nucleases (TALENs), and meganucleases, CRISPR/Cas9 allows for simple and effortless precise targeting of a host sequence.

The very first steps, towards the development of the CRISPR/Cas9 system was in 1987 the discovery of unusual repeats in the genome of *Escherichia coli* interspersed by short sequences of unknown origin[190], which were later given the name CRISPR. Adjacent to the CRISPR locus, an accumulation of uniform genes was identified in 2002, which were collectively named CRISPR-associated system (Cas)[191]. Subsequently, several groups independently reported that, hyper-variable non-repeating CRISPR spacers between the repetitive sequences were homologous to foreign DNA of plasmid and of viral origin[192-194]. Thus, it was suggested that these specialized structures might play a role in the immune defense of bacteria and archaea against phages and their transmissible genetic elements. The next milestone was the establishment of the corresponding mechanisms in 2011 by the Charpentier group[195].

The mechanism by which bacterial immunity is transmitted divides into three main steps: i) adaptation by integration of foreign DNA-derived spacers into the CRISPR locus, ii) expression and processing of CRISPR RNA (crRNA) from the CRISPR locus, which consist of unique single repeat spacer units, and iii) interference or silencing of the foreign genome[195-197]. A key event in this process is the maturation of active crRNAs from a long primary RNA transcript of the CRISPR locus (pre-crRNA), and is achieved by a *trans*-activating CRISPR RNA (tracrRNA) and a RNase encoded by the host cell organism[195]. To silence the foreign genome, the active ribonucleoprotein complex with SpCas9 endonuclease is formed[198] capable of recognizing the

target sequence and inducing double-strand breaks (DSBs) of the cognate organism, preventing further proliferation and maturation.

In 2012, Charpentier and Doudna managed to repurpose the CRISPR/Cas9 system into a gene-editing tool by engineering a single tracrRNA-crRNA duplex (single guide RNA or sgRNA) containing the sequences of both tracrRNA and crRNA[189]. Shortly after, Zhang and colleagues adopted the system in mammalian cells, showing its promising applicability for targeted genome editing[199]. For their work on the CRISPR/Cas9 system, they received the 2020 Nobel Prize in chemistry.

Gene editing by the CRISPR/Cas9 system occurs in an equal manner as the initial described mechanism. After recognizing a target sequence by the SpCas9/sgRNA complex and introduction of cleavage at the position of interest (reviewed in [200]). In detail, the complex recognizes a 20 nt sequence, with the only condition of being located adjacent to a 5'-NGG-3' protospacer adjacent motif (PAM, Figure 5A)[189,201]. Main factors for the nearly parallel cutting mechanisms of both DNA strands are the two catalytic domains of SpCas9, RuvC and HNH[202]. Endogenous DNA repair processes account for the actual genetic modification and thereby lead to disruption of genes or insertion of foreign DNA cassettes by the two commonly known main mechanisms – non-homologous end-joining (NHEJ, Figure 5B) and homology-directed repair (HDR, Figure 5C). Since NHEJ is error-prone in nature, repair of the two DNA strands occurs usually in the presence of short insertions or deletions (indels) (reviewed in [203]) and thus inserting frameshifts or premature stop codons into the targeted gene-coding sequence[204]. The second mechanism, HDR, requires a homologous template (donor) for effective DNA repair and thereby offering the possibility of transferring exogenous genetic material to the locus of choice[205] (reviewed in [206]). Donor constructs can easily be generated by cloning a gene-cassette flanked by two long homology arms (HA) complementary to the regions adjacent to the cut site and are transferred as either single-stranded of double-stranded (ds) DNA[207-210]. However, mostly present at late S/G2 cell cycle phase[205], HDR was shown to possess extremely low efficiency levels compared to NHEJ [211-213].

To this end, developing new methods to improve CRISPR/Cas9-mediated knock-in is a steadily evolving field. Three main parameters can be adjusted: targeting a specific DNA repair pathway for suppression or activation, careful evaluation of the donor sequence architecture (length, ds or ssDNA), and the adequate choice of the delivery method. Bypassing NHEJ pathways can be achieved by gene silencing[214], the use of cell lines deficient in NHEJ components[215], or suppression with small-molecule inhibitors[216-219]. An up to 19-fold increase in HDR efficiency was reported after using Scr7 to inhibit DNA ligase IV and, thereby, the NHEJ pathway[214,216]. The second option, i.e., optimizing the donor architecture, has been tried to a great extent with partially encouraging results. Several groups showed that the *in situ* linearization of the donor construct can increase HDR drastically[211,220-222] compared to the incorporation of extrachromosomal plasmid DNA. Furthermore, a decrease or absence of HA regions in linearized donor DNA was studied extensively. This strategy was termed MMEJ (microhomology-mediated end joining), and described to occur NHEJ-mediated, or HDR-independently, respectively[211,222,223]. Single components of the CRISPR/Cas9 system are

introduced as single-stranded oligonucleotides (ssODN)[224,225], naked plasmid DNA, mRNA, or IDLVs[224,226,227] encoding the genetic information or via direct transfer of SpCas9/gRNA ribonucleoproteins (RNP)[228,229]. Delivery methods range from microinjection, electroporation, viral transfer, and lipid nanoparticles to cell-penetrating peptides and depends on the delivery vector of choice.

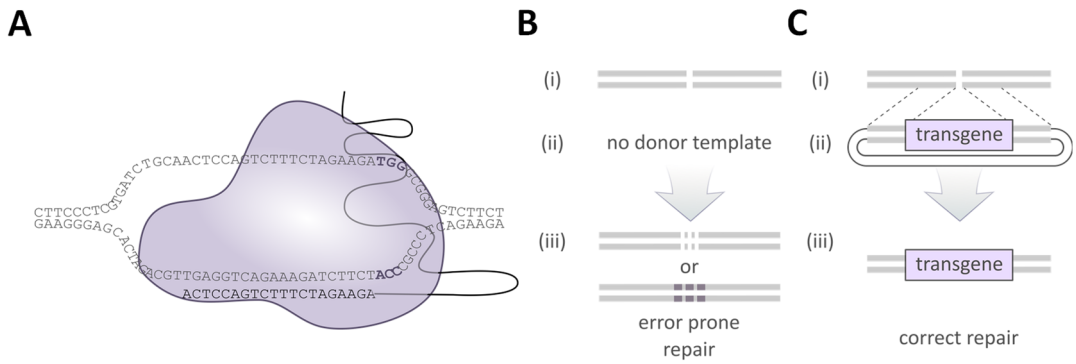


Figure 5: CRISPR/Cas9 gene editing process. DSBs are inserted by SpCas9 endonuclease, which is guided to a favored site in the genome by a short gRNA. After binding of SpCas9/gRNA complex to the site adjacent to a PAM sequence, phosphodiester bonds between bases 3 and 4 upstream of PAM are hydrolyzed. Cell-own repair mechanisms (NJEH or HDR) are used for subsequent DSB repair and can be exploited for knock-out of cellular genes and knock-in of transgenes. B) NHEJ does not require a donor template (ii) and repair is performed by minimal processing of DNA ends and direct ligation. This process introduces small indels (iii) to the site of the initial DSB and therefore is known as an error-prone repair pathway. C) In order to insert a transgene of interest, the process of HDR can be exploited by adding a donor template (ii) with flanking homology arms, homologous to the regions next to the DSB. Homologous recombination in the late S/G2 phase of the cell cycle leads to the correct insertion of the transgene into the genome (iii).
Cas: CRISPR associated, DSB: double-strand breaks, gRNA: guide RNA, HDR: homology-directed repair, indels: insertions or deletions, NHEJ: non-homologous end-joining, PAM: protospacer adjacent motif

For bacterial immunity, it seems reasonable for sgRNA binding to not rely on the exact matching of crRNA to the target DNA. Unfortunately, this promiscuity is carried over to gene-editing applications of the system, resulting in off-target effects[230-235]. Even multiple mismatches can be tolerated by the CRISPR/Cas9 complex, especially when located in regions downstream of the 3' seed region (10-12 nt adjacent to PAM)[199,236]. One obvious way to alleviate the problem is by designing guide RNAs (gRNAs) that are less likely to interact with non-specific regions. Indeed, there are several prediction algorithms available online for predicting off-targets and on-target specificity. However, a comparative analysis of different prediction tools revealed large disparities in gRNA score calculations, casting doubt over their reliability[237]. Another approach is the optimization of the experimental conditions, in particular the duration of SpCas9 exposure and SpCas9 engineering. With respect to the former, the transient exposure to SpCas9 by IDLV delivery or direct protein transfer was shown to rapidly clear SpCas9 expression after genome modification[229,238]. Regarding the latter, mutation of either of the active sites of SpCas9 enzyme leads to the generation of a nicking variant, i.e., one that induces single strand nicks instead of DSBs. It has been shown that two paired SpCas9 nickase/ gRNA complexes can induce nicks on the opposing strands, diminishing off-target effects without harming genome editing efficiency[239,240]. In a similar approach, an

inactive SpCas9 variant was fused to a FokI nuclease. Binding of two complexes in close proximity results in subsequent dimerization for successful gene modification, again diminishing off-target effects[241]. It should be mentioned that, although every CRISPR/Cas9 approach should be revised carefully for its comprehensive off-target evaluation, these new approaches usually come at the expense of genome targetability.

With respect to gene correction in HSPCs, the CRISPR/Cas9 system holds great promise for therapeutic applications and biomedical research (reviewed in [242,243]). The prospect to correct defective genes in a highly targeted fashion represents an attractive alternative to the incorporation of foreign genes into cells, due to the physiological regulation of the corrected genes and the lack of genomic integrations. However, editing efficiencies in primary cells are still unpredictable and relatively low, thus not yet sufficient for translation into clinics. Moreover, the success rate for gene delivery and gene knock-out between 26-75 % in human and murine HSPCs is heavily dependent on the applied method. A plasmid-based dual gRNA approach has been recently reported to have a rate as high as 26 %, while the use of preassembled RNP complexes display much higher success rates, in the range of 60-75 %[244,245]. Even though several groups have reported high HDR frequencies in human $CD34^+$ cells by combination of RNPs with recombinant adeno-associated virus serotype 6 (rAAV6) donor templates, only little attention has been put to murine HSPC experiments[246-251]. More recently, functional transplantation studied in immunodeficient mice reported up to 20 % targeted integration of *IL2RG* cDNA for correction of SCID-X1[252]. To date, HDR-mediated knock-in in murine HSPCs was investigated in a single study, which reported efficiencies up to 30 % in two different loci[253]. In order to achieve a therapeutic effect, protein restoration in about 10 % of the physiologically relevant cells like HSPCs is sufficient[254], suggesting that, despite low editing levels, CRISPR/Cas9 usage might be advantageous to LVs.

Alternatively to gene replacement, targeted knock-in also constitutes a useful tool for the introduction of genetic barcodes for lineage tracing studies. Uncoupling the incorporation of barcodes the semi-random LV integration would benefit from harboring barcodes on the exact same loci in all cells, as this would preclude possible effects on physiology and biases in the engraftment capacity of HSPCs. Another perspective for the optimization of the use of the CRISPR/Cas9 machinery for barcoding is to aim at equal or higher rates of marked cells than the limit present for LV transduction, due to the need for unique barcode abundance (1 barcode/ cell). As shown by the high variances in CRISPR/Cas9-dependent HDR events, there is an urgent need to adapt the CRISPR/Cas9 barcoding concept to each individual purpose, cell type, and delivery of components in order to achieve high gene editing rates.

4 Aims of the thesis

Despite the outstanding clinical success of HSCT, the success after the treatment is still unpredictable and cellular dynamics of blood reconstitution driven by HSPCs remain largely unknown. The naturally occurring surface protein CD34 is used as a surrogate marker for graft composition and quality in the clinic. Importantly, this marker is not uniquely expressed on LT-HSC, but also on a variety of downstream progenitor populations. This led to the assumption that besides HSCs other progenitors might also contribute to blood reconstitution following HSCT. Studying the influence of the exact graft composition on the clonal dynamics might add important value to existing knowledge and reveal the unique roles of the heterogenic pool of HSPC subpopulations to hematopoietic reconstitution. However, several other factors such as the damage of the niche by the pre-conditioning regimen or cytotoxic events of stem cell mobilizing agents on the engraftment ability of cells yet have to be investigated in that context. The main topic on the current work was the tracking of cells via lentiviral barcoding, and the work was divided in two parts. While the first part focused on an *in vivo* barcoding approach, the second part of the project focused on the optimization of the barcoding system by *in vitro* experiments.

The aims of the first part of the thesis were

1. Lentiviral barcoding (BC32) of four different HSPC populations in order to simultaneously track the cellular progeny in murine HSCT experiments (*in vivo*)
2. Establishing a ddPCR-based strategy to quantify barcode backbone abundance, independent of fluorescent marker expression
3. Assessing the clonal composition of engrafted cells post transplantation
4. Comparison of the clonal dynamics in similar transplantation experiments using different HSPC sources and pre-conditioning regimens.

Random lentiviral integration, as the base for the current barcoding technique, has been linked to adverse events and unpredictable genetic alterations, possibly influencing HSPC capacity to engraft post HSCT. Work, which addressed this issue, has led to contradictory results, thus aggravating to draw reliable conclusions about HSPC cell fate decisions from lentiviral barcoding strategies. A second limitation are the initial transduction rates of the target cells, which need to be kept at a low level to ensure a single barcode integration per cell. Thereby, only a certain fraction of the cell pool is analyzed. In order to prevent the random integration nature of lentiviruses, the prospect of the second part of the project was to optimize and use the CRISPR/Cas9 system for the supposedly neutral integration of barcodes in a defined locus.

To this end, the aims of the second part of the thesis were

1. Design and cloning of the vectors for a CRISPR/Cas9-based barcoding approach
2. IDLV production as non-integrating alternative delivery vehicles
3. Evaluation of the CRISPR/Cas9 editing and knock-in efficiency in a murine cell line.

5 Results

5.1 In vivo reconstitution dynamics

5.1.1 Experimental design

In order to track the contribution of HSPCs to blood reconstitution after HSCT, murine transplantation experiments were performed following the experimental set-up shown in Figure 6A. HSPCs were taken from 30 male donor animals and sorted into HSPC subpopulations, namely, HSCs, MPPs, CMPs, and CLPs, based on their phenotype. Cellular barcoding was performed by LVs containing BC32 barcode libraries and fluorescent markers (T: T-Sapphire, B: blue fluorescent protein (eBFP), G: green fluorescent protein (GFP), and V: Venus) to discriminate between the populations. Transduced cell populations were mixed with female BM support cells to support post-transplant recovery, and the transplants were distributed equally to 30 female recipient animals. Recipients were exposed to myeloablative pre-conditioning before transplantation to create a niche for the homing of transplanted cells and to avoid rejection. Post transplantation, cohorts of five animals were analyzed at different endpoints (1w, 3w, 8w, 16w, and 26w), with samples from various compartments (BM, spleen, thymus, and PB) taken for subsequent fluorescence-activated cell sorting (FACS) and NGS analyses.

In order to analyze the influence of the HSPC source and the pre-conditioning regimen on the temporal and spatial outcome, two different HSPC sources, namely BM and PB, and two different pre-conditioning regimens, namely total body irradiation (TBI) or busulfan (Bu) treatment was used. The four independent groups were designated per source of transplanted HSPCs and myeloablation regimen: BMSCs into irradiated recipients (BM-TBI), BMSCs into chemotherapy-treated recipients (BM-Bu), PBSCs into irradiated recipients (PB-TBI), and PBSCs into chemotherapy-treated recipients (PB-Bu). Unfortunately, PB-TBI mice displayed low engraftment and continuous weight loss (data not shown), thus this group was terminated prematurely. Moreover, due to unrelated problems possibly caused by the low strength of the EFS promoter, we decided to set up a new group in which the BC32 vectors contained the fluorescent marker under the control of the stronger spleen focus-forming virus (SFFV) promoter. To compare the groups, experimental conditions (HSPC source and myeloablation) were chosen equal to those used in BM-TBI, and therefore named BM-TBI^{SFFV}.

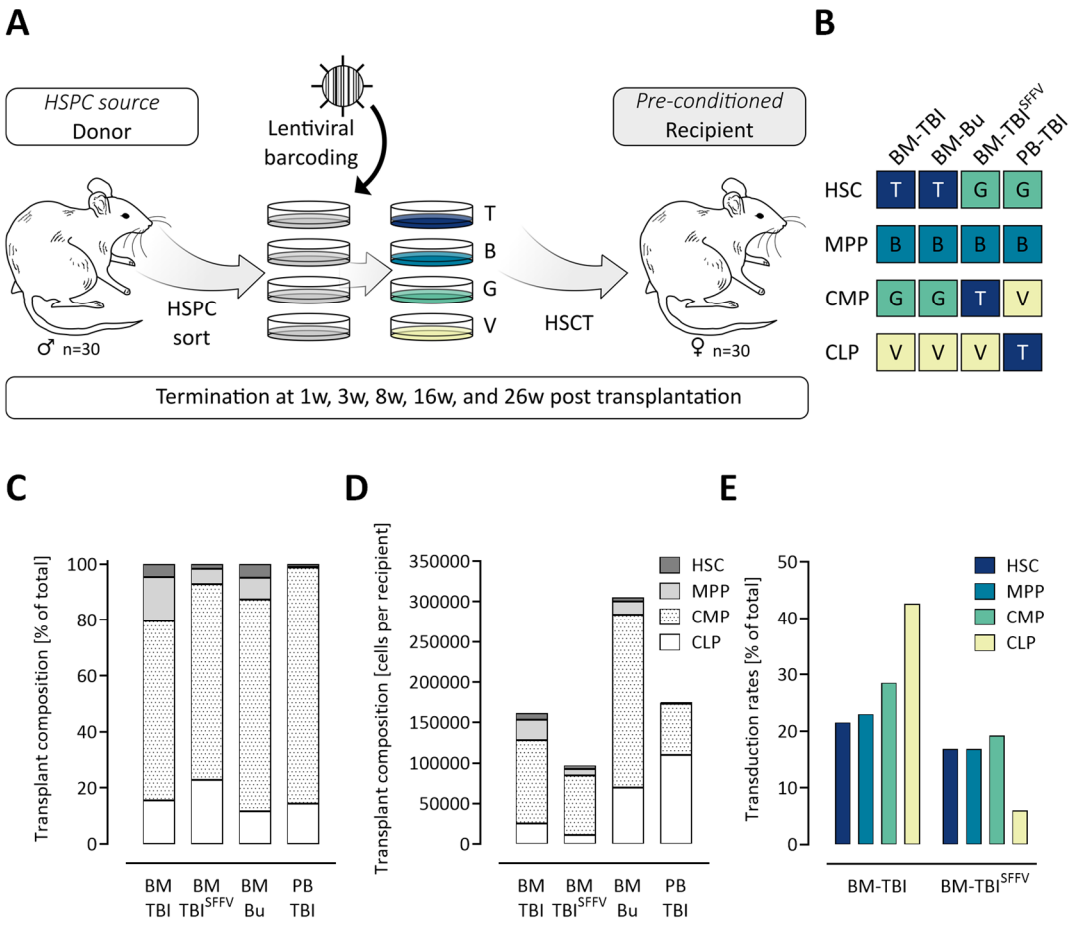


Figure 6: Experimental procedure and transplant composition. A) Murine HSCT experiments were carried out using HSPCs from donor animals, which were sorted into four subpopulations (HSC, MPP, CMP and CLP) and barcoded using lentiviral BC32 libraries with different fluorescent markers (T, B, G, and V). After mixing the transplant with a BM support, transplantation was performed by intravenous injection into myeloablated recipient animals. Final analyses were performed 1w, 3w, 8w, 16w and 24w post transplantation and BM, spleen, thymus and PB was sampled for subsequent FACS and NGS analyses. B) The transduction of HSPCs for each experiment was performed with a randomly chosen BC backbone (here schematically shown in dark blue (T), blue (B), green (G) and yellow (Y)) to avoid bias during analysis. C) Composition of the transplant comprised of HSCs (white bars), MPPs (light grey bars), CMPs (white dotted bars) and CLPs (dark grey bars). Shown are the transplant compositions of the four experiments BM-TBI, BM-TBISFFV, BM-Bu and PB-TBI. Frequency of cell population was normalized to the total count of cells before transduction. D) Total amount of the four different subpopulations per recipient animal in a transplant. Bars represent the cell populations HSCs (white bars), MPPs (light grey bars), CMPs (white dotted bars) and CLPs (dark grey bars). E) Transduction rates in the four different subpopulations analyzed by FACS analysis for BM-TBI and BM-TBISFFV groups. Note that color of the bars is not indicative of the fluorescent marker.
BM: bone marrow, Bu: busulfan, CLP: common lymphoid progenitor, CMP: common myeloid progenitor, HSC: hematopoietic stem cell, HSCT: hematopoietic stem cell transplantation, HSPC: hematopoietic stem and progenitor cells, MPP: multipotent progenitor, PB: peripheral blood, TBI: total body irradiation

In order to track the four different HSPC subpopulations, four different barcode libraries were used for the generation of the BC32 vectors, harboring the additional information of one out of four fluorescent markers. This marking system allowed to simultaneously track the four different populations, linking each unique barcode, characterized by its backbone sequence, to one of the four fluorescent markers (see also Table 11). As shown in Figure 6B transduction of different HSPC populations between experiments was performed with a randomly chosen BC32 vector in order to avoid bias in analyses. Thus, e.g. in the BM-TBI group the

HSPC-construct allocation was HSCs: T-Sapphire, MPPs: eBFP, CMPs: GFP, and CLPs: Venus, while for the PB-TBI group designation was HSCs: GFP, MPPs: eBFPs, CMPs: Venus, and CLPs: T-Sapphire. Besides fluorescent marker expression, distinction of the four subpopulations was also possible through DNA-based recovery of the backbones, as the latter contained additionally short identifier (ID) sequences. To sum up, the specific design of backbone sequences allowed for different layers of analysis: (i) fluorescent marker expression via FACS, (ii) droplet digital PCR (ddPCR)-based analysis to trace back mature blood cells to the precursor population, and (iii) in-depth analysis and tracking of single-cell clones via NGS analysis of the barcodes in hematopoietic organs (Figure 7). The complexity values of generated plasmid libraries were comparable in all experiments and were in the desired range of 10^6 to 10^8 barcodes.

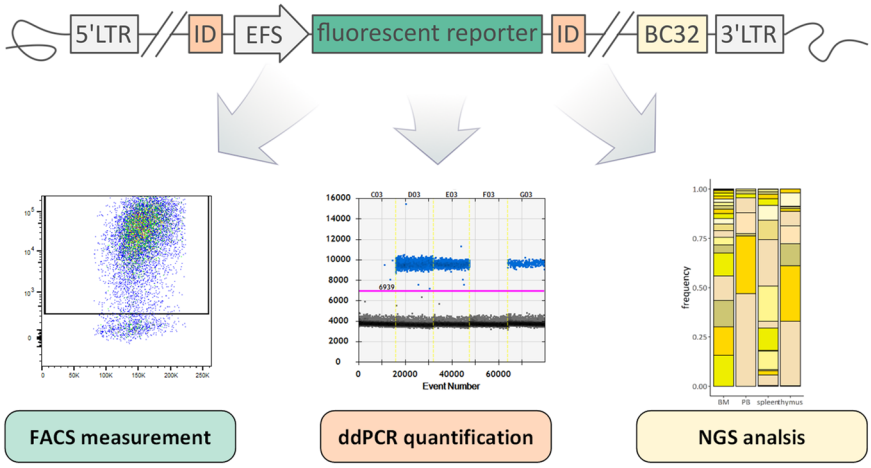


Figure 7: Methods of detection of BC32 vectors. The backbones of BC32 vector were designed to contain a fluorescent marker for FACS measurement, a short ID tag to additionally allow for ddPCR-based quantification of the backbone, as well as the BC32 barcode library for in-depth clonal analysis by NGS. Exemplary pictures taken from the experiment to depict each method are shown.
BC: barcode, ddPCR: droplet digital PCR, EFS: Elongation factor 1a short, FACS: fluorescence activated cell sorting, ID: identifier, SFFV: Spleen focus-forming virus, LTR: long terminal repeat, NGS: next-generation sequencing

5.1.2 HSPC cell counts for transplantation

After sorting the HSPCs from the donor source, the abundance and frequency of each subpopulation were determined. As shown in Figure 6C (relative abundance) and Figure 6D (absolute cell counts per transplanted animal), some similarities among the transplants of the four HSCT groups were observed, while others differed strongly. Total cell counts ranged from 2.9×10^6 to 9.2×10^6, ending up in a maximum of 9.7×10^4 to 3.1×10^5 cells of different composition per transplanted animal (before viral transduction). In all groups, transplants were mostly comprised of CMPs (64 – 84 %); HSCs and MPPs accounted for lower frequencies (1 – 20 %). However, in two groups, namely the BM-TBI^{SFFV} and PB-TBI groups, an overall lower frequency of HSC population (2.8-fold and 3.4-fold reduced) compared to the other two groups was observed (1.7 % and 0.8 % compared to a mean of 4.8 % HSCs in the BM-TBI and

BM-Bu group). Furthermore, mobilization of PBSC in the PB-TBI group resulted in 25-fold reduced MPP levels compared to the mean MPP frequency of groups using BMSC, constituting only 0.38 % of the graft.

Barcoding of each HSPC subpopulation was performed by lentiviral transduction with the respective library. As efficiencies were aimed to range below 20 % in order to ensure single barcode integration, a moderate MOI of 30 was used. Determination of the transduction rates were only performed in the BM-TBI and BM-TBI^{SFFV} groups, as enough cells for additional FACS measurement were available. Notably, rates in all respective subpopulations were higher in the BM-TBI group (Figure 6E). Surprisingly, CLPs had both the highest (42.6 % in BM-TBI) and the lowest (6 % in BM-TBI^{SFFV}) transduction rate among all subpopulations in both groups. HSC, MPP, and CMP transduction rates ranged between 21—29 % (BM-TBI), and 17—19 % (BM-TBI^{SFFV}).

5.1.3 Hematopoietic reconstitution and engraftment

To characterize the degree of hematopoietic reconstitution, total leukocyte counts were analyzed via FACS in spleen samples over time (Figure 8C). Both, BM-TBI and BM-TBI^{SFFV} showed similar reconstitution patterns of total leukocytes. Specifically, starting from frequencies of 30.7 % in the BM-TBI group at 1w, the levels increased up to 89.5 % at 16w post transplantation, when a plateau was reached. Comparable patterns were observed for the BM-TBI^{SFFV} group: leukocyte frequencies increased from 15.5 % (1w) to 93.3 % (16w) throughout the observation period. On the other hand, the BM-Bu group gave relatively stable and high leukocyte counts up until the premature end of the experiment (8w), with the lowest abundance (89.5 %) observed at 3w. Finally, the PB-TBI group exhibited a delayed recovery, with relatively low total leukocyte abundance up until 8w post transplantation, at which point the mice were sacrificed.

Besides the total leukocyte abundance, the relative representation of leukocyte subgroups was determined. Based on their immunophenotypes leukocytes were further distinguished in T cells (CD3e$^+$), B cells (B220$^+$), monocytes (CD11b$^+$), and neutrophils (Ly6G$^+$CD11b$^+$). Changes in T cells (Figure 8D) and B cells (Figure 8E) in the BM-TBI group displayed a pattern similar to the one observed in the BM-TBI^{SFFV} group. Levels of T cells in both groups showed a steady increase during the first 16w post transplantation. Shortly after transplantation (1w), low T cell levels between 5—13 % were found. After 16w, the frequencies increased up to 33 % in both groups. B cells in both groups also displayed a steady increase, starting at 0.4 – 6 % (1w), and reaching a stable abundance at 16w post transplantation (48 % -- 52 %). On the other hand, in the BM-Bu group, T cells (61 %) and B cells (30 %) started at high levels at 1w post transplantation. Both populations reached a plateau of around 40 % until the final termination point at 8w. After a slower recovery in PB-TBI compared to the other groups, levels of T cells and B cells reached maximum values of 36.9 % or 25.1 %, respectively, at the termination date. Monocyte and neutrophil counts displayed equal patterns in all groups (Figure 8F and Figure 8G). Here, the basal levels remained below 8 %, showing low variation in the reconstitution of

these cell populations. Both populations peaked at 3w post transplantation (BM-TBI and BM-TBI^{SFFV}) but declined to the basal levels at later time points.

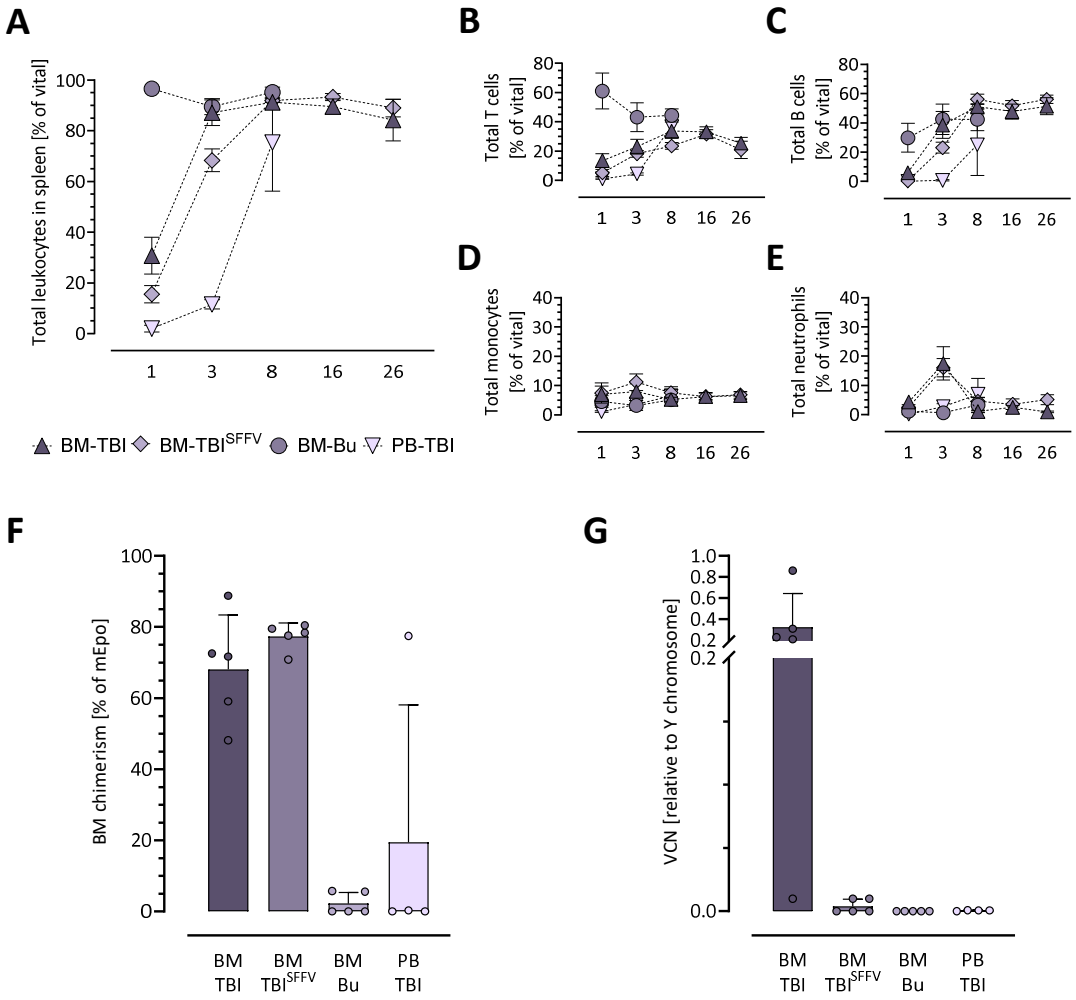


Figure 8: Engraftment and hematopoietic reconstitution after stem cell transplantation. A) Total leukocyte counts from spleens were analyzed to investigate the dynamics in hematopoietic recovery and in splenic cellular subsets. In all four mice experiments T cells (B), B cells (C), monocytes (D), and neutrophils (E) were analyzed (mean ± SD). Missing data points indicate premature termination of the experiments (n=5). F) Chimerism at 16w (BM-TBI and BM-TBISFFV) or 8w (BM-Bu and PB-TBI) post transplantation (mean ± SD). Engraftment of donor cells was analyzed in BM by measuring the amount of Y chromosome relative to the housekeeping gene mEpo via ddPCR (n=5). G) To evaluate the approximate number of barcodes per transplanted cell, VCN was measured via ddPCR at the same time points in BM (mean ± SD). The amount of Illumina sequence in the backbones was normalized to the amount of Y chromosome (n=5).
BM: bone marrow, Bu: busulfan, PB: peripheral blood, TBI: total body irradiation, VCN: vector copy number

As a measure for the success of donor engraftment, chimerism was determined in BM samples 8w or 16w post transplantation. To this end, frequencies of Y chromosome were measured via ddPCR and the determined copy numbers were normalized to a reference gene (*mEpo*). On average, chimerism levels were 68 % (BM-TBI) and 77.4 % (BM-TBI^{SFFV}) (Figure 8F). Lower chimerism levels were observed for BM-Bu (2.3 %) and PB-TBI (19.5 %), indicating inadequate engraftment of transplanted cells in the analyzed animals. Due to low engraftment and

continuous weight loss of the animals in both groups (data not shown), the experiments were terminated prematurely and the respective mice were sacrificed.

Beyond the total amount of engraftment after HSCT, determining the abundance of traceable transduced cells in recipient animals was of special interest. To this end, vector copy numbers (VCNs) were measured in BM samples by quantifying a BC32 vector-specific sequence (Illumina sequence) by ddPCR. As depicted in Figure 8G, the mean frequency of VCN 16w post transplantation was 0.32 in the BM-TBI group, which was in line with the mean of the initial transduction rates of HSPC subpopulations (28.9 %). In contrast, the three other groups showed either no (BM-Bu) or extremely low VCNs (BM-TBI^{SFFV} and PB-TBI), indicating low marking of engrafted cells.

5.1.4 Tracking of marked HSPC progeny

The influence of differently marked HSPC subpopulations on hematopoietic reconstitution was investigated by measuring the fluorescent marker expression in BM samples collected upon termination. Figure 9A-D displays the temporal changes in fluorescent marker expression derived from the transduced HSPC subpopulations divided by the experimental groups. Values below 1 % were considered non-relevant, i.e., indicative of an absence of FP expression.

In the BM-TBI group (Figure 9A), detectable fluorescent marker expression was only observed from MPP-derived cells at 1w post transplantation. Equally low rates of fluorescent marker expression was found in the BM-TBI^{SFFV} group (Figure 9B); CMP- and CLP-derived cells did not display detectable fluorescent marker expression levels at any time throughout the observation period. In contrast, both HSC- and MPP-derived cells showed detectable fluorescent levels of reporter protein expression at 1w post-transplantation. Then, expression levels dropped below the detection limit where they remained either for the rest of the experiment (HSCs) or until 26w (MPPs). Notably, the reemerging MPP signal at 26w was much weaker than the signal at 1w post transplantation. As a whole, the differences between the BM-TBI and BM-TBI^{SFFV} groups suggest that the impact of the stronger SFFV promoter in the BC32 constructs was evident but not prominent. As anticipated by the observed engraftment rates and VCNs no marked progeny of transplanted HSPCs was detected by FACS analysis in the BM-Bu group (Figure 9C). Finally, similarly to the BM-TBI group, in the PB-TBI group (Figure 9D), detectable fluorescent marker expression was only observed from MPP-derived cells at 3w post transplantation.

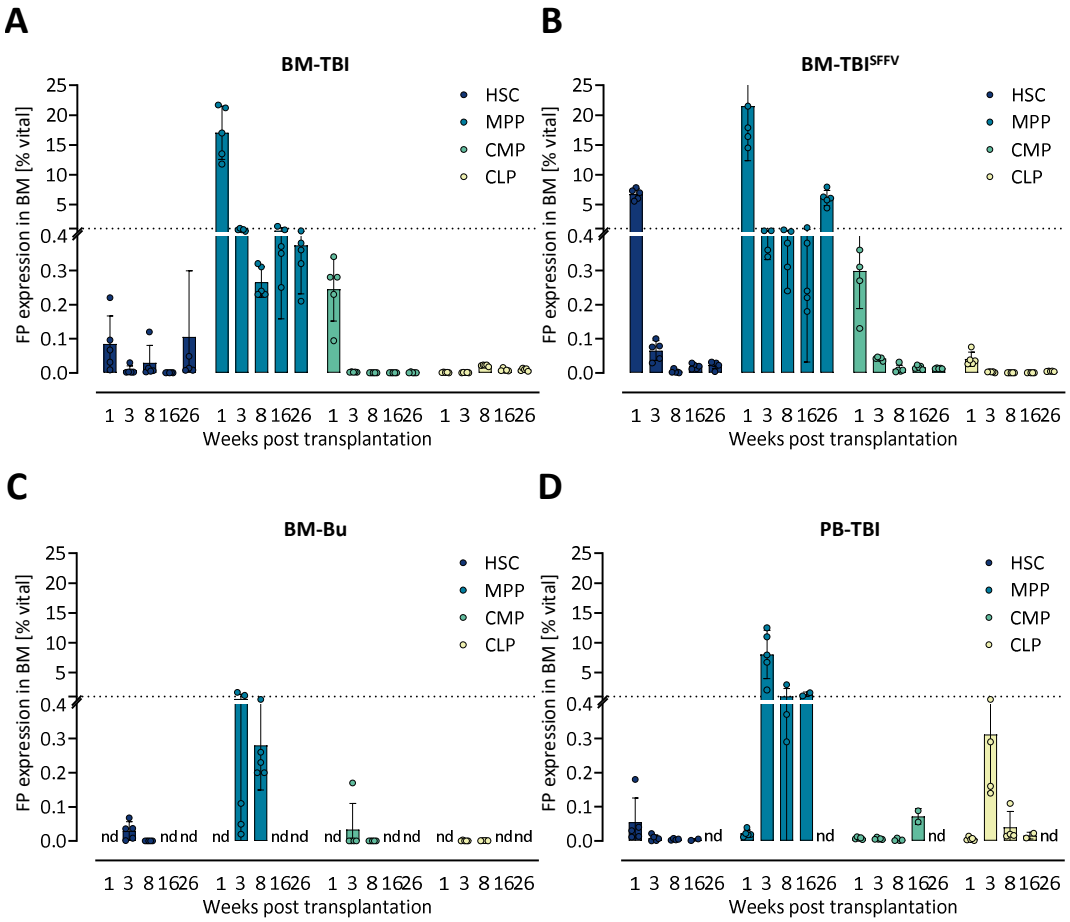


Figure 9: Contribution of barcoded HSPCs-derived cells to BM reconstitution after HSCT. Expression of fluorescent markers was measured in BM via flow cytometry analysis at 1w, 3w, 8w, 16w, and 26w post transplantation in the experimental settings BM-TBI (A), BM-Bu (B), BM-TBI^{SFFV} (C) and PB-TBI (D). Data shows fluorescent marker expression in mature tissue after engraftment of initial barcoded and transplanted HSPCs, thus the cellular output derived from the initial transplanted populations (mean ± SD). Note: depicted colors do not indicate fluorescent reporter protein colors, since designation of subpopulation and barcode backbone was distributed randomly. For better visualization of low values, a discontinuous y-axis was used, with the value 0.4 serving as both a gap and a scale break. Values below one were considered non-relevant, i.e., indicative of an absence of FP expression (ticked lines) (n=5).
BM: bone marrow, Bu: busulfan, CLP: common lymphoid progenitor, CMP: common myeloid progenitor, HSC: hematopoietic stem cell, HSPC: hematopoietic stem and progenitor cells, MPP: multipotent progenitor, PB: peripheral blood, TBI: total body irradiation
nd: not determined

Corresponding FACS analyses were also performed in the spleen samples collected upon termination of the experiment, to detect the contribution of marked HSPC subpopulations to splenic recovery after HSCT.

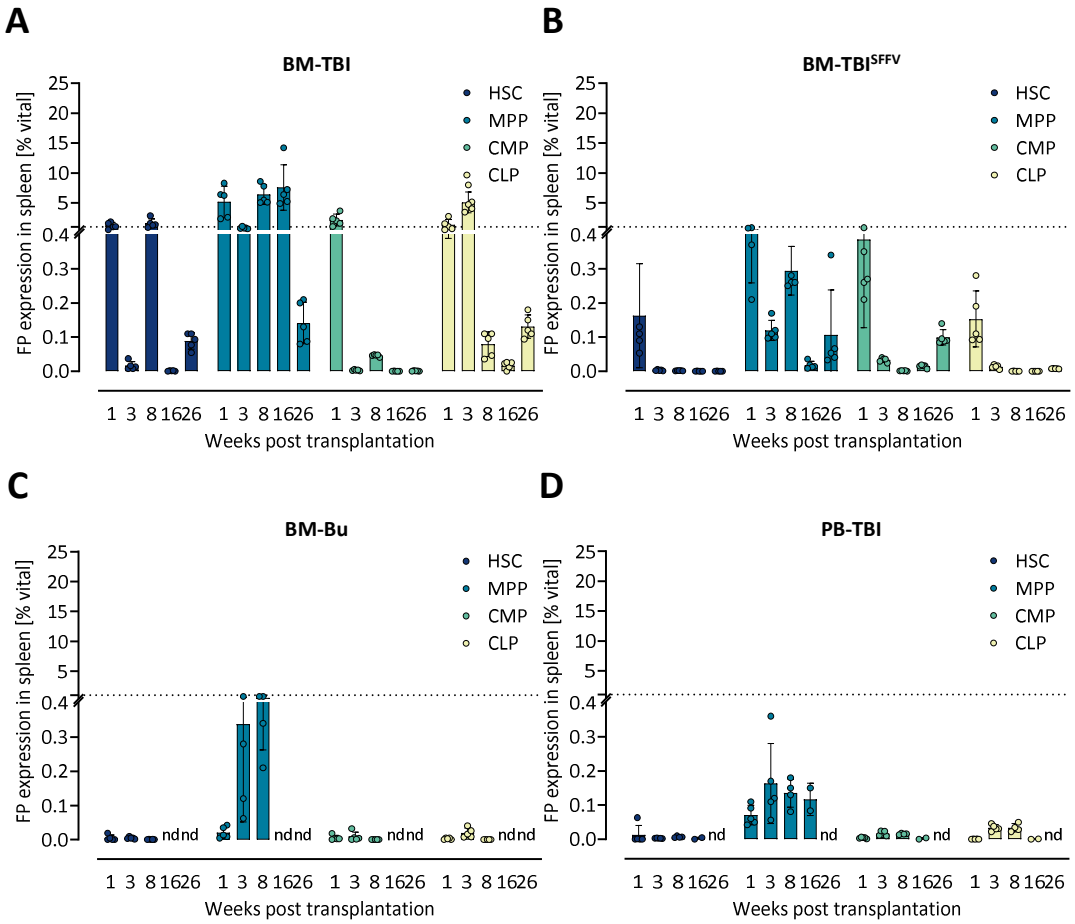


Figure 10: Contribution of barcoded HSPCs-derived cells to splenic reconstitution after HSCT. Expression of fluorescent marker was measured in spleens via flow cytometry analysis at 1w, 3w, 8w, 16w, and 26w post transplantation in the experimental settings BM-TBI (A), BM-Bu (B), BM-TBI^{SFFV} (C) and PB-TBI (D). Data shows fluorescent marker expression in mature tissue after engraftment of initial barcoded and transplanted HSPCs, thus the cellular fluorescent reporter protein expressing output derived from the initial populations (mean ± SD). Note: depicted colors do not indicate fluorescent protein colors, since designation of subpopulation and barcode backbone was distributed randomly. For better visualization of low values, a discontinuous y-axis was used, with the value 0.4 serving as both a gap and a scale break. Values below one were considered non-relevant, i.e., indicative of an absence of FP expression (ticked lines) (n=5).
BM: bone marrow, Bu: busulfan, CLP: common lymphoid progenitor, CMP: common myeloid progenitor, HSC: hematopoietic stem cell, HSPC: hematopoietic stem and progenitor cells, MPP: multipotent progenitor, PB: peripheral blood, TBI: total body irradiation
nd: not determined

The BM-TBI group showed a diverse pattern of fluorescent marker expression (Figure 10A). HSC-derived cells were visible at 1w and 8w post transplantation (1.2 % and 1.6 %, respectively). CMP- and CLP-derived cells were exclusively detected at the first (1w) or the first two (1w and 3w) time points, respectively, with frequencies up to 5 % (CLPs at 3w). In general in the early phase post transplantation more HSPC-derived marked cells were observed, but the total abundance was lower and declined to zero in the long-term phase of hematopoietic reconstitution (up to 26w). However, in contrast to the data obtained from BM, in the BM-TBI^{SFFV} (Figure 10B), BM-Bu (Figure 10C), and PB-TBI groups (Figure 10D) no fluorescent signal of marked HSPC progeny was detected in the spleen samples. These results show that the analysis of subpopulation kinetics via fluorescent marker expression is highly dependent on the adequate marker expression in the respective compartment. Despite the

high chimerism rates observed in BM-TBI and BM-TBI^{SFFV} groups, detection limit for FACS analysis was barely exceeded.

5.1.5 Clonal analysis of HSPC reconstitution

Beyond fluorescent marker expression in mature cells, NGS analysis represents a powerful alternative method to track cellular development. Thus, DNA harvested from BM, PB, spleen, and thymus samples at 1w, 3w, 8w, 16w, and 26w post transplantation was used to investigate the barcode contribution. Only samples from the BM-TBI group were analyzed, as this group seemed the most promising based on both transduction rates and fluorescent marker expression. In each NGS run approximately 1160 cells per sample were analyzed. Data was processed with respect to error correction and detection limit of barcode contribution (as described in section 8.14). Clone numbers of barcodes (count of unique barcodes of one origin per sample, regardless of clone size) were analyzed and displayed in Figure 11A-D. Despite *interindividual* variability, mean clone counts decreased over time in the analyzed compartments, independent of the BC32 vector or the subpopulation (Table 25, Appendix A, last row). This suggests an overall rather limited amount of clones contributing to stable blood formation over time (Figure 11A-D). Detailed analysis of the marked HSPC subsets showed a similar contribution pattern of MPP- and CMP-derived clones in BM, spleen, and thymus samples (Figure 11A-C and Table 25, Appendix A). While MPP-derived clones were present throughout the observation period of 26w, CMPs contributed mainly 1w post transplantation, after which barcode abundance declined to nearly zero. In the same tissue types, HSCs and CLPs contributed to a lower extent to blood reconstitution at 1w post transplantation (see Figure 11A-C and Table 25, Appendix A), after which only a small number of mostly HSC clones remained throughout the analyzed period of 26w. A different situation, compared to the rest of the analyzed tissue types, was observed in PB (Figure 11D). Up to 8w post transplantation, mostly CMP-derived barcodes were present, showing a higher number of different CMP clones contributing to the PB formation short-term after HSCT. MPP- and HSC-derived output was also detected at a moderate level up until 8w post transplantation but to a lower extent compared to CMPs. This shows that the contribution of these cells, though existent, was smaller than the contribution of the CMP population. From 16w and onwards, CMPs and CLPs were hardly detectable in any of the examined compartments. However, it is important to mention that this analysis did not take the sizes of contributing clones into account.

Next, the size of clones at two different time points (short-term - 1w and long-term - 16w) were analyzed in order to assess clonal heterogeneity and determine the proportional contribution of each clone. To this end, the frequencies of all occurring barcodes were calculated per compartment and depicted as stacked bar plots. Inclusion criteria for the analysis were minimal read counts of ≥3 and a minimal contribution of 1 % to the composition of any tissue. In Figure 12A-D representative plots of five animals at 1w post transplantation are shown for different subpopulations in the respective compartments. Presence of equal barcodes among compartments is visualized by matching color and positions in the plots.

Since plotted data does not provide information on the level of abundance (read count) of different barcodes, mean read counts of HSPC subpopulations are listed in Table 25 (Appendix A). Thus, putative dominant barcodes with low read counts can be distinguished from 'true' dominant barcodes representing high levels of read counts.

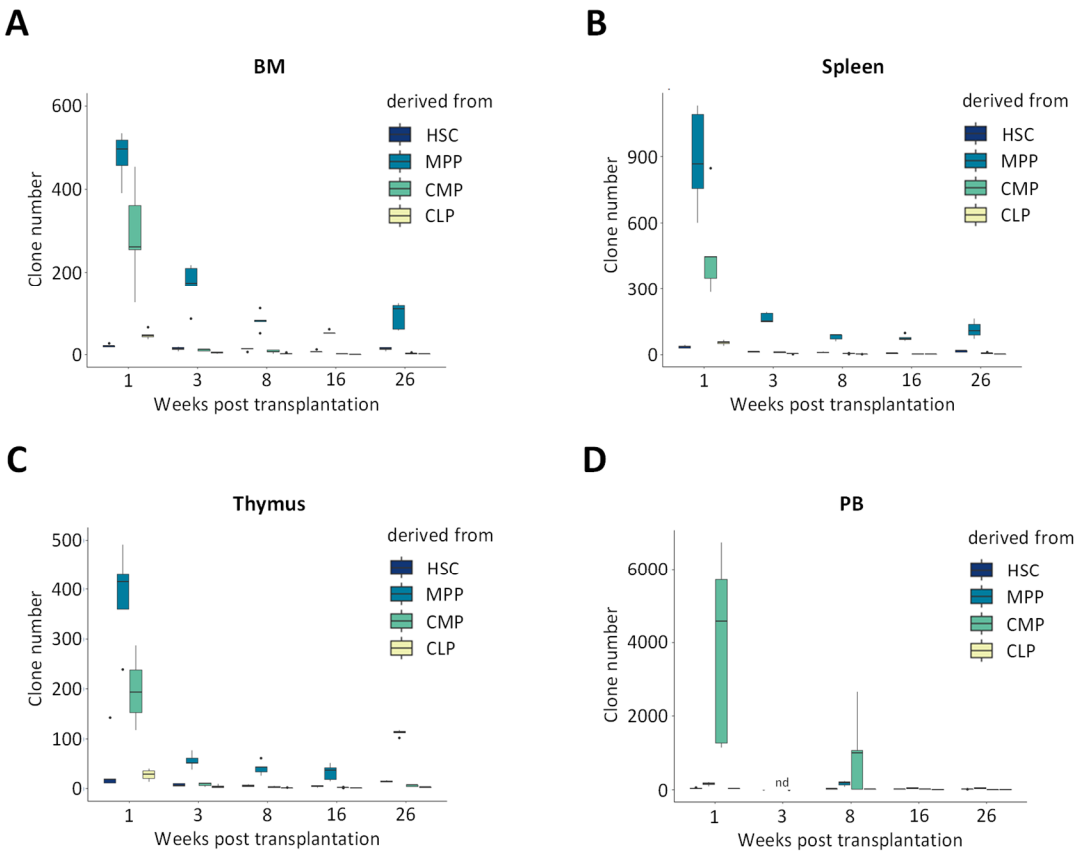


Figure 11: Quantification of the barcode backbone by NGS. For in-depth barcode analysis via NGS, DNA was isolated from BM, PB, spleen, and thymus and amplified according to the respective protocol. A)-D) The number of unique barcode clones from the initially marked subpopulations is shown in BM (A), spleen (B), thymus (C) and PB (D) samples. Presence of barcodes from the initially marked subpopulation represents progeny of the respective cell population (HSC – dark blue, MPP – blue, CMP – green and CLP - yellow) at the indicated time point (n=5).
BM: bone marrow, CLP: common lymphoid progenitor, CMP: common myeloid progenitor, HSC: hematopoietic stem cell, MPP: multipotent progenitor, PB: peripheral blood
nd: not determined

Clones of HSC-derived cells showed an overall moderate polyclonal pattern, with the majority of clones contributing equally to reconstitution in the compartments (Figure 12A). In some animals, a rather dominant clonal behavior was observed (contribution of one clone >60 %), suggesting that single major clones had already outcompeted minor clones even by the early time point of 1w post transplantation. Notably, this clonal dominant behavior seemed to arise independently of tissue type. The most characteristic examples thereof were the PB of subject #31097, where one clone contributed more than 86 %, and the spleen of subject #31098, where a single clone yielded 97 % of all reads. Other dominant clones were detected in the BM and thymus of subject #31099, as well as in the PB of subject #31121. Another interesting

observation was the overlap of clone abundances between tissue types, as demonstrated by corresponding colors of clones in the plot. HSC-derived cells produced a homogeneous output in between all compartments, with no differences in barcode abundance in a single or throughout several compartments, showing an even distribution of unique clones throughout the compartments (see also Appendix B-1). The phenomenon of barcode overlap was clearly visible in subject #31121, in which two barcodes were present in all four compartments. However, their abundance varied greatly among tissue types. Furthermore, subjects #31096 and #31098 showed the abundance of one prominent barcode each in all four analyzed tissue types, showing one clone highly contributing to the cellular output in the organism.

As expected due to the overall higher number of MPP-derived clones (as described in Figure 11), the clonal profiles of the MPP-derived clones had a much higher variation in contributing clone numbers (and thus lower contribution of each clone) compared to the profiles of HSC-derived clones (Figure 12B). Most of the samples showed a highly polyclonal pattern with the exception of one dominant clone in the thymus of subject #31121 (79 %). Other thymus samples were composed of a variety of MPP-derived barcodes. Indicative overlap between tissue types was observed more frequently, highlighting the highly variable contribution of clones (see also Appendix B-2). While in subject #31098, a highly similar MPP-derived pattern was observed in spleen and thymus samples, in subject #31099 barcode sharing occurred between PB and thymus samples.

Detailed analysis of CMP-derived clones and their contribution (Figure 12C) revealed a highly polyclonal situation. In the majority of tissue samples, a high amount of equally contributing barcodes were found. Again, in some animals, a few major clones were observed, independently of the compartment (PB, spleen, or thymus). Furthermore, significant abundance of barcode overlap in more than one compartment was found (see also Appendix B-3).

Contribution of short-lived CLP-derived clones was mainly polyclonal without any significant differences between animals (Figure 12D). CLP-derived clones were detectable in multiple tissue types, though without significant barcode overlap between compartments (see also Appendix B-4). Compared to other subpopulations, such as MPPs or CMPs, lower abundance and variety of clones was detected, as observed by less clones contributing to a higher level in the respective organs.

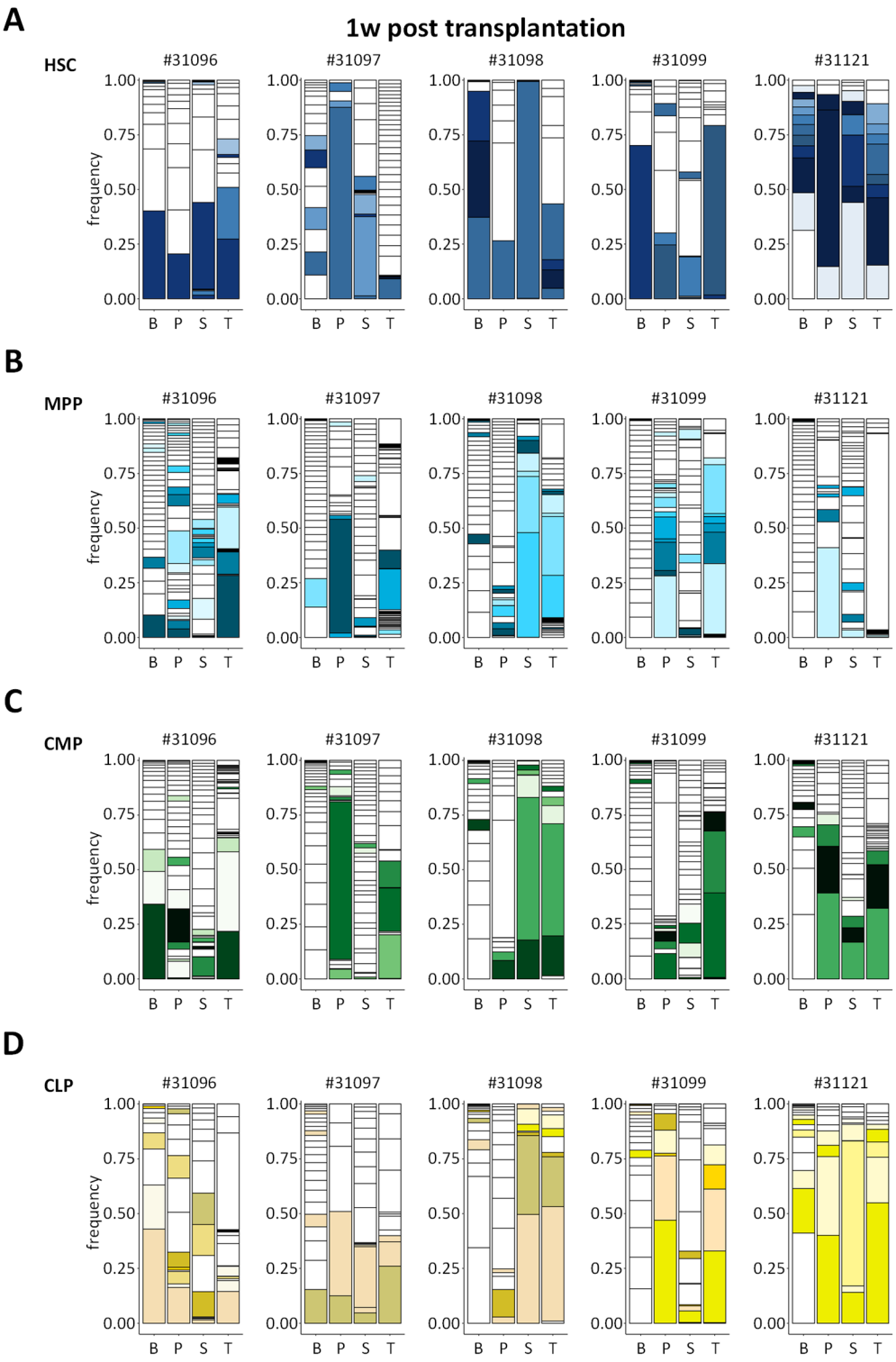


Figure 12: Short-term spatial dynamics of hematopoietic reconstitution. A-D) Stacked bar plots from five animals 1w post transplantation are shown, visualizing barcode frequencies per tissue (B: BM, P: PB, S: spleen, and T: thymus). Barcode backbones are color-coded and initial HSPC progenitor population is indicated on the left (HSCs: dark blue (A), MPPs: light blue (B), CMPs: green (C) and CLPs: yellow (D)). Equal barcodes showing overlap between compartments are visualized in the same 'horizontal' color. Each plot represents one animal. BM: bone marrow, CLP: common lymphoid progenitor, CMP: common myeloid progenitor, HSC: hematopoietic stem cell, MPP: multipotent progenitor, PB: peripheral blood

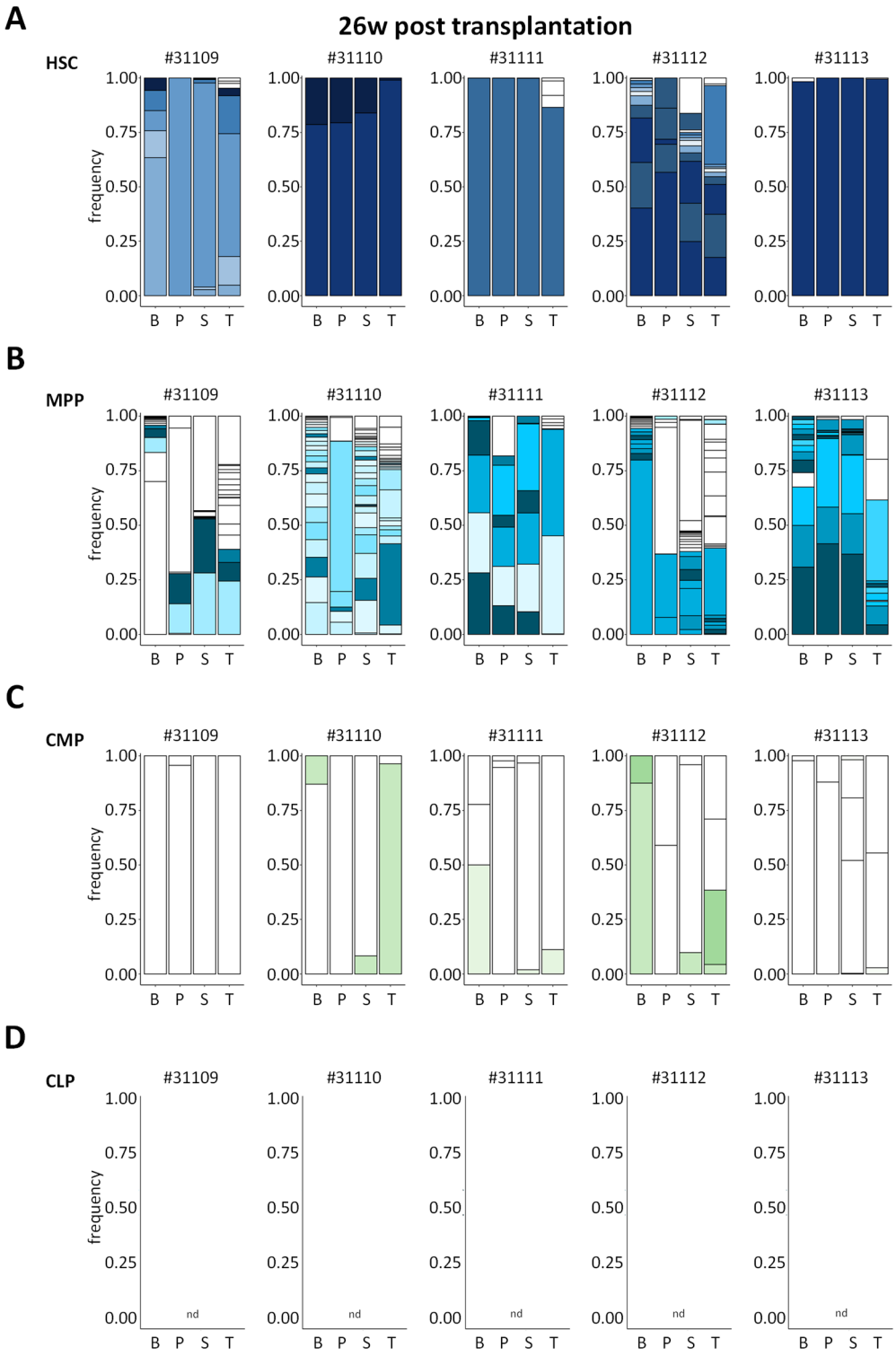


Figure 13: Long-term spatial dynamics of hematopoietic reconstitution. A-D) Stacked bar plots from five animals 26w post transplantation are shown, visualizing barcode frequencies per tissue (B: BM, P: PB, S: spleen, T: thymus). Barcode backbones are color-coded and initial HSPC progenitor population is indicated on the left (HSCs: dark blue (A), MPPs: light blue (B), CMPs: green (C) and CLPs: yellow (D)). Equal barcodes showing overlap between compartments are visualized in the same 'horizontal' color. Each plot represents one animal.

BM: bone marrow, CLP: common lymphoid progenitor, CMP: common myeloid progenitor, HSC: hematopoietic stem cell, MPP: multipotent progenitor, PB: peripheral blood

nd: not detected

Comparing the data per animal between the HSPC subpopulations indicated in some animals an overall lower amount of barcodes with higher clonal abundances in one tissue (e.g., #31097 in PB or #31098 in spleen). Apart from those rare situations, no conclusions could be drawn between the subpopulations per animal. Furthermore, at the analyzed time point no tendency of higher contribution of subpopulations linked to specific tissue types was found. Of note, here a rather early time post transplantation was analyzed, increasing the likelihood of a polyclonal situation of cells shortly after engraftment. This at later time points might decrease and early dominant clones might point towards a clonal dominance at later stages.

Stable engraftment and contribution was further evaluated in the subsequent time points, in order to draw conclusions to clonal kinetics. Over time, the clonal contribution of certain clones increased in mostly all subpopulations and tissue types: especially the total amount of CMP- and CLP-derived clones diminished as quickly as 3w post transplantation and thus showing higher frequencies of contribution of the residual barcoded cells (data not shown). To compare temporal dynamics of clonal behavior equivalent analyses were performed for five animals at 26w post transplantation. Barcode contribution data was plotted, as shown in Figure 13A-D. In general, an overall lower amount of contributing clones was observed 26w post transplantation in all compartments, as described by clone number analysis before (Figure 11 and Table 25, Appendix A). While the total number for HSC-derived clones was greatly reduced and in most of the animals a maximum of five contributing clones was detected, no significant barcode overlap was detected (Figure 13A, Appendix B-3). In some animals, a few dominant BM-present HSC clones seemed to have completely reconstituted the analyzed compartments, showing a monoclonal HSC-derived pattern in all except for one subject (#31112). However, taking low initial marking rates into consideration, this assumption only included a certain fraction of analyzable cells.

After 26w, the highest amount and variability was present in MPP-derived cells, compared to other subpopulations (Figure 13B). More clones showed active contribution to several tissue types, since highly significant barcode overlap was present in the analyzed animals (see also Appendix B-2). Some actively contributing clones (up to five) were more present than others, as seen in subjects #31111, #31112, and #31113 by a few highly abundant clones accounting for up to 98.6 % in subject #31111.

As numbers of CMP-derived clones also decreased rapidly starting from 3w post transplantation (see Figure 11 and Table 25, Appendix A), at 26w only few clones were detected in all compartments (Figure 13C). Shared barcodes between tissue types were only present in three subjects (#31110, #31111, and #31112) in two or three compartments (see also Appendix B3). No residual clones or contribution was observed of short-lived CLP population at 26w post transplantation. Throughout the experimental procedure, CLP numbers diminished consequently starting from 16w post transplantation, until their complete extinction (Figure 13D).

The overall picture, which was observed over time compared to 1w post transplantation, was a clear reduction in barcode numbers (less polyclonal) and therefore, an increase in contribution of single, stable clones. More barcode sharing between tissue types suggested

long-term output from active clones between tissue types. As the experimental set-up did not allow for consecutive sampling in the animals, temporal dynamics of reconstitution are not included in this analysis. Moreover, total clone numbers are disregarded in this analysis (Table 25, Appendix A).

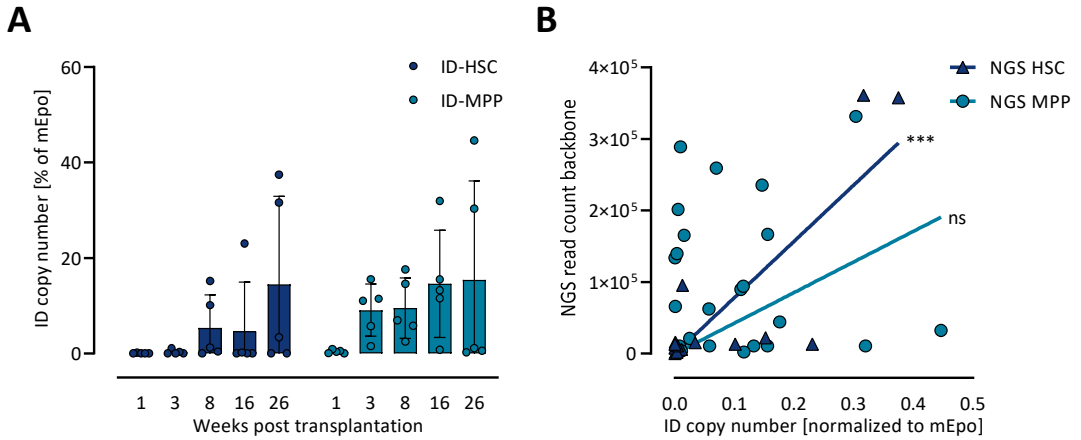


Figure 14: Quantification of the barcode backbone by ddPCR. For in-depth backbone analysis by ddPCR, DNA was isolated from BM samples and amplified according to the respective protocol. A) Barcode backbone copy numbers of the two most abundant subpopulations (HSC – dark blue and MPP – blue) in BM samples at 1w, 3w, 8w, 16w, and 26w post transplantation (mean ± SD). Backbones were equipped with a short ID tag representing the initial subpopulation (here: ID-HSC and ID-MPP) to allow for DNA discrimination (n=5). B) Both datasets of BM samples – NGS and ddPCR – were used to correlate the data (n=25). Significant correlation was only found in HSC-derived data.
ddPCR: droplet digital PCR, HSC: hematopoietic stem cell, ID: identifier, MPP: multipotent progenitor, NGS: next-generation sequencing
ns: not significant, *** $p \leq 0.001$

HSCs and MPPs are two key players of the hematopoietic hierarchy, as emphasized by the clone sizes, which contributed to hematopoietic recovery (Table 25, Appendix A). Therefore, ddPCR was used to validate the NGS data regarding these two populations. To establish the quantification of ID sequences in the backbones and compare the data to the NGS results, corresponding BM samples were taken from the BM-TBI group from all analyzed time points. Detection of HSC and MPP populations in BM samples was performed via specific primer and probe sets of IDc (ID-HSC) and IDb (ID-MPP) (see also Table 14), with *mEpo* serving as a reference gene. In Figure 14A, the ID copy numbers of the two populations are depicted over time. As expected, both populations displayed high *intersample* variances for ID frequencies. The HSC-specific ID sequence was only detected in high copy numbers (>10) in five samples, two at 8w, one at 16w, and two at 26w post transplantation. At the other indicated time points, ID-HSC levels were close to the detection limit. This result is consistent with the low abundance of HSC-derived clones retrieved from NGS data (Figure 11A-D). With respect to ID-MPP, several samples from 3w and 8w resulted in high copy numbers. However, at 26w post transplantation, only the samples of two animals showed high ID copy numbers for the respective cell population, creating a high variance in data distribution. No correlation of ID copy numbers to VCN or chimerism rates was found. It should be noted that the samples were

not consecutive samples of one animal, but every data point depicts a different animal at a different time point.

In order to validate the correlation between NGS and ddPCR, the data from each individual method were plotted and Spearman's rank correlation test was performed (Figure 14B). While MPP data showed no correlation between NGS and ddPCR samples, HSC data showed a positive correlation of r=0.63 (p=0.0006).

5.1.6 Influence of promoter strength on BC32 data interpretation

It was mentioned before that the initial transduction rate of marked HSPC populations presumably has a high impact on interpreting clonal tracking data and extrapolating the results to the whole organism. Therefore, we reasoned that elevating initial transduction rates would result in higher signals in the cellular and clonal output and thereby support interpretation of the results. In fact, the generated data from the BM-TBI group served as a comparative base for a publication[255], shedding light on age-related differences in clonal output after HSCT in the abovementioned HSPC subpopulations. More precisely, three groups were compared: HSPCs harvested from young donor animals transplanted into young recipients (data of BM-TBI, here termed Y-Y), HSPCs harvested from young donor animals transplanted into old recipients (Y-O), and HSPCs harvested from old donor animals transplanted into old recipients (O-O). Experimental conditions were designed as delineated above for BM-TBI experimental setting. All procedures were performed accordingly. Designated end point for these studies was 16w post transplantation, where chimerism rates of 68 % (Y-Y), 44 % (Y-O) and 56 % (O-O) in BM were detected in the three groups[255]. More importantly, transduction rates of the HSPC populations in the two additional groups (Y-O and O-O) exceeded the actual target limit and reached levels of approximately 50 %. Notwithstanding the likelihood of multiple integrations per cell, the experiments represented a powerful possibility to compare output data based on the differences in transduction rates. Frequencies of fluorescent marker expressing cells were measured by FACS analysis and comparative data of all three groups is shown in Figure 15A-C. As already described above (Figure 10) in the Y-Y (BM-TBI) group frequencies of fluorescent marker expressing cells in spleens ranged from 0.0—1.6 % (HSC-derived cells) and from 0.8—7.6 % (MPP-derived cells) (Figure 15A). In both groups for which higher transduction rates were described (Y-O and O-O) value ranges were 1.1—4 .5 % and 0.7—5.1 %, respectively (HSC-derived cells), as well as 0.7—5.1 % and 2.5—9.0 %, respectively, (MPP-derived cells) (Figure 15B and Figure 15C). Except for HSC-derived cells, which resulted in a slightly higher output in FACS analysis in the Y-O and O-O groups, no general differences between the groups were found. All groups showed a higher abundance of MPP-derived cells compared to HSC-derived cells and showed high *interindividual* variations in the data.

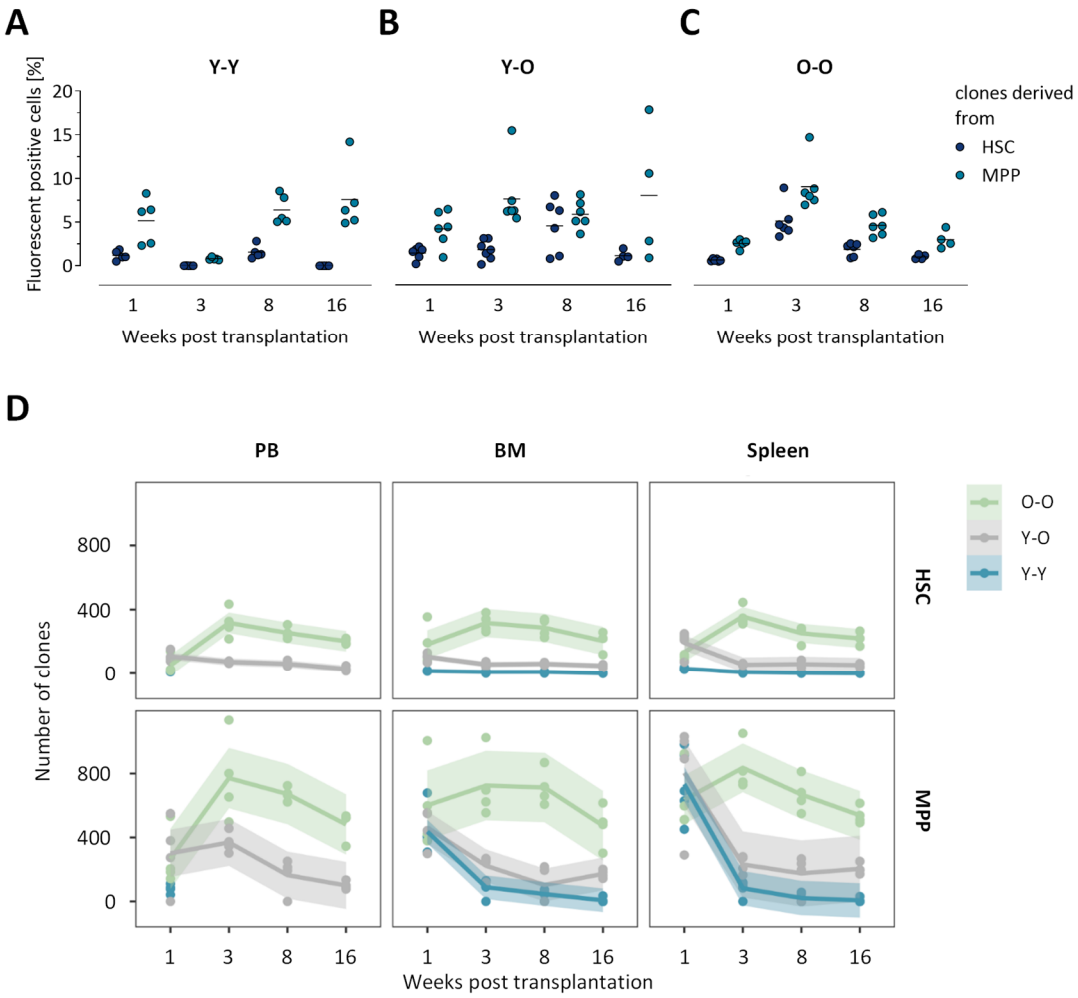


Figure 15: HSPC contribution in mice experiments with higher initial transduction rates. A)-C) Three different murine HSCT experiments (Y-Y, Y-O, and O-O) with marked cell populations were used to compare the influence of higher transduction rates on barcode abundance and fluorescent marker expression over time for up to 16w post transplantation. Contribution of HSPC-derived cells was measured by the fluorescent marker expression in mature progeny of the initially marked cells (n=4-6). Mean values are depicted by horizontal lines. D) Mean barcode content in mature compartments. DNA was extracted from PB, BM, and spleens and barcode sequences were amplified for NGS analyses. The number of unique barcode clones from HSCs and MPPs is shown. For reasons of representation of reconstitution dynamics, polynomial regression curves were fitted to the data, including their 95% confidence intervals (shaded areas) (n=4-6).
BM: bone marrow, HSC: hematopoietic stem cell, HSCT: hematopoietic stem cell transplantation, HSPC: hematopoietic stem and progenitor cells, MPP: multipotent progenitor, NGS: next-generation sequencing, PB: peripheral blood

Next, NGS data representing the number of HSC- and MPP-derived clones in blood, BM, and spleens were compared over time (Figure 15D). Overall clone numbers were higher (absolute data not shown) in particular in the O-O group compared to Y-Y with less initial transduction rates. This observation was not confirmed by the second group Y-O, which showed comparable results in comparison to the Y-Y group. As shown by flow cytometry data before, HSC-derived clones were, in general, less abundant than MPP-derived clones in all groups (Figure 15D). In particular, HSC-derived clones in the O-O group showed a more discrete pattern of HSC kinetics. Although signal intensities were slightly lower in the Y-Y group due to lower marking, patterns of MPP-derived clonal output were highly comparable in the Y-Y and Y-O groups. Despite higher clone numbers, in order to analyze the kinetics in different animals,

discrepancies in VCNs were taken into consideration. In both groups with higher transduction rates (O-O and Y-O), analysis of the sequence chromatograms revealed VCNs of two (O-O) or three (Y-O), respectively[255].

Taken together, the barcoding data of cellular tracking experiments after stem cell transplantation presented here showed that the BC32 system represents a powerful technique to label four HSPC subpopulations and track them simultaneously for up to 26w post transplantation. In all animals, a highly polyclonal development of BM, spleen, thymus, and PB post transplantation of BMSCs into lethally irradiated animals was detected. Short-lived progenitor populations showed less abundance in FACS and clonal NGS analyses, while HSC and especially MPP contribution was prominent in the indicated time points. Numbers of contributing clones decreased over time in both populations, indicating a stable but lower amount of actually contributing clones in long-term reconstitution. Due to premature termination of other groups (engraftment or pre-conditioning failure), no comparative analyses of clonal behavior was executed between different pre-conditioning regimens or stem cell sources.

5.2 CRISPR/Cas9 barcoding

5.2.1 Experimental design

In barcode tracking studies, overcoming the random lentiviral integration and the limitation of initial transduction rates is of high interest in order to increase the fraction of contributing cells included in the analysis. As explained in the introduction, by using a CRISPR/Cas9-mediated knock-in strategy of the barcodes, a genetically neutral barcoding approach outlined and implemented. The experimental part was designed in two different ways: co-transduction of wild type (WT) target cells with target and donor LVs (Figure 16A left panel) or alternatively using SpCas9-expressing cells, omitting the need for co-transduction (Figure 16A right panel). Generation of an efficient CRISPR/Cas9-mediated knock-in of barcodes into the genome of target cells relies on several interacting components and requires careful evaluation of each. Specifically, preliminary work included (i) analysis of possible gRNAs and subsequent cloning into the SpCas9-encoding target vectors, (ii) validation of on- and off-target effects of the respective SpCas9/gRNA complex, (iii) generation of barcoded donor vectors, and (iv) IDLV production for the delivery of the components. Target and donor vectors were designed expressing either mNeon or eBFP fluorescent marker under the control of an endogenous promoter (elongation factor 1α short (EFS) or SFFV). Barcoded donor vectors additionally contained left and right homology arms (LHA and RHA) of approximately 1 kb length, which support HDR-directed integration into the GSH locus (Figure 16B).

Since IDLVs lack random integration capacity of LVs, genetic information is delivered transiently. Thus, persistent fluorescent marker expression in NIH/3T3 target cells post transduction measured via FACS was considered to convey targeted knock-in of donor DNA to the expected locus. Finally, knock-in events were verified via conventional PCR or ddPCR.

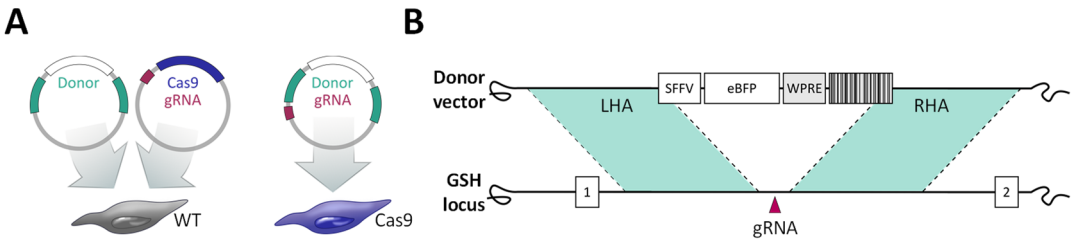


Figure 16: Schematic overview of the CRISPR/Cas9-mediated knock-in strategy. A) For CRISPR/Cas9-mediated knock-in experiments, two strategies were used to deliver the CRISPR/Cas9 components to NIH/3T3 target cells. For WT cells, co-transduction with target and donor constructs were performed. For Cas9-expressing target cells, a single construct on the base of the donor was used, additionally coding for the gRNA sequence. B) Knock-in strategy using a CRISPR/Cas9 donor vector. Design of the barcoded donor vector (above) containing an eBFP fluorescent marker under the control of an SFFV promoter and the barcode library. HDR-dependent barcode knock-in to the GSH locus is supported by the two homologous regions in the vector (LHA and RHA) flanking the SpCas9 cut site (red triangle) in the genome. Exons are depicted by white, numbered boxes in the genome.
Cas: CRISPR-associated system, CRISPR: clustered regularly interspersed short palindromic repeats, eBFP: blue fluorescent protein, HDR: homology-directed repair, gRNA: guide RNA, GSH: genomic safe harbor, LHA: left homology arm, RHA: right homology arm, SFFV: spleen focus-forming virus, WPRE: Woodchuck hepatitis virus posttranscriptional regulatory element, WT: wild type

5.2.2 Target efficiency prediction of gRNAs

For targeted integration into GSHs, the murine *Gt(ROSA26)Sor* (*Rosa26*) and hypoxanthine-guanine phosphoribosyl transferase (*Hprt*) loci were chosen, and two gRNAs per locus were designed based on their location next to a 3' PAM region in the genome. As for the *Rosa26* locus a variety of well-described gRNAs was known from the literature, two out of which were used for further analysis. Online prediction tools (CCTop[256] and CRISPOR[257]) were used to predict efficiencies of gRNA modifications and to determine possible off-target sites. As a cut-off value for off-target site prediction, three mismatches between off-target site and gRNA were tolerated. In Table 1, the calculated cutting frequency determination (CFD) specificity scores are listed for each of the four gRNAs (ROSA1, ROSA2, HPRT1, and HPRT2). High specificities corresponded to low amount of off-target sites, thus ROSA1 showed the lowest specificity score with 31 possible off-targets. Off-target prediction between the two algorithms differed slightly and CRISPOR algorithm resulted in more predicted off-targets (except for ROSA2). Predicted modifications including any type of indels[258] ranged from 43 % to 90 % independent of the corresponding specificity scores. Except for HPRT2, none of the gRNAs showed off-target predictions in exons of coding regions.

Table 1: Predicted efficiencies of gRNAs. Prediction of gRNA efficiencies was carried out with available algorithms, based on CFD specificity score and Fusi-Score[259,260]. Predicted modifications included any type of indels in the target locus[258]. Off-target predictions were performed using CRISPOR and CCTop algorithms with a cut-off of three mismatches between gRNA and off-target sequence.

Name	Specificity score	Predicted modification	Predicted efficiency	Off-targets for 0 – 1 – 2 – 3 mismatches	Off-targets (CCTop)
ROSA1	60	43	44	0 – 0 – 1 – 30	23 (no exonic)
ROSA2	91	81	58	0 – 0 – 0 – 04	4 (no exonic)
HPRT1	88	90	60	0 – 0 – 1 – 09	6 (no exonic)
HPRT2	87	82	64	0 – 0 – 4 – 07	5 (1 exonic)

5.2.3 IDLV delivery of CRISPR/Cas9 components

Chosen gRNAs were used for subsequent target construct generation. Donor constructs were generated by amplifying 800—1000 kb regions adjacent to the gRNA recognition sites in the murine genome and cloned into the well-established BC32 vector (see Figure 16B). In case of close proximity of two gRNA sequences (ROSA1 and ROSA2), only one donor construct was generated and used for both target sites. In order to create a suitable delivery vehicle for CRISPR/Cas9 both constructs, IDLVs were generated. After production of ILVs and IDLVs, titers were determined by infection of the producer cells line (HEK293T) and the number of infectious particles per mL was calculated. Produced IDLV supernatants showed remarkably reduced viral titers compared to their integrating counterparts (Figure 17A, see also Table 26, Appendix C). Similar trends were observed for all constructs, independent of locus, except for one outlier production (HPRT2 donor ILV). Additionally, mean titer values of ILVs (2.15×10^7 IU/ mL), as well as IDLVs (6.88×10^5 IU/ mL) produced with target constructs were

lower, when compared to donor constructs (ILV 2.89×10^9 and IDLV 1.7×10^7 IU/ mL). All calculated values are listed additionally in Table 26 (Appendix C). Titers below 1×10^5 IU/ mL were excluded from calculations due to inefficient virus production and supernatants were discarded.

Next, to prove neutral integration behavior, transient abundance of IDLVs was performed in a time course by FACS analysis. Therefore, target cells were transduced with ILV or IDLV constructs. Fluorescent marker expression was measured by flow cytometry analysis at 6h, 24h, 48h, 72h, and 144h (6d) post transduction (Figure 17). After transduction with ILVs, FP^+ cells steadily increased up until 48 h post transduction, where a stable abundance of FP^+ cells was reached. Similar patterns were observed for cells transduced with IDLVs: a steady increment of FP^+ cells was detected until 48 h post transduction. As expected, due to the transient IDLV abundance, after this time point the amount of FP-expressing cells decreased to a basal level until 6d post transduction.

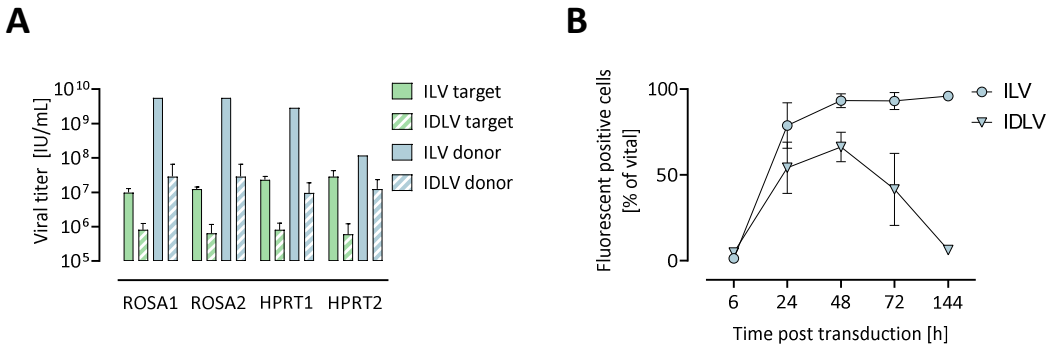


Figure 17: Comparison of ILVs and IDLVs. A) Cells were transduced with viral supernatant and 72h post transduction the amount of fluorescent positive cells was measured for titer calculation. Bar plots of viral titers (mean ± SD) of all 14 constructs, either produced as ILV or as IDLV are shown. Green bars (filled or dashed) depict mNeon target constructs and blue bars (filled or dashed) show eBFP donor constructs (n=1-14). Donor constructs of ROSA1 and ROSA2 show corresponding values since only one donor construct was used for both gRNAs. Viral titers show a strong reduction of titer when IDLVs were used, independently of the construct. B) Time course of ILV and IDLV transduced target cells measured by FACS analysis (mean ± SD). Persistence of ILVs is shown by stable expression of fluorescent marker, while IDLV transduced cells show a transient expression and slow removal of the vector. Measurement was performed at 6h, 24h, 48h, 72h, and 144h (6d) post transduction (n=2-6).
eBFP: blue fluorescent protein, IDLV: integrase-deficient lentiviral vector, ILV: integrating lentiviral vector

5.2.4 SpCas9 target specificity

In order to examine the SpCas9-induced modification at the chosen loci, NIH/3T3 cells were transduced with equal MOIs and $mNeon^+$ cells were sorted 48 h post transduction. On day seven, DNA was extracted from transduced cells and non-transduced control cell bulk and used for PCR amplification of on-target and off-target sites. From the predicted off-targets, each three top hits of the cutting frequency determination (CFD) off-target score were chosen for detailed analysis. Off-target site sequence amplification was performed accordingly. Sequence chromatogram data of ILV and IDLV transduced samples was taken for *tracking indels by decomposition* (TIDE) analysis and compared to control WT cells. Figure 17A shows

two representative Sanger sequence chromatograms of control and CRISPR/Cas9 edited cells. The target site is depicted by a grey square, after which the edited sequence abruptly started to show aberrant sequence signal, indicating heterogeneous CRISPR/Cas9 modifications in the cell bulk. Nine days post transduction CRISPR/Cas9 induced genetic modifications at on-target sites were analyzed by TIDE tool. Transduction with a stably persisting target ILV yielded in overall slightly higher efficiency rates than transduction with target IDLVs (Figure 17B). Mean efficiency rates from ILV cell bulk of 80.3 % (ROSA1), 73.8 % (ROSA2), 97.8 % (HPRT1) and 93.3 % (HPRT2) were detected. With target IDLVs showing greater variance in data, mean efficiency rates were 59.1 % (ROSA1), 26.5 % (ROSA2), 46.7 % (HPRT1) and 85 % (HPRT2). At *Rosa26* locus in general IDLV productions resulted in highly reduced efficiency rates (26.4 % reduction for ROSA1 and 64.1 % reduction for ROSA2). At *Hprt* locus, efficiencies resulting from IDLV productions were similar to those generated with ILV vector productions at one of the analyzed sites (52.2 % reduction for HPRT1 and 8.9 % reduction for HPRT2). Note that ILV data needs careful evaluation, since efficiency rates were only calculated from n=1 sample for three of the four indicated loci. Calculated R2 values from TIDE algorithm were >0.9 in all samples. Except for ROSA1, with a predicted modification of 43 %, these results matched with their predicted values in the efficiencies of genetic modification.

Next, samples were processed for amplification of three predicted off-target sites. After sequencing of the amplicons, Sanger sequence chromatograms were used with TIDE algorithm to assess total CRISPR/Cas9 efficiency scores. All off-target site efficiencies were expectedly low (Figure 17C). While at each off-target site analyses revealed low efficiencies in the range of 0.8—3.1 %, only one off-target site of ROSA1 slightly differed from this range, with a total efficiency of 7.5 %.

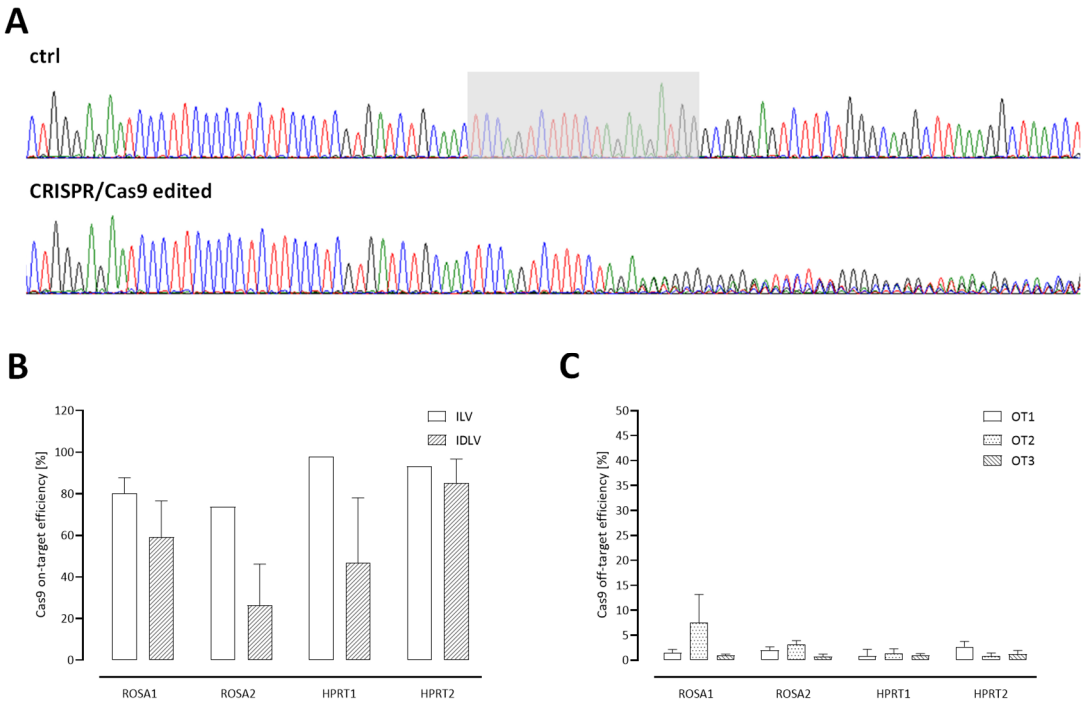


Figure 18: Total efficiency and specificity of Cas9 to induce DSBs at different loci. A) Sanger sequence chromatograms of *Rosa26* locus after CRISPR/Cas9 modification. DNA of control cells and cells transduced with CRISPR/Cas9 target ILV was used for amplification of the target site in the genome. The chromatogram below shows aberrant sequence signals 3‘ to the DSB induction site, implying a mixture of different sequences in the bulk population after CRISPR/Cas9 modification. Recognition sequence of gRNA and PAM site (5‘-TGG-3‘) are underlined in the control sequence. B) TIDE analysis of NIH/3T3 cell bulk of amplified CRISPR/Cas9 target sequences 9d after transduction with integrating (white bars) or non-integrating (dashed bars) target vectors (mean ± SD). The total amount of aberrant sequences as a measure for SpCas9 efficiency is shown for the respective gRNAs (ROSA1, ROSA2, HPRT1, and HPRT2) (n=1-7). C) TIDE analysis of predicted CRISPR/Cas9 off-target sites of NIH/3T3 cell bulk 7d after transduction (mean ± SD). For each gRNA three predicted off-target sites (OT1-OT3) were chosen for detailed analysis (n=4-8).
DSB: double strand breaks, Cas: CRISPR-associated system, CRISPR: clustered regularly interspersed short palindromic repeats, gRNA: guide RNA, IDLV: integrase-deficient lentiviral vector, ILV: integrating lentiviral vector, OT: off-target, TIDE: tracking of indels by decomposition, PAM: protospacer adjacent motif

5.2.5 CRISPR/Cas9 barcoding efficiency

After successful validation of the CRISPR/Cas9 components, co-delivery of target and donor constructs was performed for targeted barcode knock-in. To deliver the vectors simultaneously, the respective IDLVs were used to co-transduce a cell bulk. Three days post transduction fluorescent marker expressing cells were sorted for subsequent experiments and sorted bulk population was used for FACS analysis 21d post sort (Figure 19A, white bars). To additionally include a delivery method independent of LV, chemical transfection was used (Figure 19A, grey bars). First, overall transduction/ transfection rates were determined by flow cytometry analysis based on the fluorescent marker expression. IDLV transduction resulted in 9.1 % transduced cells and chemical transfection using naked plasmid DNA resulted in 28.8 % transfected cells (data not shown). Putative CRISPR/Cas9-mediated knock-in at target loci were identified based on stable eBFP expression at 21d post sort. Transduction revealed knock-in rates of 1.8 and 1.5 % (ROSA1 and ROSA2), respectively, as well as 5.8 % and 6.4 % (HPRT1 and HPRT2), respectively. Delivery of the components via transient transfection yielded in slightly higher knock-in levels between 4.3 % (HPRT2) and 7.7 % (HPRT1). Despite

high variance in data, only at HPRT2 locus delivery via transfection did not exceed delivery via transduction.

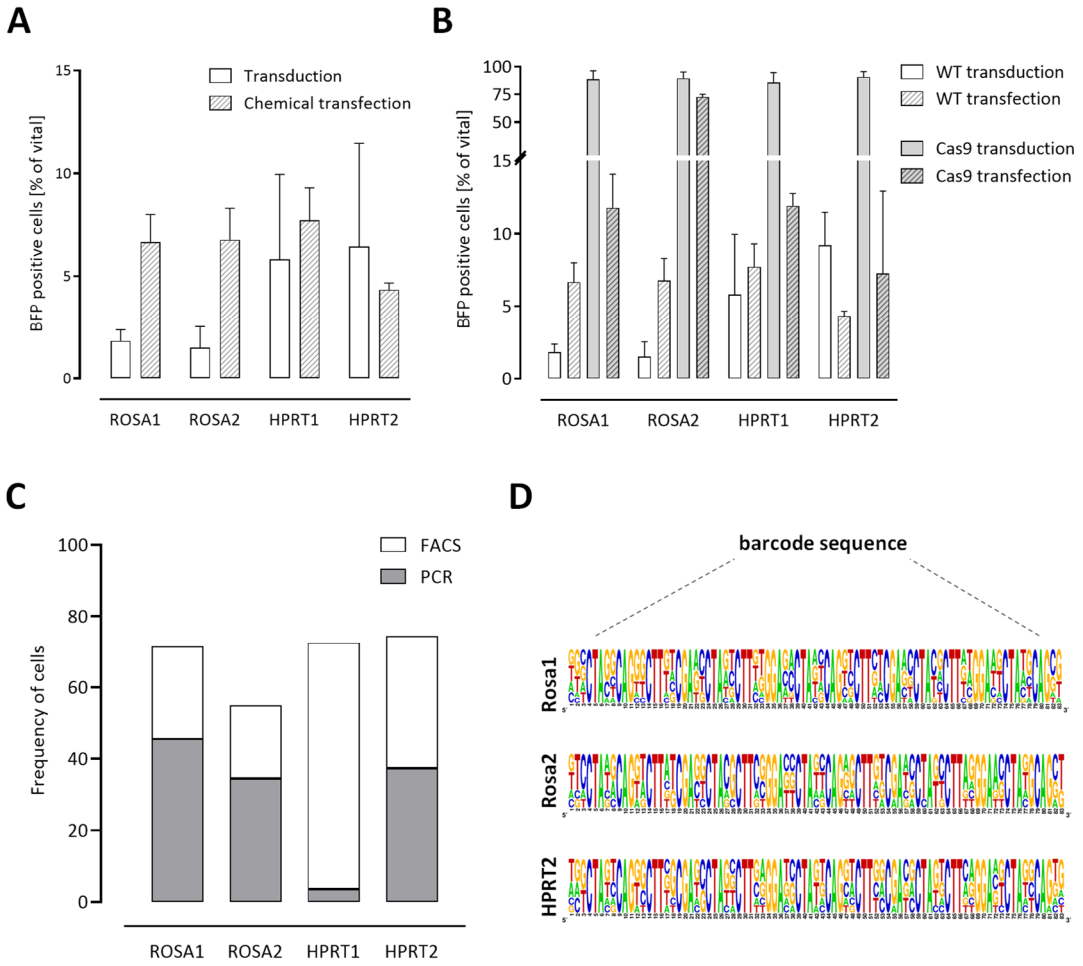


Figure 19: Knock-in efficiency in bulk and single cells based on the delivery method. Target and donor plasmids were delivered to NIH/3T3 target cells via IDLV transduction and chemical transfection. A) Fluorescent marker expression was measured by FACS analysis and stable abundance of fluorescence protein 21d post-delivery was considered to reflect a knock-in. For each locus (*Rosa26* and *HPRT*) two different gRNAs were used and are shown in the graph (mean ± SD, n=3). B) Cas9 expressing cells were generated by lentiviral transduction of NIH/3T3 cells. gRNA-Donor plasmids were delivered via IDLV transduction or chemical transfection and fluorescent reporter protein expression was measured 21d post-delivery (mean ± SD, n=3). C) Single NIH/3T3 target cells were sorted after co-transduction and eBFP expression was measured 28d post transduction (white bars). eBFP^{+} single cell clones were used for subsequent PCR validation (grey bars). PCR results were considered to reflect successful knock-in in the analyzed cell clones. D) Seqlogos created from the Sanger sequences of single cell barcode amplification. Results of Rosa1, Rosa2 and HPRT2 are shown, since at HPRT1 locus only one BC could be recovered. Barcode backbones can be distinguished from randomly inserted nucleotides by the presence of only one base in the respective position in the frequency plot.Cas: CRISPR-associated system, eBFP: blue fluorescent protein, FACS: fluorescence-activated cell sorting, gRNA: guide RNA, IDLV: integrase-deficient lentiviral vector, PCR: polymerase chain reaction

In order to overcome shortcomings in SpCas9 delivery due to its size, an alternative system was used for the targeted barcoding approach. Besides the conventional co-transduction for delivery of CRISPR/Cas9 components, the abovementioned system was validated in SpCas9-expressing target cells to reduce the plasmid amount to be delivered. For a stable SpCas9 expression, NIH/3T3 target cells were transduced with SpCas9-encoding target ILV (w/o gRNA) and three days post transduction single-cell clones were sorted and expanded

(NIH/3T3^{SpCas9}). Since off-target analysis of SpCas9-expressing cells was performed as described above, NIH/3T3^{SpCas9} cells were used for subsequent knock-in experiments. To this end, donor vectors were further modified by means of cloning an U6-sgRNA cassette 5' of the LHA of donors. Cells were transduced with gRNA-donor IDLVs and analyzed by flow cytometry analysis at 21d post sort (Figure 19B). When comparing the knock-in data of WT cells to NIH/3T3^{SpCas9} cells, higher knock-in rates were achieved when SpCas9 was stably expressed in the cells. Transduction of SpCas9-expressing cells resulted in knock-in events in 85.9—90.7 % of the cells, independent of the selected target site. Thereby, knock-in events occurred 49-, 60-, 15- and 10-fold (ROSA1, ROSA2, HPRT1 and HPRT2) more frequent compared to co-transduction of WT cells. Transfection of NIH/3T3^{SpCas9} cells resulted in similar output compared to the transfected WT cells, with the exception of ROSA2 target site, where 11-fold more knock-in events were detected (72.6 %).

To further quantify and analyze the barcoding in target cells, single-cells were sorted from cell bulks after IDLV co-transduction and expanded for 21 to 28 days. In total, 49 (ROSA1), 40 (ROSA2), 51 (HPRT1) and 55 (HPRT2) clones were used and examined based on their fluorescent marker expression. After 28d, 71.4 % (ROSA1), 55 % (ROSA2), 72.5 % (HPRT1) and 74.5 % (HPRT2) of single-cell clones showed stable eBFP expression (light grey bars), measured by FACS anaylsis (Figure 19C, white bars). Of those, subsequent barcode analysis via construct/ genome-specific PCR primers was additionally performed, to validate site-specific knock-in events. Figure 19C shows the PCR analysis of the eBFP^{+} clones (grey bars). In total, 45.5 % (ROSA1), 34.4 % (ROSA2), 3.6 % (HPRT1), and 37.3 % (HPRT2) were confirmed by knock-in specific PCR. Except for the outlier in HPRT1 PCR data, fluorescent marker expression and PCR-validated knock-in data resulted in similar knock-in efficiencies of the barcodes in all loci. Clones showing stable eBFP expression but lacking signals by PCR analysis were considered to reflect residual integration events from IDLVs. Comparing the two methods for knock-in analysis, PCR-based method resulted in 0.3-fold (0.5- or 0.9-fold) reduced detectable knock-in events in eBFP^{+} clones showing higher specificity and reliability of PCR data.

Furthermore the expanded single-cell clones, with PCR-verified knock-in events were used for subsequent barcode sequencing (ROSA1: 14 clones, ROSA2: 10 clones, HPRT1: 1 clone, HPRT2: 13 clones). This was performed in order to prove retrieval of barcodes and investigate the frequency of barcode knock-in. Sanger sequencing revealed both, mono or bi-allelic knock-in of barcodes, as analyzed by the number of base calls in each chromatogram at the random positions of the barcode sequence. Representative barcode sequences after mono-allelic knock-in events served for the generation of sequence logos (Weblogo[261]), showing the preserved barcode backbone structure in the samples after barcode retrieval (Figure 19D).

In conclusion a comprehensive study of two different GSH sites for neutral integration of barcodes was performed. The choice of different gRNAs showed no influence on the barcode knock-in efficiencies in the current work, showing that the exact position in the GSH loci is negligible for experimental design. Thus, initial off-target site prediction can be given more weight in order to reduce possible side effects after CRISPR/Cas9 modification. HDR

efficiencies were generally low in *in vitro* experiments using NIH/3T3 cell line, but could be enhanced drastically when using a generated SpCas9-expressing cell line with the same background. IDLVs represented a powerful delivery platform for CRISPR/Cas9-mediated knock-in approaches, as comparison to chemical transfection resulted only in a slight increase of knock-in rates. Finally, barcodes were retrieved from single-cell clones after gene editing, confirming the adaptability of the CRISPR/Cas9 system for genetic barcoding and the persistence of mono- and bi-allelic knock-in events over time.

6 Discussion

6.1 In vivo reconstitution dynamics

The unique ability of the hematopoietic system to engraft in a recipient organism following HSCT is nowadays used as a standard treatment worldwide for a myriad of malignant and non-malignant diseases. HSCT has become an indispensable tool to replenish the blood-forming system in order to cure formerly challenging genetic defects in pediatric patients or as an autologous rescue option of the patients own hematopoietic system after exposure to high-dose radiochemotherapy. Despite outstanding success in treatment of the indicated malignancies, the predictability of the outcome post HSCT and the underlying cellular processes of blood reconstitution are still not completely resolved. A certain number of patients die of adverse events arising in the aplastic phase between myeloablation and eventual engraftment of the transplanted cells[80,81]. In the human system, the composition and quality of the graft are commonly assessed by the surface protein CD34 on HSCs. Interestingly, this marker is not uniquely expressed on HSCs exclusively, but also on a variety of downstream progenitor populations[262]. This raised the question as to which extent other HSPC populations besides HSCs contribute to hematopoietic reconstitution in the short- and long-term period after HSCT. This knowledge could then hypothetically lead to an amelioration of the recovery process. Here, four independent murine HSCT experiments were designed based on the donor stem cell source and the recipient pre-conditioning regimen. Prior to transplantation, HSPC populations were marked with a lentiviral barcode library (BC32[156]) in order to track the differences of spatial dynamics and clonal kinetics of engraftment.

Pretreatment of a recipient is mandatory prior to HSCT in order to prepare the microenvironment for subsequent homing and engraftment of donor-derived HSCs. However, this significantly perturbs the BM niche and influences the hematopoietic reconstitution, aggravating the comparison to the steady-state situation[263]. The here discussed experiments were specifically designed to investigate the hematopoietic recovery in perturbed niches, which needs to be mentioned.

6.1.1 HSPC frequencies in the stem cell grafts

In the four experimental set-ups (BM-TBI, BM-TBI^{SFFV}, BM-Bu, and PB-TBI), high numbers of HSPCs were harvested from BM or PB, independent of the stem cell source. However,

proportions of HSPC populations in the different grafts varied substantially (Figure 6C and Figure 6D). While for BM-TBI a relative amount of 20 % HSCs/ MPPs per graft was calculated, comparable experiments, where BM was used as the respective stem cell source, resulted in lower relative amounts of 12 % and 7 % of HSCs/ MPPs per graft, respectively. Since these observations can hardly be explained by the experimental set-up (identical settings, except for the utilization of a different promoter in the BC32 vectors), variances in HSPC counts seem to show up due to technical procedures regarding staining, preparation, and storage of the cells. When mobilizing HSPCs to the blood, as performed in the clinical practice, only a relatively low amount of 1 % HSCs/ MPPs per graft was recovered by this procedure. In the literature, diverging frequencies of Lin^- in BM were reported, with ratios from 1.25 %[264] to 5—8 %[26,265] of which 0.08 %[264] to 0.15—5 %[26,265] belonged to the LSK population. In the murine transplantation experiments performed here, Lin^- in BM cells ranged from 1—7 %, of which 0.3—1.2 % were subsequently characterized as the LSK cell population (here HSCs and MPPs). One exception was found in the group that used mobilized PBSCs as the stem cell source (PB-TBI): in total 24 % of PB cells were defined as Lin^-, of which 1 % belonged to the LSK population. As a consequence of the different harvesting methods, the experiments presented here used highly diverging cellular frequencies in the grafts. Specifically cell populations driving hematopoietic reconstitution, such as HSCs and MPPs were underrepresented in the PB-TBI group, accompanied by an elevated CLP frequency. As shown earlier, engraftment and complete recovery is conceivable after HSCT with a limiting dilution or even with single transplanted cells[15,34]. Thus, low relative cell numbers in grafts do not seem to play a critical role for engraftment in these studies. However, to which extent varying population frequencies affect reconstitution dynamics in order to shorten the aplasia period remains to be elucidated.

6.1.2 Drawbacks in achieving experimental baseline conditions

Two of the murine transplantation experiments were not performed according to the experimental design (PB-Bu and PB-TBI, respectively). The PB-TBI group using mobilized PBSCs was terminated prematurely due to intolerable weight loss and eventual graft failure of the animals. This was confirmed by the lack of chimerism in samples of transplanted animals as well as the lack of any fluorescence positive signals in FACS analysis (Figure 8F, Figure 9 and Figure 10). Even though mature cell populations in spleens recovered similarly as in animals with a successful engraftment (Figure 8A-E), survival of the residual animal cohorts did not extent beyond 16w post transplantation (data not shown). On the one hand, it was previously shown that the stem cell mobilization is highly dependent on the mouse strain and that the here used C57Bl/6 mouse strain are poor responders to the mobilizing effects of G-CSF[77,266,267]. However, this effect should be partially reversed by the additional use of the CXCR4 antagonist AMD3100, as reported before[77].

During the sorting of the mobilized PBSCs using identical panels and gating strategies as for the BMSCs, a shift in surface marker expression was noticed, which hampered the HSPC separation based on the previously established settings (data not shown). This phenotypic shift in FACS analysis could be possibly explained by the pretreatment with mobilizing agents

and the different environment these cells are exposed to, once they abandoned their original niche. Of note, it has been shown, that G-CSF mobilized cells suppress normal c-kit expression patterns[268-270], which is a fundamental surface marker used for the discrimination of different cell populations. In order to confirm the accuracy of the sorting strategy to discriminate the HSPC subpopulations, it would be reasonable to analyze the residual BM of the mobilized animals. After mobilization, an up to 90 % depletion of LT-HSCs can be expected[271]. Furthermore, *in vitro* colony assays can be used to describe the potential of the mobilized cells, showing rather the potential of the containing progenitor cells than HSCs[34]. This would have been of special interest to compare mobilization efficiencies between the different murine transplantation experiments. Due to limited cell numbers, these additional assays could not be performed. Finally, since engraftment of mobilization could not be confirmed during PB-TBI, a second group using PBSCs was omitted from the initial experimental study design.

Since irradiation causes severe damage to the hematopoietic system and full immune reconstitution in general takes longer than with other methods[272], in another set of experiments the influence of pre-conditioning regimen using busulfan was investigated. Other than the abovementioned graft failure in PB-TBI, in the BM-Bu group, no chimerism was observed in BM samples despite the quick recovery of the animals after the transplantation (Figure 8F). Inadequate chimerism rates in BM found in this experiment can be explained by insufficient pre-conditioning with the chemotherapy agent busulfan. This hypothesis was confirmed by reconstitution of blood cells in spleen samples, where cell counts at the beginning of the experiment were unexpectedly high, showing no complete depletion of blood and marrow cells in tissue samples (Figure 8A-E). However, after premature termination of the experiment, equal transplantation conditions were used in an additional experiment utilizing lentivirally transduced Lin^- cell instead of HSPC subpopulations (data not shown). In this experiment, BM chimerism levels of 84—86 % were detected in all animals at 8w post transplantation, suggesting no general influence of the busulfan pre-conditioning on the success of the transplant engraftment[273,274]. Despite that, successful engraftment with high chimerism levels after transplantation with Lin^- cells was also confirmed by others[275], no reports were found on transplanting low numbers of specific subsets of HSPCs after busulfan pre-conditioning, as presented here.

Two main aspects might have influenced the successful pre-conditioning and engraftment in the prematurely terminated group; the stability and the bioavailability of the drug. In aqueous media, busulfan shows a reduced stability due to hydrolysis[276] and precipitates can be found in the solution[277]. Stability data provided by Pierre Fabre Laboratories are 8 h at room temperature or 12 h at 2–8 °C in the reconstituted working solution of 0.55 mg/ mL. These storage conditions were marginally extended by a study of Houot et al. assessing the stability of busulfan in different storage containers[278]. Thus, in experimental conditions using busulfan special care needs to be taken in handling and storage of the drug. Another concern is the bioavailability, which for pre-conditioning prior to transplantation is a fundamental parameter to predict efficacy of BM ablation and possible toxicities. Busulfan bioavailability was reported

to vary substantially interindividually in adults and even more in pediatric patients[279], showing the urgent need for constant drug monitoring in PB and individual dose adjustment. In mice, daily blood tests during the five consecutive days of busulfan administration are not feasible due to limitations in blood sampling and blood volume. Besides, increasing the concentrations, which possibly leads to higher rates of BM ablation can cause toxic side effects like veno-occlusive disease[280]. Furthermore, in children a high rate of engraftment failure despite low toxicity was observed[281,282], which held also true for mice using repeated moderate doses of the drug[283]. However, others reported high levels of engraftment and minimal transplant-related mortality in murine transplantation studies[284-286]. Furthermore, the influence of busulfan on subsets of primitive stem and progenitor cells were intensively studied. While cytotoxic effects in quiescent hematopoietic progenitors[287,288] and reduction of the frequency of HSPCs clones contributing to hematopoiesis[289] were described, the effect of the drug on the clonality of donor subsets of HSPCs after HSCT is largely unknown and could not be resolved here. In conclusion, the experimental comparison between different stem cell sources and pre-conditioning regimens with regard to the clonal output during hematopoietic reconstitution could not be performed in detail according to the initial experimental design.

6.1.3 Lentiviral transduction of HSPC populations

The BC32 system applied here relies on the fundamental prerequisite of a single lentiviral integration per cell in order to discriminate between barcode harboring clones. Therefore, MOIs used for lentiviral transduction were kept at low levels to achieve ≤20 % transduction rates, in order to assume the integration of one barcode per cell[176,290,291]. Others reported even higher transduction rates (50 %) to yield >95 % of cells with a single integration[158]. However, depending on the experimental design no conclusions can be drawn from the transduction rates of a heterogeneous cell pool (i.e., LSK cells) to each HSPC subpopulation. Separate transduction of highly pure cell populations, as performed in the current experiments, is beneficial in order to ensure equal transduction rates. Of note, no conclusive bias in transduceability between cell populations was observed (Figure 6E). This could be of special interest for recent gene therapy approaches.

Since the method of barcoding is based on lentiviral integration, the extrapolation of the data to predict repopulation ability of HSPCs *in vivo* requires careful consideration. Prior to transplantation, cells were exposed to sorting, cell culture conditions, and lentiviral infection *ex vivo*, which required several hours (up to 15 h) of culture conditions. However, lentiviral transduction did not show genotoxic effects that may alter HSC function[292-294] or integration bias in any HSC subtype[295]. In contrast, others have reported higher engraftment potential of non-transduced human stem cells in comparison to cells transduced with a gamma retroviral vector[296]. Furthermore, lentivirally transduced lymphoid-primed multipotent progenitors (LMPPs) have also been reported to lose their differentiation capacity, hinting to an altered cell fate as a consequence of the transduction[176]. Taking single integrations and low amount of cells during barcoding procedures into account, these effects still seem unlikely. However, these results indicate that a possibly cell fate influencing integration to crucial genomic regions cannot be ruled out. Due to time constraints, these questions were not answered in

the current work and need to be investigated in further studies. In order to avoid the risk of these side effects, targeted barcoding to a GSH represents a promising alternative to unpredictable lentiviral integration.

6.1.4 Technical issues of cell recovery and barcoding procedure

Others reported a total clone number recovery of 60—80 clones[176], or up to 50 clones from the initial pool[158,160,163]. Although, clone numbers observed here were comparable to these studies, severe limitations to detect the progeny of the initially barcoded cells via FACS analysis were experienced (Figure 9 and Figure 10). These limitations can be explained by the relatively low cell count compared to the whole BM of the analyzed samples, thus requiring the processing of a great number of cells in order to detect enough low-abundant cells, diluted by the presence of other cells.

Fluorescent marking to track cells via FACS was performed with wavelength-shifted variants of GFP (GFP, eBFP, T-Sapphire, and Venus), in order to avoid possible cytotoxic effects of red variants of fluorescent proteins[297,298]. However, the isolation of specific cell populations often requires sophisticated panels of antibodies conjugated with fluorescent dyes that might interfere with the relatively dim fluorescent markers and thus, hamper the FACS analysis. Even though previous work focused on enhancing the brightness of fluorescent proteins, some promoters might not be sufficiently active to drive fluorescent protein accumulation, needed for low frequencies of cells[299,300].

As promoter activity was reported to have no impact on clonal diversity after HSCT[155,301], it was reasoned that the EFS promoter could be exchanged by the stronger SFFV promoter[302-304], to induce a more distinctive and brighter pattern in fluorescence positive cell distribution for FACS analyses. Thus, a second experiment (BM-TBI^{SFFV}) using a different promoter was performed. In this experiment, it was observed that the discrimination of FP^+ and FP^- cells did not differ compared to the BM-TBI group. Consequently, the different promoter did not positively affect FACS analysis.

To overcome the limitation of protein expression-dependent FACS analysis, a ddPCR assay, quantifying short ID tags in the initial construct sequences was introduced. Compared to other methods, ddPCR provides the highest sensitivity for quantification of rare genetic events at frequencies as low as 0.05—0.01 %[305-307]. Interestingly, only the HSC data obtained from ddPCR analysis correlated with barcode counts observed by NGS analysis. For MPP-derived ddPCR data, no correlation with NGS counts was observed, although barcode counts were above the ones obtained for HSC-derived cells (Figure 14A and Figure 14B).

The recently developed barcode system BC32 employed here has been under continuous optimization over the past years to overcome technical limitations, mostly involving PCR-induced errors[155,156,161]. This system was designed in a semi-random fashion, allowing for a balanced GC content during generation[156,308], and according to its defined backbone structure, is recognizable and assignable to a certain cellular population. This design has been used by others successfully in the past[136,160,309]. Assuming no skewing in the initial library, by working with a total amount of 32 variable positions in the barcodes, up to 1.8×10^{19} different

barcodes can be theoretically achieved in the libraries. However, in the current experiment generated libraries ranged from 6.4×10^6 to 5.4×10^7. Technical steps during barcode generation like the amount of starting material or the subsequent purification steps may account for this decrease in library complexity. Of note, barcode complexity was deduced from the number of colonies after bacterial transformation of the library[309], though this method is only an estimate of the real amount present.

Using highly complex libraries, exceeding the total number of HSPCs to be barcoded, avoids the repeated use of single barcodes which leads to incorrect lineage relationships[176]. Barcodes with low read counts (≥3) were also included to prevent the exclusion of low-abundant clones, unlike previous reports[160,173,310]. Besides, filtering barcode data solely based on the read threshold was shown to be insufficient[308,311] and consequently, other methods should be reconsidered. Nonetheless, filtering remains a crucial issue in deep sequencing data, especially since in the current system the input reference library, needed to remove false negative barcodes, was unknown. A potential way to circumvent this issue is using fully annotated libraries instead of semi-randomly synthesized libraries[159,312,313] at the expense of smaller library sizes and time-consuming library preparation. Thereby a so-called whitelist of barcodes actually present in the library is created to identify the likely true barcodes and remove false positive results. Currently another technique to reduce PCR amplification bias has been described: labeling of PCR templates prior to amplification with short DNA molecules, called unique molecular identifiers (UMIs). As a result, unique fragment-UMI combinations are created prior to amplification and amplification bias as well as errors can be resolved by this to identify PCR clones or real biological duplicates[171,174,314,315].

A low amount of barcodes was retrieved from final samples in murine tissue after stem cell transplantation, especially at later time points. Barcode count comparison with those of high transduction rate experiments did not reveal any systemic bias (Figure 15D). Moreover, it should be noted that all experiments yielded similar chimerism levels, which makes the low cell count unlikely to be a result of insufficient cell marking efficiency or poor engraftment. However, when comparing the NGS data to the same experiments as mentioned above an unusually lower NGS read quality in barcode samples was observed, regardless of the flowcell run (data not shown). Most of the low quality read counts were then assigned to miscellaneous sequence reads, showing the appearance of non-barcode carrying reads. Thus, this might have affected the sequencing depth, though not the analyses of barcoded backbones. The classification as non-barcode carrying reads possibly arose due to presence of imperfect sequences (during barcode generation or PCR amplification), which could not be allocated to one of the BC32 backbones by the analysis pipeline.

6.1.5 Clonal contribution to hematopoietic recovery

The initial transduction rates, which could only be determined for two murine transplantation experiments, were in accordance with a calculated mean VCN of ≈0.3 in mature BM cells in the BM-TBI group (Figure 8G). Taking the rates of transduced HSPC populations into account the animals received a maximum of 784 (BM-Bu) or 1648 (BM-TBI) marked donor HSCs and

1290 (BM-Bu) and 5830 (BM-TBI) marked donor MPPs. Since transduction rates could not be detected in the BM-BM-TBI^{SFFV} and PB-TBI groups, they were not considered for calculations. Although, the LSK cell pool (here HSCs and MPPs) is enriched for cells with hematopoietic reconstitution activity, it does not actually reflect the total engraftment potential of the cells. Interestingly, out of the LSK cell population only 10 % of them were described to be *bona fide* LT-HSCs[21,23]. Taking the frequencies of transplanted cells mentioned above into account this would result in a total frequency of 748 LT-HSCs in the BM-TBI group capable of engrafting and contributing to the clonal output. Furthermore, the loss of stem cell frequency caused by *ex vivo* stem cell manipulation and culturing needs to be taken into consideration. As reported by Bystrykh and colleagues, 1 to 2-log losses were observed in retrovirally transduced BM cells. After transplanting 1 – 5×10^6 transduced BM cells, they retrieved 4—13 active stem cell clones per animal in their experiments[136]. Comparing these numbers with the ones obtained by the NGS pipeline used here, mean clone numbers of 21, 15, 13, 7, and 14 (1w, 3w, 8w, 16w, and 26w, Table 25, Appendix A) represent a proportion of 9—28 % of the LT-HSCs with engraftment potential. Previous reports communicated homing frequencies of single transplanted HSCs of 10—20 %[316-318]. Thus, the data presented here is in line with the observed numbers, with only small discrepancies in the frequency values. Taken together, since barcoding captures a certain fraction of marked cells with engraftment potential, the generated data needs to be carefully extrapolated to draw conclusions to the complex hematopoietic system of an organism.

Comprehensive analysis of the sum of barcodes derived from a certain initially marked HSPC subpopulation gives insights about the contribution of the respective population to blood reconstitution over time in absolute numbers. Every barcode represents one clone and regardless of the size of the clone, conclusions can be drawn on how many of the clones give rise to mature cell populations in BM, PB, spleen and thymus (Figure 11A-D and Table 25, Appendix A). The BM is known to be the primary site of hematopoiesis, but additionally extramedullary hematopoiesis occurs in spleens in response to stress conditions like disease, myeloablation, and destruction of the BM niche[319]. Furthermore, the lymphatic organs spleen and thymus play important roles in the maturation of lymphocytes, the output of which can be evaluated in the PB of the animals post transplantation to elucidate temporal as well as spatial dynamics of engraftment and reconstitution. In murine models, it has already been shown that over one third of HSCs seed and repopulate in spleens directly on transplantation[320]. In line with this, and given the fast immune reconstitution of two weeks observed in murine models, MPPs have been hypothesized to provide early splenic reconstitution[43].

The observations in clonal contribution after HSCT were similar for BM, spleen and thymus, while contribution in PB differed slightly. In general, HSC clone numbers were found to be below MPP clone numbers and the findings were confirmed by comparative analysis to former experiments investigating the effect of age on hematopoietic reconstitution[255]. The higher amount of marked MPPs in perturbed hematopoiesis might also be explained by their high proliferation rate (4.014 per day in murine unperturbed hematopoiesis) and the fact that only

one in 110 HSCs differentiates into a ST-HSC[133]. In order to confirm the long-term repopulation and self-renewal capacity of the cellular output observed, secondary transplants would have to be conducted. Highlighting the relevance of this cell population, MPP contribution was high during the whole observation period in all tissue types. In line with this, it was suggested earlier that in unperturbed hematopoiesis MPPs might be much more relevant in driving hematopoietic reconstitution than HSCs than previously assumed[164]. Regarding this, recent studies emphasized the heterogeneity of this population by dividing them into subgroups (MPP1-4) with different lineage-biased fates and temporal reconstitution abilities[46,47,321]. Attending to the immunophenotype of the population, by sorting for $CD48^+/CD150^+$ cells, only MPP1 and MPP2 subpopulations were considered in the current analysis. Although little is known about their biological function, MPP1 were described to resemble more IT/ST-HSC-like characteristics, although devoid of self-renewal. On the other hand, MPP2 is a myeloid-biased subset in the functionally heterogeneous population. The results show, that the highly heterogeneous population of MPPs and the more slowly dividing HSCs both contributed to a stable reconstitution in the analyzed period. Detailed analysis of the exact mechanisms of their spatiotemporal interaction displays an interesting field for further research.

Short-lived progenitors like CMPs and CLPs play a critical role in short-term reconstitution of the blood system, as they contribute to the generation of mature cells of the innate immune system. Shortly after HSCT, NK cells and lymphocytes arising from CLPs are the first recovering populations[322,323]. The cellular output of both populations were only visible at one week post transplantation to a low extent in BM, spleen, and thymus (Figure 12A-C). This observation is consistent with the finding, that despite generating myeloid colonies *in vivo*, CLPs show a limited myeloid potential *in vivo* after HSCT[324]. Meanwhile, in blood a high CMP contribution was shown stable until 8w post transplantation (Figure 12D). The discrepancy between the compartments can be explained by the high-turnover rate of cells found in the blood. The initial high amount of marked CMPs present accounted for high rates of mature cells, observed in the blood. As the CMPs from the initially marked pool decrease over time, the FP^+ output in the blood vanished as well. Unfortunately, no information was found on the exact turnover rates of CMPs and CLPs, though they are in general characterized by a high turnover rate and a short life span. These data indicate a short-term homing of both populations in the BM of the analyzed mice, generating myeloid output in PB up to 8w post transplantation. The contribution of CLPs could not be followed past 1w post transplantation in the tissue types. Assuming equal behavior of the transduced (marked) and non-transduced subsets in the transplanted subpopulations, this strongly suggests an even lower level of CLP compared to CMP contribution and a shorter lifespan of these cells. Thus it is likely that only extremely low amounts of cellular CLP output were detected by NGS experiments.

By considering the clone size in the NGS data analyses, no cases of clonal dominance in the hematopoietic reconstitution patterns were observed. In mostly all of the analyzed compartments, a stable polyclonal situation was observed from all analyzed HSPC populations. As expected some major clones accounted for higher output than minor ones in the initiation phase, leading to an ‘uneven distribution’ of active clones (Figure 12). Clones of

all subpopulations were detected throughout different tissue types, indicating the stable output of an ancestry population and the migration of mature cells to different tissue types. Additional sorting of final BM samples into the subpopulations, could have added an additional level of assessing clonal differentiation relationships between the HSPCs. A study by Lu et al. in 2011 addressed this question by analyzing barcoded HSC differentiation *in vivo* and found that HSC barcodes correlated less with barcodes from mature blood cells 'at the bottom of the hierarchy' but contributed more to their immediate downstream progenitor cell population MPPs[158]. The phenomenon that LT-HSCs do not contribute to lineage output during the first six months post transplantation (short-term reconstitution) was also confirmed in a lineage tracing study in humans[325].

6.1.6 Concluding remarks

Whether the data obtained here is transmittable to human hematopoietic reconstitution after stem cell transplantation remains to be elucidated. Mice represent an indispensable model organism to gain fundamental insights on basic hematology or malignancy. Overall, *in vitro* and xenograft studies using primary human cells confirmed conserved hematopoietic differentiation schemes between mice and humans[326,327]. However, differences in genetic diversity and a higher proliferative demand of cells in humans, as well as phenotype and regulatory mechanisms between the two species need to be considered. Thus, complementing murine studies with analysis of primary human material remains crucial, though not always feasible, as demonstrated by the example of clonal kinetics and lineage tracing studies. Addressing this, studies in the human system have only been conducted by analyzing lentiviral VIS in patients subjected to gene therapy[182,325,328]. In those trials, only one in 1×10^5—10^6 transplanted HSPCs were found capable of steady long-term repopulation and reconstitution[325], suggesting that human hematopoiesis post transplantation is driven by several distinct HSPC subsets with differing lineage-biases[182,328]. These data give important insights in human hematopoiesis following stem cell transplantation; however, they are based on a diseased state and concomitant treatment, which might have further impact on the clonal reconstitution patterns.

The in-house developed and optimized BC32 system used here offers a straightforward method to equip different HSPC subpopulations with genetic barcodes to track cell development simultaneously in a cell pool. Thus, hundreds of HSPCs were tracked *in vivo*, resembling the heterogeneous pool of contributing cells after HSCT. The single-cell precision level and high throughput allowed tracking even low abundant cellular events like HSC, CMP, and CLP subpopulations, which gave a limited output to mature cell populations in the observed time. Specifically, due to the low marking and amount of analyzable cells, the conclusions drawn from this experiment shows an extrapolation to the whole system. However, the study emphasized the short-term relevance of MPP subpopulation and it suggests a possible effect for this heterogeneous cell pool on long-term reconstitution as well.

Finally, future directions of the BC32 system to track the contribution of HSPC populations to hematopoietic reconstitution should meet two important criteria: considering a targeted

knock-in approach of barcodes to a defined and genetically inert locus and increase the total amount of marked cells. Ongoing optimization of the BC32 system holds great promise to further unravel the dynamics of every contributing cell population and the cellular interplay and study the defects in blood formation in malignant blood disorders.

6.2 CRISPR/Cas9 barcoding

The prospect of this project was to introduce genetic barcodes at a specific genomic site in a targeted and therefore supposedly 'neutral' fashion. One of the major concerns in current techniques used for genetic barcoding is the random lentiviral integration causing unpredictable genetic alterations on cellular behavior. Yet not fully investigated, studies addressing this issue revealed contradictory results on the effects of lentiviral integration on the behavior and potential of the transduced cell population (see section 6.1.3). Therefore, it was reasoned that optimization of the genetic barcoding by a targeted approach to a defined locus would be beneficial in terms of comparability between barcoded cell populations and barcode accessibility for subsequent NGS analyses. Furthermore, overcoming the earlier discussed limitation of moderate transduction rates could result in an increase of the actually barcoded cell population and therefore give insights on a much larger pool of cells. This is of great interest especially in hematology research, since the heterogeneity of the HSPC populations recently was increasingly investigated and confirmed by the results here, as discussed earlier. For the targeted barcoding, the CRISPR/Cas9 system was chosen, as it holds great promise to target genomic sites for precise and site-specific gene knock-out and knock-in approaches. Importantly, CRISPR/Cas9 gene editing is mainly dependent on the choice of the locus and the delivery method of the needed components. Of note, knock-in experiments require the CRISPR/Cas9 components and an additional donor sequence, as a template for HDR-directed DNA repair. Ultimately, the long-term prospect of this project was to study the applicability of the CRISPR/Cas9 system to barcoding strategies of HSPC populations for an HSCT *in vivo* reconstitution dynamics approach (as shown before).

6.2.1 Choice of appropriate loci for targeted barcoding

Initial steps to establish the CRISPR/Cas9 barcoding were the choice of locus and delivery platform. In the literature, several GSH sites were described, thoroughly investigated and known to be safe sites for integration given the lack of affecting surrounding genes. For the recent study, the *Rosa26* locus[329-333] was chosen since it is one of the best described and most commonly used GSHs for transgenic mouse models. In addition, the *Hprt* locus was also selected. The locus is characterized by its X-chromosomal location, ensuring monoallelic, targeted barcoding. Furthermore, *Hprt* disruption leads to resistance to 6-thioguanine (6-TG) cytotoxicity[334], thereby allowing for enrichment and selection of the edited cells[335,336]. Using the latter as *in vivo* selection method for edited BM in transplanted mice, Hacke and colleagues showed that 6-TG might additionally serve as a myelosuppressive agent for pre-conditioning prior to HSCT. They demonstrated that *Hprt* disruption does not lead to adverse effects and transplantation of *Hprt*-deficient cells, and resulted in high engraftment

rates and stable long-term reconstitution[335]. This would add additional value to use the locus for barcoding experiments to study clonal reconstitution of gene-edited HSPCs post HSCT. Notably in humans, hereditary *Hprt* deficiency has been associated with Lesch-Nyhan syndrome[337], hematopoiesis was described to be functionally and phenotypically normal in *Hprt* knock-out animals[335,338]. Others have recently claimed that *Hprt* expression and the associated non-essential purine salvage pathway is crucial for HSPC function and engraftment[339]. However, this could not be confirmed for steady state hematopoiesis and neither *Hprt* nor the purine salvage pathway play critical roles in cell growth and proliferation[340]. Although both loci have been used for CRISPR/Cas9 gene editing approaches, most of the current studies lack evidence on the safety and off-target evaluation of the used gRNAs and do not offer sequences of the exact sites. In contrast to the *Hprt* locus, for *Rosa26* two well-described and commonly used gRNAs were available. Thus, sophisticated *in silico* prediction tools[237,256,257] were chosen to design and evaluate the used gRNAs based on different algorithms. These predictions were reported to reliably contain 98 % of off-target sites. Correlations between predictions and predicted genetic modifications were reported to vary considerably[237], which was also observed in the CRISPR/Cas9 efficiencies at the chosen loci of the current experiments. Here, for each gRNA the three top scores of off-targets were investigated in detail and, except for one, revealed negligible modification at the indicated site, as analyzed by TIDE[341] algorithm (Figure 18C). According to the developers of the algorithm, quantitative measurement of indels is traceable with a sensitivity down to 2.5 %. Likewise, editing events at off-target sites below 2.5 % can be disregarded, which was the case in most of the analyzed sites. For future applications, of the analyzed gRNAs those lacking off-target effects or showing reduced editing efficiencies at the respective sites were chosen (ROSA2, HPRT1). However, the amount of analyzed off-target sites here only depicts a snapshot of the most likely events and still does not reflect the *in vivo* situation. Intensive studies on all off-target sites need to be performed to give in-depth insights on the genomic alterations and off-target effects caused by SpCas9.

6.2.2 IDLVs as delivery vehicle

In order to reduce the total amount of time the cells are exposed to the CRISPR/Cas9 machinery, IDLVs were used as a transient delivery vehicle of the components. Especially for episomal gene delivery of donor templates to induce HDR in HSPCs in a hit-and-run fashion, IDLVs have been used extensively in research[224,226]. The most commonly used and well-described mutation in the catalytic domain of the enzyme is D64V[342-346], and therefore was used in the current study as well. For the IDLV constructs, exceptionally low titers were obtained, impeding the subsequent *in vitro* work by limiting the MOIs being used (Figure 17A and Table 26, Appendix C). However, using low-titer IDLV preparations, transduction of the cells did not achieve high and constant rates of reporter gene expression (data not shown). To overcome low transduction rates after IDLV infection, FACS sorting was performed on fluorescent marker expressing cells, simulating high initial percentages of fluorescent positive cells and enrich the initial population in possibly gene-edited cells. A clear reduction in titers of IDLVs and fluorescent marker expression compared to ILVs was reported already by several

groups in varying ranges[347,348]. Taking the cargo size of the virus into account, a maximal used size of 7920 bp (target vector) between the LTRs of the viral expression plasmid is likely to affect the output of infectious particles in the virus production[349]. Using a reduced-size target vector with a cargo size of 3630 bp (lacking SpCas9) dramatically increased titers of IDLV production (data not shown). Aiming to reduce the size of SpCas9 several work focused on various approaches, ranging from a 133 amino acids deletion mutant[202] to split versions of the enzyme[350,351] or *trans* protein splicing[352] leading to SpCas9 encoded on two different vectors though mediating lower rates of mutagenesis. However, this is not feasible in an HDR-mediated approach, since by the use of an additional donor construct the amount of different constructs might already be exhausted. Moreover, using SpCas9 orthologues from different bacterial or archeal species could provide the potential to reduce the cargo size by ≈300 amino acids (i.e., SaCas9). It has been used successfully in gene editing of HSPCs[353] with comparable results to SpCas9[354,355]. Another popular alternative for delivering CRISPR/Cas9 donors is the application of rAAV6 vectors, with the reemerging problem of limitations in cargo size and the lack of capacity to transduce a broad range of cells. Thorough comparison of donor AAVs and IDLVs has not been performed to date. Since the use of rAAV6 vectors would have required a splitting of large cargos (and thus, increasing the number of vectors needed), this option was not considered for optimizing the delivery strategy. Finally, as an alternative to IDLVs as transient delivery platforms, chemical transfection was performed to compare cellular outcome and knock-in rates.

Working with IDLVs in general raises the question for retaining lentiviral integration capacity, which the vectors might possess. This is crucial in the performed experiments since successful knock-in rates in target cells were in the first place confirmed by and dependent on stable fluorescent marker expression after eight weeks. The residual background integration from IDLV productions harboring the D64V mutant varies substantially, with reported reduced integration potentials compared to ILVs from 10^4-fold[356,357] to 10^1-fold[346]. Integrase-independent or illegitimate integration events were observed in transduced cells (0.2—0.5 %)[357,358] and prolonged SV40-dependent transgene expression (up to 56 days) was reported in 293T cells independent of the transient nature of IDLV transgene expression[347]. Yet not completely resolved, illegitimate integration is believed to occur due to linear episome formation during reverse transcription process[358]. Although effort has been taken in improving IDLV design, remaining integration capabilities of IDLVs cannot be precluded.

Exploiting the presence of linear DNA intermediates or episomal DNA derived from IDLV vectors after reverse transcription in the cells is commonly used as CRISPR/Cas9 donors in knock-in approaches. After linearization of the donor construct, the DNA is supposed to facilitate HDR via transgene-flanking homologous regions. Interestingly, recent work found evidence on an HDR-independent mechanism from IDLVs, not resulting from HDR, but from recombination of the complete IDLV genome to the target region. This was demonstrated by the presence of the concatemeric structure of the IDLV genome at the target site, instead of integration of the transgene exclusively. The integration seemed highly efficient, as up to 80 %o of human embryonic stem cells were targeted[359]. This was explained by the recruitment

of cellular proteins supporting HDR from the non-functional IDLV pre-integration complex and could be a possible hint of the advantage of using donor IDLVs over plasmid transfection. The here used combination of genome- and construct-specific primers for identification of a successful knock-in does not fully exclude the possibility of concatemer formation in the samples. Verification of knock-in was conducted by amplification of the correct insertion of the 5' end of the donor sequence. Since conventional HDR can still take place on one donor sequence end, the concatemer phenomenon would remain undiscovered on the residual end of the donor vector.

6.2.3 CRISPR/Cas9 barcoding evaluation

In vitro knock-in of barcode donor constructs into two different loci in a murine adherent cell line bulk (NIH/3T3) resulted in low barcoding rates of 1.7 % (*Rosa26*) and 7.5 % (*Hprt*) via IDLV transduction. Upon chemical transfection with plasmids equivalent to IDLVs barcoding rate increased to 6.7 % (*Rosa26*) or remained at equal levels of 6.0 % (*Hprt*). Despite analyzed differences in initial CRISPR/Cas9 target activities, no differences in knock-in frequencies in the loci were observed, potentially showing target activities to be highly variable and dependent on experimental conditions. In the literature, only few accounts exist reporting on actual CRISPR/Cas9-mediated *in vitro* knock-in rates using IDLV donor templates without prior selection. A study comparing different donor template architectures found knock-in events in up to 10 % of HEK293T cells[220], while the group of Doudna increased knock-in rates in cell lines from 9 to 20 % by induction of cell cycle synchronization prior to treatment[228]. The actual integration events discovered here were slightly lower compared to others reported in literature achieved by *in vitro* gene manipulation of cell lines. Of note, most of the currently used CRISPR/Cas9 knock-in approaches employ different delivery platforms of the components, which are mostly either SpCas9/gRNA RNPs or a combination of SpCas9 mRNA, gRNA and donor nucleic acid template (plasmid, AAV or IDLV), but less likely two vector constructs encoding all components. The latter requires co-transduction, which in turn might evoke two challenging issues: homologous viral interference and receptor downregulation or receptor block. Viral interference occurs when the abundance of two viruses of the same genetic background aggravate the replication of one virus, which can happen dependently or independently of receptor downregulation[360], which prevents the entry of the second virus[361]. A study by Schott et al., evaluating kinetics of fluorescent proteins expression after IDLV infection found no evidence on receptor blockade after retransduction[362]. The exact underlying mechanisms of viral interference have yet to be elucidated. However, it was suggested that high MOIs and pseudotyping of the viral particles with VSV-G should prevent those issues[363]. To address the conceivable effect of the co-transduction itself on CRISPR/Cas9 barcoding efficiencies, an SpCas9 stably expressing cell line was generated using the same target cell line. Barcoding of this cell line therefore only relies on the delivery of gRNA harboring donor constructs, and it was reasoned to be a suitable alternative for the co-transduction experiments. Notably, a highly artificial system was created not reflecting the physiology of the ideal target cell population, murine HSPCs. By using a single transduction experimental design, barcoding of SpCas9 expressing cells was increased 54-fold (*Rosa26*) or

12-fold (*Hprt*) to barcode knock-in rates in ≈89 % of the cells This suggests a spatiotemporal influence of SpCas9 protein expression during entry of gRNA and donor constructs, thus possibly leading to augmented knock-in events (Figure 19A and Figure 19B). However, off-target activity and toxicity in SpCas9 stably expressing organisms is still a matter of debate with controversial opinions[364,365]. Recently a correlation between p53 activation status and SpCas9 expression was discovered, leading to a selection of cells with p53 inactivating mutations[366]. In line with those results, others proposed a relation between p53 inhibition and an improvement of HDR-mediated DNA repair after CRISPR/Cas9-mediated gene editing[367]. This unfavorable modification in cellular properties remains a major concern in CRISPR/Cas9 gene editing technology but it might partially explain the elevated knock-in rates observed in SpCas9 stably expressing cell lines. Thus, mice strains with inducible SpCas9 expressions have been generated[368,369], which would pose a possibility to translate the CRISPR/Cas9 barcoding to an *in vivo* system.

Analysis of CRISPR/Ca9-mediated knock-in events was performed via sequencing of single-cell clones and via flow cytometry in bulk samples, measuring the expression of the knock-in fluorescent marker. A standard workflow to assess HDR- or NHEJ-editing events in bulk samples is the PCR amplification of an appropriate region surrounding the target site, followed by cloning and Sanger sequencing of the plasmids[220]. However, this method is laborious and time-consuming, as a certain amount of different clones needs to be analyzed for reliable quantification. Therefore, an alternative ddPCR-based approach was pursued, which did not reveal analyzable results. Technical issues arose due to the template length, which ideally ranges between 60 to 200 bp, albeit larger products (donor HA length ≤450 bp) can be amplified. The homologous regions of 1 kb and GC-low, repetitive sequences in the genomic regions of *Rosa26* and *Hprt* limited the options on shorter amplicon lengths. Even though the manufacturer supplies data on amplification of products of up to 2.5 kb by adapting PCR cycling conditions[370], any optimization of the protocol did not result in separable droplets in the readout of the PCR (data not shown). Knock-in of the barcodes in the correct orientation was verified by conventional PCR, but exact determination of knock-in frequency in bulk samples could not be performed yet and would be an important step in evaluating the outcome of the methods. If no specially designed reporter systems are used and PCR-based approaches can be excluded due to amplicon length, a reliable alternative, though cost-intensive is the use of deep-sequencing methods.

6.2.4 Delivery and knock-in in murine HSPCs

The efficient delivery and an adequate level or HDR in primary cells such as HSPCs is critical for clinical translation of the results. While for human $CD34^+$ cells the induction of a CRISPR/Cas9 knock-in to replace and restore gene function was reported to work with a sufficiently high amount of HDR frequency[245-251,371], fundamental knowledge on knock-in strategies in murine HSPCs is currently lacking. One of the first successful experimental trials on nuclease-mediated HDR in murine HSPCs *in vitro* was reported before, albeit it was shown to occur at unquestionable low frequencies[343]. Although, the objective of CRISPR/Cas9

barcoding of HSPCs for lineage tracing was not achieved here, several parameters need to be taken into account for future applications.

As discussed above, LSKs, or more precisely HSPCs are a rather heterogeneous pool of cells with disparate proliferation and differentiation abilities and thus, presumable differences in susceptibility to viral transduction and gene editing. Regarding HDR efficiencies, knock-in frequency was shown to be more efficient in downstream progenitor cells, compared to LT-HSCs[224-226,247,252,372]. Comparing delivery efficiencies in human HSC and MPP populations to the total HSPC cell pool ($CD34^+$) resulted in a 38 % reduction in reporter positive cells[247]. Others reported 8—10 % HDR targeting in SpCas9 expressing HSCs and MPP1 subset compared to 25—30 % HDR-edited cells in the MPP2 and MPP3/4 LSK subset[253]. Thus, lower HDR efficiencies were associated with phenotypically more primitive or quiescent cells, such as HSCs and MPPs[47,373]. This correlation is likely due to proliferation rates of the cell populations, since HDR repair pathway is favored in S and G2 phase of the cell cycle[374,375]. All of those experiments additionally confirmed unaltered engraftment and multi-lineage reconstitution upon transplantation with no detected off-target effects. However, comprehensive lineage profiling to detect putative differentiation bias after *ex vivo* editing is still lacking. The multilineage differentiation capacity was mainly demonstrated by the mature cellular progeny of CRISPR/Cas9 edited cells after engraftment, disregarding the comparison to the physiological properties of a control group[247,253].

From a therapeutically point of view low knock-in efficiencies do not pose a major hurdle, because restoration of WT protein in approximately 10 % of physiologically relevant cells is sufficient to induce the desired therapeutic response[254]. Alternatively, for approaches aiming to overcome limited amounts of gene edited human LT-HSCs, Dever at al. found that prior sorting of a $reporter^{high}$ expressing cell population resulted in 90 % of HDR-edited cells in murine tissue 16 weeks post transplantation[247]. This would be of great interest for the barcoding procedure in order to elevate the amount of traceable cells in the mice after HSCT. However, CRISPR/Cas9 editing and sorting of cells are critical factors regarding HSC survival for downstream applications, such as HSCT. A 20 % reduction in cell viability was observed after mock electroporation of HSPCs[245], further impaired by the delivery of CRISPR/Cas9 components[247,376]. In the murine transplantation setting, resultant low cell counts of transplantable HSPC subsets would represent a technical challenge for this strategy, as shown by the *in vivo* results, discussed above. For this approach, the amount of donor animals would need to be increased, in order to obtain appropriate cell numbers after additional sorting.

6.2.5 Future directions

Targeting high amounts of HSPC subpopulations via a neutral CRISPR/Cas9-induced barcoding approach holds great promise for augmented future lineage tracing studies. In order to obviate the discussed limitations, such as delivery platforms or undesired integrative vector based approaches, new strategies need to be considered. The majority of promising experimental work focused on direct delivery of SpCas9 mRNA or RNP complexes, bypassing the need for a second viral vector construct and resulting in high efficiencies in a hit-and-run

fashion (and thereby, reduced off-target effects). Thus, it was reasoned that a single-component SpCas9 protein-based approach would yield in the maximum possible barcoding efficiencies in murine HSPCs. Recently, several groups emerged with the idea of viral-like extracellular vesicles (EV) for transient transfer of SpCas9 protein. This was achieved by fusing endonucleases (such as SpCas9 or ZFNs) to HIV and MLV Gag protein[377-380], the use of high affinity EV dimerization systems[381,382], or passive incorporation of SpCas9 to EVs[383]. Alternatively, mouse strains with an inducible SpCas9 expression are under current investigation and are optimized to reduce off-target effects and genotoxicity of the components needed. These strategies, benefiting from a transient SpCas9 expression will be the methods of choice for further optimization of CRISPR/Cas9 barcoding and need to be assessed for their compatibility in a knock-in approach harboring an additional donor sequence for HDR.

7 Materials

7.1 Kits

Commercial kits (listed in Table 2) were used for DNA isolation and purification, lineage depletion of BM cells, and cloning according to the manufacturer instructions unless otherwise specified in section 8.

Table 2: Ready-made commercial kits utilized.

Product name	Company (City, Country Headquarter)	Cat. No.
QIAamp DNA Mini Kit (250)	Qiagen (Hilden, Germany)	51306
QIAamp DNA Micro Kit (50)	Qiagen (Hilden, Germany)	56304
DNeasy Blood & Tissue Kit (250)	Qiagen (Hilden, Germany)	69506
NucleoBond Xtra Midi kit (50)	Macherey-Nagel (Düren, Germany)	740410.50
QIAquick Gel Extraction Kit (250)	Qiagen (Hilden, Germany)	28706
Direct Lineage Cell Depletion Kit, mouse	Miltenyi Biotec (Bergisch Gladbach, Germany)	130-110-470
QIAquick PCR Purification Kit (250)	Qiagen (Hilden, Germany)	28106
Qubit™ dsDNA HS Assaykit	Thermofisher Scientific (Waltham, MA, USA)	Q32854
NEB® 5-alpha Electrocompetent *E.coli*	NEB (Ipswich, MA, USA)	C2989K
One Shot™ ccd B Survival 2 T1R Kompetente Zellen	Thermofisher Scientific (Waltham, MA, USA)	A10460
SOC Outgrowth Medium	NEB (Ipswich, MA, USA)	B9020S
Q5® Site-Directed Mutagenesis Kit	NEB (Ipswich, MA, USA)	E0554S
NEBuilder® HiFi DNA Assembly Cloning Kit	NEB (Ipswich, MA, USA)	E5520S
QIAGEN Multiplex PCR Plus Kit (100)	Qiagen (Hilden, Germany)	206152
Agencourt AMPure XP beads	Beckman Coulter (Brea, CA, USA)	A63880
jetOPTIMUS DNA Transfection Reagent	Polyplus Transfection (New York City, NY, USA)	117-01

7.2 Antibodies

All antibodies used for the sorting of HSPCs and at the final stage of *in vivo* reconstitution dynamics experiments are listed in Table 3.

Table 3: Antibodies utilized for sort and staining of murine cells.

Antibody name [clone number]	Company (City, Country Headquarter)	Cat. No.
TruStain fcX™ (anti-mouse CD16/32) Antibody [93]	Biolegend (San Diego, CA, USA)	101320
PE/Cy7 anti-mouse Ly-6A/E (Sca-1) Antibody [E13-161.7]	Biolegend (San Diego, CA, USA)	122514
APC anti-mouse CD150 (SLAM) Antibody [TC15-12F12.2]	Biolegend (San Diego, CA, USA)	115910
Pacific Blue™ anti-mouse CD117 (c-Kit) Antibody [2B8]	Biolegend (San Diego, CA, USA)	105820
PE anti-mouse CD34 Antibody [MEC14.7]	Biolegend (San Diego, CA, USA)	119308
APC/Cy7 anti-mouse CD117 (c-kit) Antibody [2B8]	Biolegend (San Diego, CA, USA)	105826
PE anti-mouse CD150 (SLAM) Antibody [TC15-12F12.2]	Biolegend (San Diego, CA, USA)	115904
PE anti-mouse Ly6G/Ly6C (Gr-1) [RB6-8C5]	Biolegend (San Diego, CA, USA)	108408
APC anti-mouse/human CD11b Antibody [M1/70]	Biolegend (San Diego, CA, USA)	101212
Brilliant Violet 785™ anti-mouse CD3ε Antibody [145-2C11]	Biolegend (San Diego, CA, USA)	100355
PE/Cy7 anti-mouse/human CD45R/B220 Antibody [RA3-6B2]	Biolegend (San Diego, CA, USA)	103222

7.3 Enzymes

For cloning procedures, the enzymes shown in Table 4 were used according to the manufacturer recommendations. Amplification of DNA fragments to be used for subsequent cloning was performed using proofreading polymerases (Phusion or Q5 polymerase), while PCR for validation of the presence of fragments was performed with either ddPCR Master Mix (Biorad, Hercules, CA, USA) or DreamTaq DNA Polymerase (Thermofisher Scientific, Waltham, MA, USA).

Table 4: Enzymes used for amplification and cloning.

Enzyme	Company (City, Country Headquarter)	Cat. No.
DreamTag DNA Polymerase (5 U/ µL)	Thermofisher Scientific (Waltham, MA, USA)	EP0703
Phusion High-Fidelity DNA Polymerase (2 U/ µL)	Thermofisher Scientific (Waltham, MA, USA)	F530L
Q5® High-Fidelity DNA Polymerase	NEB (Ipswich, MA, USA)	M0491S
T4 DNA-Ligase (5 U/ µL)	Thermofisher Scientific (Waltham, MA, USA)	EL0011
T4 DNA-Ligase (30 U/ µL)	Thermofisher Scientific (Waltham, MA, USA)	EL0013
ddPCR™ Supermix for Probes (No dUTP)	Bio-Rad (Hercules, USA)	1863024
T4 Polynucleotide Kinase (10 U/ µL)	Thermofisher Scientific (Waltham, MA, USA)	EK0031
FastAP Thermosensitive Alkaline Phosphatase (1 U/ µL)	Thermofisher Scientific (Waltham, MA, USA)	EF0652
FastDigest MreI	Thermofisher Scientific (Waltham, MA, USA)	FD2024
FastDigest MauBI	Thermofisher Scientific (Waltham, MA, USA)	FD2084
FastDigest NotI	Thermofisher Scientific (Waltham, MA, USA)	FD0596
FastDigest BamHI	Thermofisher Scientific (Waltham, MA, USA)	FD0055
FastDigest EcoRI	Thermofisher Scientific (Waltham, MA, USA)	FD0274
FastDigest ApaI	Thermofisher Scientific (Waltham, MA, USA)	FD1414
FastDigest MfeI (MunI)	Thermofisher Scientific (Waltham, MA, USA)	FD0753
FastDigest BsrGI (Bsp1407I)	Thermofisher Scientific (Waltham, MA, USA)	FD0933
FastDigest BsmBI (Esp3I)	Thermofisher Scientific (Waltham, MA, USA)	FD0454
Eco72I (PmlI) (10 U/ µL)	Thermofisher Scientific (Waltham, MA, USA)	ER0361

7.4 Cell culture media, buffers, and compounds

All ingredients and reagents used for the preparation of cell culture media, as well as the buffers for all protocols are listed in Table 5 (commercially available) or prepared as described in the subsections below (section 7.4.1 – 7.4.5).

Table 5: Media, buffer and additives used for their preparation.

Medium/ Buffer/ Additive (conc.)	Company (City, Country Headquarter)	Cat. No.
AMD3100	MedChemExpress (Monmouth Junction, NJ, USA)	HY-10046
Busulfan	Sigma Aldrich (St. Louis, MO, USA)	B2635-25G
DMEM high Glucose	Thermofisher Scientific (Waltham, MA, USA)	11965092
Enrofloxacin	Bayer AG (Leverkusen, Germany)	770-060
Fetal Bovine Serum	Thermofisher Scientific (Waltham, MA, USA)	10270-106
Sodium Pyruvate (100 mM)	Thermofisher Scientific (Waltham, MA, USA)	11360070
HEPES (1 M)	Thermofisher Scientific (Waltham, MA, USA)	15630080
Penicillin-Streptomycin (10,000 U/ mL)	Thermofisher Scientific (Waltham, MA, USA)	15140122
L-Glutamine	Thermofisher Scientific (Waltham, MA, USA)	25030024
Chloroquine diphosphate salt	Sigma Aldrich (St. Louis, MO, USA)	C6628-25G
Polybrene Infection / Transfection Reagent	Sigma Aldrich (St. Louis, MO, USA)	TR-1003-G
StemSpan Medium SFEM	Stemcell Technologies (Vancouver, Canada)	9650
Gibco DPBS	Thermofisher Scientific (Waltham, MA, USA)	14190250
mSCF	PeproTech (Rocky Hill, NJ, USA)	250-03
LB medium	Carl Roth (Karlsruhe, Germany)	X964.2
Ampicillin	Carl Roth (Karlsruhe, Germany)	K029.1
Fast-Media® Amp Agar	InvivoGen (San Diego, CA, USA)	fas-am-s

7.4.1 Cell culture media

For outgrowth and maintenance of HEK293T cells Dulbecco's Modified Eagle Medium (DMEM), supplemented with 10 % fetal bovine serum (FBS), 1 % sodium pyruvate and 1 % penicillin-streptomycin (P/S) was prepared. If used for transient transfection and virus production, 20 mM HEPES and 25 µM chloroquine were added just before use. As growth medium for NIH/3T3 cells, DMEM supplemented with 10 %FBS and 1 % P/S was used. FBS was heat-inactivated for 10 min at 56°C for all purposes unless otherwise stated. The StemSpan medium for the transient culture of HSPCs was composed of 1 % P/S, 1 % sodium pyruvate, 1 % L-glutamine and 50 ng/ mL mSCF. Transduction of cell lines and primary cells was performed after the addition of 8 µg/ mL polybrene to the respective medium.

7.4.2 Transient transfection buffer

For transient transfection, 2× HEPES-buffered saline (HEBS) Buffer was prepared according to the following recipe: 275.8 mM $CaCl_2$, 10.2 mM KCl, 1.41 µM Na_2HPO_4, 42 mM HEPES, and 1.1 mM Glucose. The pH was adjusted to 7.05. Buffer was sterilized using 0.2 µM filters. Indicated salts were purchased from Sigma Aldrich unless otherwise mentioned.

7.4.3 Buffers used for cell preparation and analysis

Lysis of erythrocytes during mouse sample handling was performed using ACK buffer, prepared as follows: 155 mM NH_4Cl, 10 mM $KHCO_3$, and 87 µM Mg-Titriplex*$2H_2O$ (87µM); pH was adjusted to 7.4. Magnetic-activated cell sorting (MACS) buffer for lineage depletion

procedures was prepared by adding 0.5 % BSA and 2 mM EDTA to Dulbecco's Phosphate-buffered Saline (DPBS) solution adjusting the pH to 7.2. Both buffers were sterilized via 0.2 µM filtration. Subsequent FACS analysis mouse samples was perfomed using DPBS supplied with 2 % FBS (FACS buffer).

7.4.4 Media used for bacterial culture

Lysogeny broth medium for bacterial culture was prepared by dissolving 20 g LB medium powder in 1 L of distilled water. Medium bottles were autoclaved and stored at 4°C until further usage. Ampicillin-containing selection medium (LB-Amp) and LB-Amp agar plates were prepared by addition of 0.1 mg/ mL ampicillin freshly before use.

7.4.5 Buffers used for DNA purification and analysis

TE buffer for elution of DNA contained 1 mM EDTA and 10 mM Tris-Cl; pH was adjusted to 8.0.

Tris-borate-EDTA (TBE) buffer for electrophoresis was prepared as 10×stock by mixing Tris base (1 M), boric acid (1 M) and EDTA (0.02 M). Buffer was diluted 1:10 for proper working concentration. The working solution was stored for up to six months at room temperature.

7.5 Primer

Desalted of HPLC-purified primers and oligonucleotides were purchased from Metabion (Planegg, Germany). All used primers are listed in Table 6 and primers used for specified purposes are mentioned in the respective methods section (see section 8).

Table 6: Primer and oligonucleotides sorted by their specific purposes. N positions indicate random integration of A, T, C, or G during oligonucleotide generation and thus creating a 25 % chance of each base to be inserted.

Primer name	Sequence (5'-3')	Purpose
	Barcoding	
32BC-Poly-fw	GGTGCATCTAGAACACTC	Amplification of BC
32BC-Poly-rev	CATAGTGCACCTCGAG	Amplification of BC
Poly-GFP-BC-fw	GGTGCATCTAGAACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNACTN NCGANNCTTNNCGANNCTTNNGGANNCTANNACTNNCGANNCTTNNCGANNC TTNNGGANNCTANNACTNNCGANNCTCGAGGTGCACTATG	BC32 GFP Backbone
Poly-Venus-BC-fw	GGTGCATCTAGAACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNCGAN NAGANNCTTNNCGANNCTANNGGANNCTTNNCGANNAGANNCTTNNCGANNC TANNGGANNCTTNNCGANNAGANNCTCGAGGTGCACTATG	BC32 Venus Backbone
Poly-Cerulean-BC-fw	GGTGCATCTAGAACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNCAGN NATCNNCTTNNCGANNGGANNCTANNCTTNNCAGNNATCNNCTTNNCGANNG GANNCTANNCTTNNCAGNNATCNNCTCGAGGTGCACTATG	BC32 TSapphire Backbone
Poly-Cherry-BC-fw	GGTGCATCTAGAACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNCTAN NCAGNNCTTNNCGANNCTANNCTTNNGGANNCTANNCAGNNCTTNNCGANNC TANNCTTNNGGANNCTANNCAGNNCTCGAGGTGCACTATG	BC32 eBFP Backbone
	Cloning	
NotI-ID.A-EFS_fw	ATAGCGGCCGCTAGGATAGGATAGGATAGGACTACATCCCGATTGGCTCCGG TGCCCGTCA	Cloning Identifier (a) sequence in LeGO backbone
NotI-ID.B-EFS_fw	ATAGCGGCCGCCTCAGCTCAGCTCAGCTCAGCTCAGCTAAAATTGGCTCCGG TGCCCGTCA	Cloning Identifier (b) sequence in LeGO backbone
NotI-ID.C-EFS_fw	ATAGCGGCCGCGGACAGGACAGGACAGGACAGGACAGTAAAATTGGCTCCGG TGCCCGTCA	Cloning Identifier (c) sequence in LeGO backbone
NotI-ID.D-EFS_fw	ATAGCGGCCGCTTCAGTTCAGTTCAGTTCAGAGCCCGAGGGATTGGCTCCGG TGCCCGTCA	Cloning Identifier (d) sequence in LeGO backbone
EFS-BamHI_rv	GCGGGATCCCGCGTCACGA	Cloning Identifier sequences in LeGO backbone

Primer name	Sequence (5'-3')	Purpose
BamHI-FP_fw	ATAGGATCCCGCCACCATGGTGAGC	Cloning Identifier sequences in LeGO backbone
FP-ID.A-EcoRI_rv	TACGAATTCCGGGATGTAGTCCTATCCTATCCTATCCTATTACTTGTACAGC TCGTCCATGCC	Cloning Identifier (a) sequence in LeGO backbone
FP-ID.B-EcoRI_rv	TACGAATTCTTTAGCTGAGCTGAGCTGAGCTGAGCTGAGTTACTTGTACAGC TCGTCCATGCC	Cloning Identifier (b) sequence in LeGO backbone
FP-ID.C-EcoRI_rv	TACGAATTCTTTACTGTCCTGTCCTGTCCTGTCCTGTCCTTACTTGTACAGC TCGTCCATGCC	Cloning Identifier (c) sequence in LeGO backbone
FP-ID.D-EcoRI_rv	TACGAATTCAAGTCAAGTCAAGTCAAGTCTCGGGCTCCCTTACTTGTACAGC TCGTCCATGCC	Cloning Identifier (d) sequence in LeGO backbone
pMDLg-IN(D64V)_fw	TGGCAGCTAGtgTGTACACATTTAGAAG	Mutagenese Integrase in pMDLg-pRRE
pMDLg-IN(D64V)_rv	TATTCCTGGGCTACAGTC	Mutagenese Integrase in pMDLg-pRRE
U6-sgRNA_fw	AGTTTGGTTAGTACCGGGCCGAGGGCCTATTTCCCATG	Amplification U6-gRNA-tracrRNA fragment
U6-sgRNA_rev	GACCTCGACTAGAGCGGGCCCACTCGAGTTCTAGAGGAG	Amplification U6-gRNA-tracrRNA fragment
	ddPCR	
ID.A_WPRE_fw	TAGGATAGGATAGGATAGGACTACATC	ddPCR IDa
ID.B_WPRE_fw	CTCAGCTCAGCTCAGCTCAGC	ddPCR IDb
ID.C_WPRE_fw	GGACAGGACAGGACAGGACAG	ddPCR IDc
ID.D_WPRE_fw	TTCAGTTCAGTTCAGTTCAGAGCC	ddPCR IDd
WPRE_rv	AGTGGAACCGGAACCCTTAAAC	ddPCR ID
WPRE_p	CGTCGAGGGACCTAATAACTTCGTATAGCATACATT	ddPCR ID (probe)
mROSA26-Intr1_fw	atcgggactctggcgg	ddPCR ROSA26 knock-in
mROSA26-Intr_p	acctgggagcgcctcctcgcc	ddPCR ROSA26 knock-in (probe)
preSFFV_rv	tagcttgccaaacctacagg	ddPCR ROSA26 knock-in
mHPRT-Intr3_fw	GCGTCTGGGAATTGAACG	ddPCR HPRT1 knock-in
preSFFV_p	cggccgctgaaagaccccacctgt	ddPCR HPRT1 knock-in (probe)
SFFV_rv	accgcagatatcctgtttgg	ddPCR HPRT1 knock-in
Y Chromosome fw	ACAAGTTTTGGGACTGGTGACAA	ddPCR Y chromosome
Y Chromosome Rv	ACCACGGGACCACACCATAA	ddPCR Y chromosome
Y Sonde	TGTCAAGCGCCCCATGAATGCA	ddPCR Y chromosome (probe)
Illu_fw	CTAGAAAAACATGGAGCAATCACAA	ddPCR Illumina sequence
Illu_rv	AGCGTCGTGTAGGGAAAGAGTG	ddPCR Illumina sequence
Illu_p	CTGAATGATACGGCGACCACCGTCTAG	ddPCR Illumina sequence (probe)
Epo_fw	GCAGGCGGGGTCGCTACTC	ddPCR mEpo
Epo_rv	CGCCTGTGCAGATCCGATAA	ddPCR mEpo
Epo_p	TTCTGAGGCGCCACTTTTGCAAGACC	ddPCR mEpo (probe)
	CRISPR/Cas9 Cloning	
sgROSA26.1_fw	CACCGACTCCAGTCTTTCTAGAAGA	gRNA ROSA1
sgROSA26.1_rv	AAACTCTTCTAGAAAGACTGGAGTC	gRNA ROSA1
sgROSA26.2_fw	CACCGATGTCTTTAATCTACCTCGA	gRNA ROSA2
sgROSA26.2_rv	AAACTCGAGGTAGATTAAAGACATC	gRNA ROSA2
sgHPRT.1_fw	CACCgCAACTAGAATGATCAGTCAA	gRNA HPRT1
sgHPRT.1_rv	AAACTTGACTGATCATTCTAGTTGC	gRNA HPRT1
sgHPRT.2_fw	CACCGAAGTACTCATTATAGTCAA	gRNA HPRT2
sgHPRT.2_rv	AAACTTGACTATAATGAGTACTTC	gRNA HPRT2
mROSA26-LHA_fw	taagggttccaagcttaagcCAGGCCCTCCGAGCGTGG	Amplification LHA ROSA1/2
mROSA26-LHA_rv	tacaggtggggtctttcagcCTAGAAAGACTGGAGTTGCAGATCACGAG	Amplification LHA ROSA1/2
mROSA26-RHA_2_fw	agggactgtggcttctctcgcgcgcgCTAGAAGATGGGCGGGAG	Amplification RHA ROSA1/2
mROSA26-RHA_2_rv	cttgtgctcttccgatctcgCAGTTTATAAATGGAGAAAAAGGAGAG	Amplification RHA ROSA1/2
mHPRT-LHA1_fw	taagggttccaagcttaagcCTTACATTTTCTTCTAATCCAC	Amplification LHA HPRT1
mHPRT-LHA1_rv	tacaggtggggtctttcagcggccgcTCATTCTAGTTGAAAAAAGAAC	Amplification LHA HPRT1
mHPRT-RHA1_fw	agggactgtggcttctctcgcgcgcGGGACATAAAAGTTATTGGTGG	Amplification RHA HPRT1
mHPRT-RHA1_rv	cttgtgctcttccgatctcgTCTGGCAGTTGTACACCAAATG	Amplification RHA HPRT1
mHPRT-LHA2_fw	taagggttccaagcttaagcTAGACTAGTCTCTATGTGATTAAAATTG	Amplification LHA HPRT2
mHPRT-LHA2_rv	tacaggtggggtctttcagcggccgCATATCCAACAACAAACTTG	Amplification LHA HPRT2
mHPRT-RHA2_fw	agggactgtggcttctctcgcgcgcgTAATGAGTACTTCAGGGATTTG	Amplification RHA HPRT2
mHPRT-RHA2_rv	cttgtgctcttccgatctcgCTGAAACATGTTTATTCCAC	Amplification RHA HPRT2
	CRISPR/Cas9 Analysis	
mROSA26.1-edit_fw	TTTGGAGGCAGGAAGCACTT	Amplification ROSA1 target site (TIDE analysis)
mROSA26.1-edit_rv	GAGGCGGATCACAAGCAATA	Amplification ROSA1 target site (TIDE analysis)
mROSA26.2-edit_fw	AAAGGCTAACCTGGTGTGTG	Amplification ROSA2 target site (TIDE analysis)
mROSA26.2-edit_rv	GGGGTTGGATAAGCCAGTAT	Amplification ROSA2 target site (TIDE analysis)
mHPRT.1-edit_fw	TGCCCACTCAGAAATAGATG	Amplification HPRT1 target site (TIDE analysis)
mHPRT.1-edit_rv	TTGGCTATAAAGTCCAGGAG	Amplification HPRT1 target site (TIDE analysis)
mHPRT.2-edit_fw	GGCCAGACTGTAAGTAGATGC	Amplification HPRT2 target site (TIDE analysis)
mHPRT.2-edit_rv	GTTCAACTACTTCAGCCAGCAA	Amplification HPRT2 target site (TIDE analysis)
mROSA26(Intr)-edit_fw	ctcgtcgctgattggcttct	Verification ROSA1/2 knock-in
mHPRT(Intr)-edit_fw	ACAACCATCTGTTATGGGATCTG	Verification HPRT1 knock-in
mHPRT(Intr)1_rv	TGCCAGCCCCGAAAACGGATTA	Verification HPRT1 knock-in
mHPRT(Intr)2_rv	AAGTGGGAAAATACAGCCAACACTG	Verification HPRT2 knock-in
	Other Anaylses	
lvLTR_fw	GAGCTCTCTGGCTAACTAGG	LM-PCR: Identification VIS
lvLTR-nested1_fw	AGCTTGCCTTGAGTGCTTCA	LM-PCR: Identification VIS
lvLTR-nested2_fw	AGTAGTGTGTGCCCGTCTGT	LM-PCR: Identification VIS
lvLTR-nested1_rv	CCTAACTGCTGTGCCACT	LM-PCR: Identification VIS
lvLTR-nested2_rv	GAATTCAGATCTCCCGCG	LM-PCR: Identification VIS
lvLTR-nested2_rv	GAATTCAGATCTCCCGTTAA	LM-PCR: Identification VIS
	Sanger sequencing	
bGlobin_fw	TCTTCCTCCCACAGCTC	Sequencing Cas9-LP
Cas9-seq1_fw	CTGACCCCCAACTTCAAG	Sequencing Cas9-LP
Cas9-seq2_fw	AAGGCCATCGTGGACCT	Sequencing Cas9-LP
Cas9-seq3_fw	ATCGTGCCTCAGAGCTT	Sequencing Cas9-LP
Cas9-seq4_fw	CACCGTGGCCTATTCTGT	Sequencing Cas9-LP

Primer name	Sequence (5'-3')	Purpose
gag-seq_fw	TAAGAGCCGAGCAAGCTT	Sequencing Cas9-LP
mHIV1-pol_rv	ctggccatcttcctgctaat	Sequencing D64V mutation pMDLg-pRRE
bGlobin_fw	TCTTCCTCCCACAGCTC	Sequencing Cas9-LP packaging Plasmid pCas9-PH-pMDLg-pRRE(D64V)
Cas9-seq1_fw	CTGACCCCCAACTTCAAG	Sequencing Cas9-LP packaging Plasmid pCas9-PH-pMDLg-pRRE(D64V)
Cas9-seq2_fw	AAGGCCATCGTGGACCT	Sequencing Cas9-LP packaging Plasmid pCas9-PH-pMDLg-pRRE(D64V)
Cas9-seq3_fw	ATCGTGCCTCAGAGCTT	Sequencing Cas9-LP packaging Plasmid pCas9-PH-pMDLg-pRRE(D64V)
Cas9-seq4_fw	CACCGTGGCCTATTCTGT	Sequencing Cas9-LP packaging Plasmid pCas9-PH-pMDLg-pRRE(D64V)
gag-seq_fw	TAAGAGCCGAGCAAGCTT	Sequencing Cas9-LP packaging Plasmid pCas9-PH-pMDLg-pRRE(D64V)
BC-PCR-Seq	ACAGCAGCTACCAATGCTGA	Sequencing BC
lvLTR-seq	ACTAGAGATCCCTCAGACCC	Sequencing LM-PCR
	NGS Sequencing	
Illu_P2 (43)	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT	Bridging oligo NGS
Illu_MPLX35 (164)	CAAGCAGAAGACGGCATACGAGAT TCTGAG GTGACTGGAGTTC	NGS sequencing: Index35
Illu_MPLX36 (165)	CAAGCAGAAGACGGCATACGAGAT CCTTGC GTGACTGGAGTTC	NGS sequencing: Index36
Illu_MPLX37 (166)	CAAGCAGAAGACGGCATACGAGAT TGGAGC GTGACTGGAGTTC	NGS sequencing: Index37
Illu_MPLX38 (167)	CAAGCAGAAGACGGCATACGAGAT TCGGGA GTGACTGGAGTTC	NGS sequencing: Index38
Illu_MPLX39 (168)	CAAGCAGAAGACGGCATACGAGAT AAACCT GTGACTGGAGTTC	NGS sequencing: Index39
Illu_MPLX40 (169)	CAAGCAGAAGACGGCATACGAGAT CTCTAC GTGACTGGAGTTC	NGS sequencing: Index40
Illu_MPLX41 (170)	CAAGCAGAAGACGGCATACGAGAT CGGCCT GTGACTGGAGTTC	NGS sequencing: Index41
Illu_MPLX42 (171)	CAAGCAGAAGACGGCATACGAGAT CCGGTG GTGACTGGAGTTC	NGS sequencing: Index42
Illu_MPLX43 (172)	CAAGCAGAAGACGGCATACGAGAT CAGCAG GTGACTGGAGTTC	NGS sequencing: Index43
Illu_MPLX44 (173)	CAAGCAGAAGACGGCATACGAGAT AAGTGC GTGACTGGAGTTC	NGS sequencing: Index44
Illu_MPLX45 (174)	CAAGCAGAAGACGGCATACGAGAT CAGGCC GTGACTGGAGTTC	NGS sequencing: Index45
Illu_MPLX46 (175)	CAAGCAGAAGACGGCATACGAGAT GGTAGA GTGACTGGAGTTC	NGS sequencing: Index46
Illu_MPLX47 (176)	CAAGCAGAAGACGGCATACGAGAT CCAGCA GTGACTGGAGTTC	NGS sequencing: Index47
Illu_MPLX48 (177)	CAAGCAGAAGACGGCATACGAGAT GCGCCA GTGACTGGAGTTC	NGS sequencing: Index48
Illu_MPLX49 (178)	CAAGCAGAAGACGGCATACGAGAT GGAACT GTGACTGGAGTTC	NGS sequencing: Index49
Illu_MPLX50 (179)	CAAGCAGAAGACGGCATACGAGAT GCGGAC GTGACTGGAGTTC	NGS sequencing: Index50
Illu_MPLX51 (180)	CAAGCAGAAGACGGCATACGAGAT CGAAAC GTGACTGGAGTTC	NGS sequencing: Index51
Illu_MPLX52 (181)	CAAGCAGAAGACGGCATACGAGAT CCACTC GTGACTGGAGTTC	NGS sequencing: Index52
ILL_Dual_P5-01	AATGATACGGCGACCACCGAGATCTACAC AGCTTAGT ACACTCTTTCCCTACACGACGCTCTTCCGATCxT	NGS sequencing: Dual Index01
ILL_Dual_P5-02	AATGATACGGCGACCACCGAGATCTACAC GCTACTTG ACACTCTTTCCCTACACGACGCTCTTCCGATCxT	NGS sequencing: Dual Index02
ILL_Dual_P5-03	AATGATACGGCGACCACCGAGATCTACAC CTAGGCAC ACACTCTTTCCCTACACGACGCTCTTCCGATCxT	NGS sequencing: Dual Index03
ILL_Dual_P5-04	AATGATACGGCGACCACCGAGATCTACAC TAGCAGCA ACACTCTTTCCCTACACGACGCTCTTCCGATCxT	NGS sequencing: Dual Index04

7.6 Plasmids

Table 7 lists original plasmids and the applications for which they were used, while plasmids generated during this thesis are listed in the cloning section (section 8.12). Maps of the cloned constructs are attached in the Appendix D-1-D-5.

Table 7: Original plasmids used for cloning and lentivirus packaging. Plasmid abbreviations were chosen based on the main characteristics (P – initial plasmid, GP – gag/pol, R – rev and V – VSV-G). The designation tSt (toxic stuffer) is as placeholder for plasmids containing different barcode fragments, which are available as barcoded constructs but are not listed below.

Abbreviation	Name	Purpose
P1	LeGO-SFFV-eGFP-tSt	plasmid for cloning
P2	LeGO-SFFV-BFP-tSt	plasmid for cloning
P3	LeGO-SFFV-TSapphire-tSt	plasmid for cloning
P4	LeGO-SFFV-Venus-tSt	plasmid for cloning
P5	LeGO-EFS-GFP-tSt	plasmid for cloning
P6	LeGO-SFFV-eBFP-smSt	plasmid for cloning
P7	pL40C-U6-EFS-Cas9-mNeon	plasmid for cloning
GP1	pMDLg-pRRE	gag/pol packaging plasmid
GP2	pMDLg-pRRE-D64V	mutated gag/pol plasmid
R1	pRSV-rev	rev packaging plasmid
V1	pCMV-VSV-G	VSV-G packaging plasmid

7.7 Software and webtools

Software or available online tools used for FACS data acquisition and analysis, gRNA prediction and target analysis, DNA measurement and imaging, cloning procedures, ddPCR, and bioinformatics analysis are listed in Table 8.

Table 8: Software and online tools used for data handling.

Software/ Online tool	Company/ Instiution
Nanodrop 2000/ 2000c Operating Software	Thermofisher Scientific (Waltham, MA, USA)
Biometra BioDocAnalyze 2.2	Analytik Jena (Jena, Germany)
QuantaSoft version 1.7	Bio-Rad (Hercules, CA, USA)
NEBase Changer online tool	NEB (Ipswich, MA, USA)
CCTop CRISPR.cos	University of Heidelberg
TIDE online tool	Netherlands Cancer Institute
FlowJo version 10.6.1	Becton, Dickinson & Company (Franklin Lakes, NJ, USA)
GraphPad PRISM version 8	GraphPad Software (San Diego, CA, USA)
CRISPOR	Sorbonne University (Paris, France) & University of California at Santa Cruz (USA)
RStudio	RStudio (Boston, USA)
BD FACSDiva 8.0.2 Software	Becton, Dickinson & Company (Franklin Lakes, NJ, USA)

7.8 Instruments

Instruments crucial for the execution and analysis of the experiments are listed in Table 9. Mainly, these instruments were used for DNA analysis, ddPCR, TBI of the animals, flow cytometry and NGS approaches. Common instruments like centrifuges, thermocyclers or incubators were excluded from the list.

Table 9: Instruments used for analysis and experimental setup. All instruments were used according to the manufacturers' instruction, while cleaning and maintenance were performed on a regular basis.

Software/ Online tool	Company/ Instiution
Nanodrop 2000c spectrophotometer	Thermofisher Scientific (Waltham, MA, USA)
Qubit 3.0 Fluorometer	Thermofisher Scientific (Waltham, MA, USA)
Biometra Biodoc GelAnalyzer	Analytik Jena (Jena, Germany)
QX200 Droplet Digital PCR system	Bio-Rad (Hercules, CA, USA)
PX1™PCR Plate Sealer	Bio-Rad (Hercules, CA, USA)
Biobeam2000 (Cs-138 source)	Eckert & Ziegler BEBIG (Berlin, Germany)
BD LSRFortessa Cell Analyzer	Becton, Dickinson & Company (Franklin Lakes, NJ, USA)
BD FACS Canto II Cell Analyzer	Becton, Dickinson & Company (Franklin Lakes, NJ, USA)
BD FACS AriaFusion Cell Sorter	Becton, Dickinson & Company (Franklin Lakes, NJ, USA)
Fragment Analyzer	Advanced Analytical Technologies (Ankeny, IA, USA)
MiSeq NGS	Illumina (San Diego, CA, USA)

7.9 Consumables

Consumables used for *in vivo* reconstitution dynamics experiments, ddPCR, and DNA measurements are listed in Table 10. Standard consumables like reaction tubes and cell culture plates were excluded from the list.

Table 10: Consumables used for ddPCR and mouse experimental procedures.

Medium/ Buffer/ Additive (conc.)	Company (City, Country Headquarter)	Cat. No.
Microvette 500 K3E	Sarstedt (Nümbrecht, Germany)	20.1341
Pre-Separation Filters (30 µm)	Miltenyi Biotec (Bergisch Gladbach, Germany)	130-041-407
LS colums	Miltenyi Biotec (Bergisch Gladbach, Germany)	130-041-407
DG8 Cartridges and Gaskets	Bio-Rad (Hercules, CA, USA)	1864007
96-well PCR plate	Eppendorf (Hamburg, Germany)	951020389
PCR Plate Heat Seal foil	Bio-Rad (Hercules, CA, USA)	1814040
Qubit assay tubes	Thermofisher (Waltham, MA, USA)	Q32856
Droplet Generation Oil for Probes	Bio-Rad (Hercules, CA, USA)	1863005
Droplet Reader Oil	Bio-Rad (Hercules, CA, USA)	1863004
CELLSTAR EASYstrainer 70 µm	Greiner BioOne (Kremsmünster, Austria)	542070
LoBind microcentrifuge tubes	Eppendorf (Hamburg, Germany)	022431021
Gene Pulser/MicroPulser Electroporation Cuvettes	Bio-Rad (Hercules, CA, USA)	165-2089
VSWP dialysis filter membranes	Merck Millipore (Burlington, VT, USA)	VSWP02500

8 Methods

8.1 DNA Isolation

Murine tissue samples for final barcode analyses were processed with QIAamp DNA Mini Kit (BM and spleen), QIAmp DNA Micro Kit (thymus and blood samples) or DNeasy Blood & Tissue Kit (cultured cells) according to the manufacturer's instructions. In brief, up to 5×10^7 cells or up to 100 µL of blood were diluted in 200 µL DPBS and mixed with 20 µL proteinase K. For efficient lysis, 200 µL AL buffer was added to the samples, mixed by pulse-vortexing, and

incubated at 56°C for 10 min. In the case of samples with low DNA amounts, namely thymus and blood, carrier RNA was added to AL before lysis at a final concentration of 10 µg/ mL.

After the addition of an equal volume of 96 % ethanol, the mixture was loaded onto separation columns, which were subsequently centrifuged at 6000 × *g* for 1 min. Columns were washed by the addition of buffers AW1 and AW2, followed by an additional centrifugation step to remove residual liquids. To ensure the removal of residual ethanol, column membranes were air-dried before elution. DNA was eluted in 200 µL (BM and spleen samples), 50 µL (thymus and blood samples), or 100 µL (cultured cells) of AE buffer.

For plasmid DNA isolation, bacterial cultures were harvested, pelleted at 3800 × *g* at 4°C for 15 min, and processed using the NucleoBond Xtra Midi Plus Kit as follows. First, bacterial cells were resuspended in RES buffer and cell lysis was carried out by the addition of 8 mL of LYS buffer. To stop cell lysis, an equal volume of neutralization buffer was added to each sample, and samples were passed through the supplied equilibrated filter before being loaded onto columns. After two washing steps for filter and columns, elution of DNA was performed with 5 mL of ELU buffer. In order to precipitate DNA from the eluate, 0.7 volumes of isopropanol were added, samples were vortexed and centrifuged at 3800 × *g* at 4°C for 30 min. The precipitate was then washed with 70 % ethanol. After centrifugation, DNA pellets were air-dried for 15 min at room temperature and dissolved in 400 µL TE buffer.

Nucleic acid concentrations were measured using a Nanodrop 2000c spectrophotometer. In the case of plasmid DNA, concentration was adjusted to 1 µg/ µL for standardized use in subsequent protocols.

8.2 Gel electrophoresis

Analysis of PCR amplicons and restriction digest products during cloning procedures was performed using 1—1.5 % agarose gels, depending on the size of the expected bands. For optimal resolution, the strength of the electrical field in the TBE buffer filled chambers was set to 10 V/ cm. Ethidium bromide was added to the melted agarose in a final concentration of 0.05 µg/ mL, and DNA was visualized under ultraviolet (UV) light using Biodoc GelAnalyzer and the corresponding gel imaging software.

If separated DNA fragments were to be used for subsequent reactions, the corresponding bands were excised using a disposable scalpel, weighted, and dissolved in three volumes QG buffer from QIAquick Gel Extraction Kit. After incubation at 50°C for 10 min under gentle shaking, one volume isopropanol was added, and samples were subsequently loaded on the spin columns and washed consecutively by addition of additional QG or PE buffer. Columns were centrifuged for 1 min at maximum speed to remove residual wash buffer. Extracted DNA was eluted using 30—50 µL of provided 10 mM Tris-Cl buffer (pH 8.5).

8.3 Cell culture

For transient transfection and viral titer determination HEK293T was used as the producer cell line. Cells were cultured in DMEM high Glucose with L-Glutamine, supplemented with 10 % heat-inactivated FBS, 1 % sodium pyruvate, 20mM HEPES, and 1 % penicillin/streptomycin at 37°C in a humidified atmosphere providing 5 % CO_2. Passaging of the cells was performed every other day at a 1:10 ratio.

The murine fibroblast cell line NIH/3T3 was used as a model for the establishment of CRISPR/Cas9-mediated knock-in experiments. This cell line has the advantage of a male genotype, which, according to precious experiments of our group, allows for optimal single barcode knock-in to the X-chromosomal *Hprt* locus (unpublished data). Cells were cultured in DMEM high Glucose with L-Glutamine, supplemented with 10 % heat-inactivated FBS and 1 % penicillin/streptomycin at 37°C, in a humidified atmosphere providing 5 % CO_2. For culture maintenance, cells were passaged at a 1:10 split ratio.

Cultivation of murine HSPCs after transduction was performed in StemSpan medium with supplements (see section 7.4) for a maximum of 24 h.

8.4 Animal maintenance, husbandry and care

All animal procedures were conducted after approval by the local authorities and were carried out in accordance with the regulations of University Medical Center Hamburg Eppendorf (UKE, Hamburg), influenced by the Animal Welfare Act (TierSchG) and the Animal Welfare Laboratory Animal Regulations (TierSchVerV). Mice (C57BL/6J) were bred in-house, i.e., in the animal facility of UKE, and housed in individually ventilated cages (IVC) under pathogen-free conditions, 12 h light:12 h dark cycles, and controlled temperature and humidity. Ad libitum access to autoclaved water and mouse chow was ensured, by controlling food supply and water bottles daily. Nesting material was provided.

For experimental procedures, mice were sex- and age-matched between 8—10 weeks and a minimum of five mice was assessed per group to allow for statistical analysis. After the onset of the experiment, mice were monitored daily and body weight was determined daily (in the first two weeks) or twice a week (from beginning of the third week). Furthermore, mice received antibiotic therapy, as described in detail in section 8.6.

8.5 HSPC isolation

In order to obtain HSPCs for SCT from BM, male mice were anesthetized with CO_2 and sacrificed by cervical dislocation. Every BM-donor group was comprised of n=30 animals to ensure a donor:recipient ratio of 1:1 for transplantation. Femur, tibia, and pelvis were excised with surgical scissors and stored on ice until further processing. Soft tissue was removed by gently rubbing the bones with a sterile paper towel. Cleaned bones were placed in DMEM, supplemented with 1 % penicillin/streptomycin at 4°C, divided into sets of six with each set

containing bones from four animals, and crushed using mortar and pestle. Cell suspensions were washed three times with DPBS and separated through 70 µm cell strainers to withhold tissue debris. After a short centrifugation step at 4°C and 300 × *g* for 5 min, cell pellets were treated with 30 mL ACK buffer for 10 min on ice to ensure complete lysis of present erythrocytes. After the addition of DPBS up to 50 mL to stop the reaction, the remaining cells were pelleted at 4°C and 300 × *g* for 5 min. Obtained whole BM cells were counted using a Neubauer chamber and used for subsequent lineage depletion. In the case of PBSC isolation, cells were first collected by retro-orbital bleeding of the mice, stained for FACS analysis (see below).

A number of 1×10^8 cells per tube were resuspended in 400 µL MACS buffer and stained with 100 µL of Direct Lineage Cell Depletion Cocktail containing magnetically labeled antibodies against lineage markers (CD5, B220, CD11b, Ly-6G/C, 7-4, and Ter-119) for 10 min at 4°C and then separated by pre-separation filters to avoid clogging of the columns before applying onto LS columns installed in the magnetic separator, which were prepared by rinsing with 3 mL MACS buffer. The flow-through containing enriched lineage-negative cell population was collected and combined with effluent from three additional washing steps, each with 3 mL of MACS buffer. After haemocytometrical determination of their number, cells which underwent preparation were used for HSPC sorting.

This process consisted of antibody staining, followed by a FACS panel, and gating the cellular subpopulations (modified according to [1]). In brief, 1 µL anti-Sca-1 PE-Cy7 antibody, 1 µL anti CD150 PE antibody, 0.7 µL anti-cKit Pacific Blue antibody and 5 µL anti-CD34 PE antibody were added per 1×10^7 lineage-negative cells and incubated for 60 min at 4°C. After cell suspensions were washed three times with FACS buffer, HSCs (Sca-1^+ cKit$^+$ CD150$^+$ CD34$^-$), MPPs (Sca-1^+ cKit$^+$ CD150$^+$ CD34$^+$), CMPs (Sca-1^- cKit$^+$ CD150$^-$ CD34$^+$) and CLPs (Sca1^{int} cKit^{int} CD150$^-$ CD34$^-$) were sorted based on their surface marker profile using a BD FACSAria Fusion Cell Sorter housed in the FACS Sorting core facility of UKE. Cell counts were determined by the sorter and to take cell loss throughout the method into account, a so-called *sorting factor* was determined, calculated by dividing the manual CLP cell count divided by the given count. Cell numbers were adjusted by multiplying with the calculated factor to ensure reliable calculations for subsequent lentiviral transduction procedures.

Sorted cell populations were transduced with four distinct LVs, assigned to one of the four available barcodes and their respective fluorescent markers, which were encoded by the same construct (Table 11). The procedure was as follows. First, cells were seeded in StemSpan growth medium in 24- or 96-well dishes, depending on the cell count and transduced with an MOI (i.e., the number of transducing lentiviral particles per cell) of 25 according to the standard transduction procedure (see section 8.16). To avoid bias effects due to vector and fluorescent markers, each respective HSPC population and BC32 vector were arbitrarily distributed in the individual mouse experiments. The following day, transduced cells were detached by washing the wells several times with DPBS and prepared for transplantation. To determine the initial transduction rates, the leftover (approximately 20 µL) of transduced cell suspension was measured at the third day (d3) post transduction via FACS.

Table 11: Assignment of different barcode backbones to fluorescent markers in the LV constructs. Four different LVs encoding GFP, eBFP, Venus or T-Sapphire fluorescent markers were combined with four distinct plasmid libraries in order to produce lentiviral particles. Positions N represent any of the bases A, T, C or G. Additionally, unique IDs consisting of known 30 bp sequences were cloned to the constructs to allow for analyses on DNA and RNA level of a known sequence. Methods of analysis are indicated below.

Fluorescent marker	Barcode backbone	ID
GFP	NNNACTNNCGANNCTTNNCGANNCTTNNGGANNCTANNACTNNCGANNCTTNNCGANNCTTNNGGANNCTANNACTNNCGANN	A
eBFP	NNNCTANNCAGNNCTTNNCGANNCTANNCTTNNGGANNCTANNCAGNNCTTNNCGANNCTANNCTTNNGGANNCTANNCAGNN	B
T-Sapphire	NNNCAGNNATCNNCTTNNCGANNGGANNCTANNCTTNNCAGNNATCNNCTTNNCGANNGGANNCTANNCTTNNCAGNNATCNN	C
Venus	NNNCGANNAGANNCTTNNCGANNCTANNGGANNCTTNNCGANNAGANNCTTNNCGANNCTANNGGANNCTTNNCGANNAGANN	D
FACS analysis	NGS analysis	Nucleic acid detection

8.6 Stem cell transplantation

Thirty female recipient animals per experiment underwent myeloablative pre-conditioning using either total body irradiation (TBI) or administration of the chemotherapeutic agent, busulfan. With respect to irradiation, mice in groups of four were exposed to 9.5 Gy at a dose rate of 0.99 Gy/ min using a ^{138}Cesium (Cs) source one day before transplantation. On the other hand, busulfan was dissolved in DMSO at 62.5 mg/ mL and, before use, the stock solution was diluted in DPBS preheated at 37°C to a working concentration of 2.5 mg/ mL. The working solution was administered by daily intraperitoneal (i.p.) injection at doses of 10 µL (i.e., 25 µg busulfan) per g of body weight for five consecutive days prior to transplantation. On the day of transplantation, BM cells from one female C57BL/6 mouse (donor) were isolated according to the standard protocol (see section 8.5) with 150 µL suspension, containing around 3×10^5 BM support cells and the transduced HSPCs. To avoid bacterial infections and support recovery, myeloablated mice received enrofloxacin (Baytril, Bayer, Leverkusen, Germany) via their drinking water (0.25 mg/ mL) for 2-4 weeks post transplantation.

8.7 Blood sampling

Starting from the sixth week (6w) post transplantation, blood samples were taken every four weeks from the submandibular vein of recipient mice via cheek punch using a lancet with 20—50 µL of blood sample being collected in EDTA-treated collection tubes. Bleeding was stopped by applying pressure with gauze to the puncture site. After lysis of red blood cells by adding 1 mL of ACK buffer, white blood cells were pelleted and stored at -20°C until further processing.

8.8 Final sampling

At defined endpoints (week one (1w), week three (3w), week eight (8w), week 16 (16w) and week 26 (26w) post transplantation), five animals per group were weighted, observed for

internal organ abnormalities, sedated using 5 % CO_2 and 95 % O_2, and sacrificed via retro-orbital bleeding using a glass capillary and EDTA-treated collection tubes. Spleens and thymuses were resected and single-cell suspensions were obtained by mincing the tissue and passing it through a 70 µm cell strainer into a sterile 50 mL conical tube. Each sample was handled separately so as to analyze the organs of each individual mouse separately. Cells were washed by adding 10 mL DPBS through the strainer and, if necessary, erythrocyte lysis was performed as described in section 8.5. Cells were counted and used for subsequent antibody staining and DNA isolation. Preparation of single-cell suspensions from BM was conducted via the standard protocol (see section 8.5).

A total amount of 2×10^6 (spleen/ thymus) or 5×10^6 (BM) cells were stained with the respective antibody panels for 20 min at 4°C. BM cells were stained with anti-Sca-1 PE-Cy7, cKit APC-Cy7, and CD150 PE in order to discriminate between the differently differentiated HSPC populations. Spleen and thymus cells were stained with B220 PE-Cy7, CD3 BV785, Ly6G PE, and CD11b APC to distinguish between the cells of the lymphoid (T and B cells) and the myeloid lineage (granulocytes and monocytes). After incubation with the antibody cocktails, cells were washed twice with FACS buffer and fixed with 1 % paraformaldehyde. Measurements were carried out using BD LSRFortessa cell analyzer, while special care was taken to keep flow rates and cell count/ s at a low level to avoid loss of single-cell information. All samples were kept on ice throughout the procedures and centrifugation was carried out at 4°C.

8.9 Droplet digital PCR

Chimerism rates, VCNs, IDs and CRISPR/Cas9 knock-in rates were quantified using the QX200 ddPCR system. To this end, 25—100 ng of DNA from mouse samples representing different time points and different organs diluted in 10 µL sterile water (DNase/ RNase free). Due to the low DNA yield of blood samples, DNA from blood cells was only used for NGS analyses and excluded from ddPCR measurements. Reactions were prepared according to Table 12 using 2× ddPCR Supermix for Probes (no dUTP). Premixed HEX- or FAM-labeled primer/ probe stocks (final concentration 900 nM/ 250 nM) for duplex amplification of target and reference locus were prepared for simplicity and used in the indicated concentrations.

Table 12: Master mix for ddPCR reaction. All reagents were used in the indicated concentrations in a final reaction volume of 25 µL. †HEX and FAM stocks include forward and reverse primer (final concentration 900 nM), as well as the dye-labeled probe (final concentration 250 nM).

	Volume [µL]	Final concentration
DNA	10	2.5-10 ng/ µL
Master mix (2x)	12.5	1x
HEX stock†	0.76	-
FAM stock†	0.76	-
$MgCl_2$ [1M]	0.07	2.8 mM
EcoRI/ HindIII	0.5	0.2 U/ µL
H_2O	ad 25	-
total	25	-

Table 13: Program for ddPCR amplification. An incubation step at 37°C was preceded before PCR amplification for restriction digest. Annealing temperatures were adapted based on the primers used for PCR reaction.

Temperature [°C]	Time [mm:ss]	Cycles
37	15:00	
95	10:00	
94	00:30	10×
60	02:00	
98	10:00	
4	pause	

For droplet generation, 20 µL of the reaction mixture was loaded in the designated wells of disposable DG8 Cartridges, followed by the addition of 70 µL Droplet Generation Oil. Rubber Gaskets were attached to the cartridges and placed in the QX200 Droplet Generator, generating up to 20,000 droplets per sample. Created droplets were pipetted into a 96-well PCR plate and heat-sealed at 180°C using Bio-Rad's PX1™PCR Plate Sealer and foil heat seal. PCR plates were placed into a TProfessional Standard PCR Thermocycler and the program was adjusted as indicated in Table 13.

After the completion of the PCR reaction, the sealed PCR plate was immediately placed into the QX200 Droplet Reader or stored overnight at 4°C. Data analysis was carried out using QuantaSoft Software Version 1.7.4 with the threshold set globally for PCR reactions with the same conditions. Since each reaction contained a reference amplicon, such as the murine erythropoietin receptor or Y-chromosomal sequences, output data (copies/ well) were normalized to the respective internal reference. Chimerism rates and ID quantification were calculated as the percentage of mEpo copies divided by two, while VCN was calculated by dividing the number of copies of Illumina sequences (in the BC32 vector) by the number of copies of Y-chromosome (engrafted donor cells). For comparisons between animals, a factor dependent on the DNA amount used for PCR reaction was included in the analysis.

8.10 Lentiviral vector production

HEK293T cells were used as producer cells of integrating (ILVs) and non-integrating (IDLVs) particles by transient transfection. One day prior to transfection, cells were seeded in 10 cm cell culture dishes at a density of 5×10^6 cells/ plate. Three plates per construct were prepared in order to obtain sufficient viral supernatant. When cells reached approximately 80 % confluence, the reagents needed for transfection were thawed and brought to room temperature. The medium was carefully changed to pre-warmed DMEM (37°C) containing 25 µM chloroquine to avoid disturbing the cell growth area. Plasmids needed for the production of third generation lentiviral particles were mixed in a molar ratio of 1:1:1:0.2 (transfer plasmid : pMDLg-pRRE or pMDLg-pRRE∆IN : pRSV-rev : pCMV-VSV-G). The amount

of transfer plasmid used per preparation was 15 µg, while the mix was diluted to a total volume of 1350 µL. Prior to calcium-phosphate precipitation, plasmids were mixed with 150 µL calcium chloride (2.5 M) and added to 1500 µL 2× HEBS buffer in a dropwise manner, under constant mixing. Plasmid/ HEBS mixtures were incubated for 15 min at room temperature and subsequently added to cell culture dishes, gently shaking. The medium was replaced after approximately 16 h of incubation at 37°C. An additional 24h later, supernatants were harvested and filtered with a disposable syringe through a 0.45 µm filter unit. Concentration of viral supernatants was performed by overnight centrifugation at 8000 × *g* and 4°C in a fixed angle rotor (Beckman Coulter Centrifuge). The supernatant was discarded and pelleted viral particles were dissolved in 1 mL medium. Aliquots of 100 µL were stored at -80°C until further use.

8.11 Viral titer estimation

Measurements of viral infectivity were performed using HEK293T cells. To this end, a total of 50,000 cells per well were seeded in 24-well plates and allowed to attach for 2—5 h. Shortly before infection with viral supernatants, the medium was changed to fresh polybrene containing medium (8 µg/ mL) to aid viral entry into cells. Cells were exposed to 10 µL, 1 µL, 0.1 µL, and 0.01 µL of viral supernatant via spin-infection (or spinoculation) at 1000 × *g* for 1 h. Experiments were performed in duplicates. The following day, the medium was exchanged with fresh medium, and cells were incubated at 37°C for another 24 h (ILV) or 36 h (IDLV), respectively. Transduction rates were measured using a BD FACS Canto II analyzed and viral titers were calculated according to the following equation

$$infectous\ units\ (IU)/mL = \frac{\%\ transduced\ cells\ P \times plated\ cells\ N}{volume\ V\ [mL]}$$

where P is the proportion of transduced cells (decimal), N is the number of plated cells (5×10^6) and V is the volume of initially added supernatant. For example, a GFP-positive cell percentage of 17 % initially infected with 1 µL supernatant corresponds to

$$infectous\ units\ \frac{IU}{mL} = \frac{0.17 \times 50{,}000}{0.001\ mL} = 8.5 \times 10^6\ IU/mL$$

Notably, cell populations showing a 5—20 % transduction rate were used for further calculations to increase the possibility for single integrations and thereby avoid calculating with cell populations harboring several integrations[290].

8.12 Cloning

For *in vivo* reconstitution dynamics experiments, the initial BC32 LeGO plasmids (P1-P4)[156] were kindly provided by Tim Aranyossy. Two modifications of the initial plasmids were performed: (i) replacement of the internal promoter SFFV to an EFS promoter and (ii) insertion of 30 nt ID sequences in order to allow DNA-based differentiation between the four constructs. Using the coding sequences of the fluorescent reporter proteins (GFP, eBFP, T-Sapphire, and Venus) for the latter, i.e., discriminating the four constructs by primers specific for each respective coding sequence, was not feasible due to their high sequence similarity. To this end, four different ID sequences (designated IDa-IDd, see Table 14) were designed based on balanced GC content (40-60 %) and reduced probability of secondary structures. Aiming to insert two copies of the same IDs to the plasmids, chosen sites were 5' of the promoter and 3' of the fluorescent marker (Figure 20A). Amplification was performed in two steps: (i) amplifying ID-EFS fragment from plasmid P5 (Figure 20B) and (ii) amplifying FP-ID fragments with the respective FPs from plasmids P1-P4 by primers with 5'-ID and restriction enzyme recognition sequence overhangs (Figure 20C). Proofreading PCR was performed using Q5 High Fidelity DNA Polymerase according to the provided instructions, resulting in ID-EFS and FP-ID fragments visualized on an agarose gel. After gel extraction (see section 8.2), fragments were combined by overlap extension (OE) PCR – possible due to an internal overlap of 12 bp between the two amplified fragments – using the forward primer of the ID-EFS and the reverse primer of the FP-ID amplification reaction.

Table 14: Plasmid ID sequences for DNA discrimination. ID sequences of 30 nt were designated a, b, c, and d and assigned to the four different plasmid constructs, here indicated by the fluorescent proteins GFP, eBFP, Venus and T-Sapphire.

ID	ID sequence	Fluorescent marker
a	TAGGATAGGATAGGATAGGACTACATCCCG	GFP
b	CTCAGCTCAGCTCAGCTCAGCTCAGCTAAA	eBFP
c	GGACAGGACAGGACAGGACAGGACAGTAAA	T-Sapphire
d	TTCAGTTCAGTTCAGTTCAGAGCCCGAGGG	Venus

The internal SFFV promoter was excised from P1-P4 by NotI/EcoRI restriction digest, and the ID-EFS-FP-ID fragments were digested using the same enzyme mix. Construct backbone and inserts were separated by agarose gel electrophoresis and then extracted from agarose gel. DNA from each excised band was eluted in 30 µL of TE buffer. Subsequently, backbone and insert were ligated at a 3:1 molar ratio using T4 DNA ligase at 22°C for 2h. Since all plasmids earmarked for barcoding experiments harbored a toxic stuffer (tSt) containing a ccdB suicide gene under the control of a lac promoter, barcoding preceding cloning and bacterial transformation had to be perfomed using a ccdB resistant bacteria strain (producer: One Shot ccdB Survival 2 T1R, Thermo Fisher Scientific, self-made batch $>5 \times 10^6$ cfu/ µg, kindly provided by Tim Aranyossy) and heat-shock induced transformation. In brief, each ligation reaction was mixed with 40 µL of bacteria and incubated for 30 min on ice. Next, a heat-shock

was induced for 30 sec, 950 µL superoptimal broth with catabolite repression (SOC) medium was added and samples were incubated for 1 h at 37°C before plating on LB-Amp plates. The next day, colonies were picked, DNA was isolated from a 2 mL culture and correct insertion of the fragments was verified by Sanger sequencing (Microsynth AG). Plasmids LE1-LE4 were used for subsequent barcoding or stored at -20°C until further use.

During the course of the practical work of the thesis, fluorescent markers emerged to be expressed at remarkably low levels, leading to difficulties in setting appropriate gating parameters for FACS analyses. Based on that observation, it was decided to replace the promoter by the stronger SFFV promoter and use these newly generated constructs for the remaining *in vivo* reconstitution dynamics experiments. To this end, plasmids LE1-LE4 (backbones) and plasmid P1 (insert) were digested with NotI/BamHI in order to exchange ID-EFS for the SFFV promoter from plasmid P1. After purification of the fragments, ligations and following bacterial transformations were performed as described above. The new constructs were Sanger-sequenced and designated LS1-LS4.

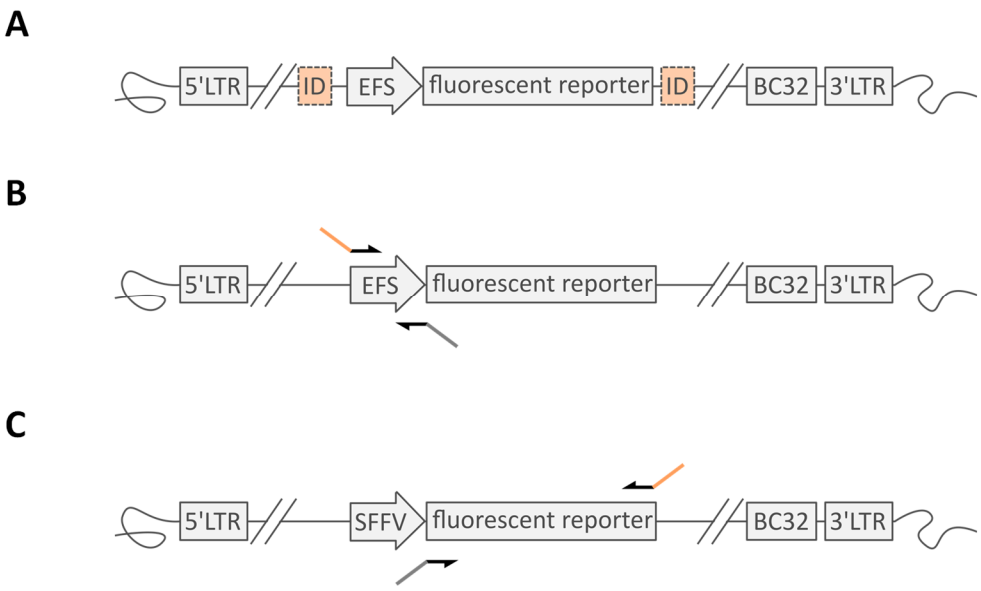


Figure 20: Cloning procedure for ID-containing BC32 constructs. A) BC32 vector backbone with the two additional ID sequences 5' of the EFS promoter and 3' of the fluorescent marker. Cloning was conducted by first amplification of ID-EFS fragment from plasmid P5 (B) and amplification of FP-ID fragments from plasmids P1-P4 (C). Primers contained sequences for the IDs and restriction enzyme recognition sequences needed for cloning as overhangs. Fragments from both PCRs were combined by overlap extension (OE) PCR using the forward primer of the ID-EFS and the reverse primer of the FP-ID amplification reaction.
BC: barcode, EFS: elongation factor 1a short, ID: identifier, SFFV: spleen focus-forming virus, LTR: long terminal repeat

CRISPR/Cas9 target plasmid P7 expressing SpCas9 under the control of an EF1α promoter connected to mNeon fluorescent protein via P2A sequence was kindly provided by Dirk Heckl[384]. The P7 plasmid was designed containing two neighboring BsmBI restriction sites 5' to a tracrRNA sequence, which facilitates cloning of chosen gRNAs by annealing short oligonucleotides (25 nt) with BsmBI overhangs. Target gRNAs were either extracted from literature for *Rosa26* locus or designed for *Hprt* locus and validated using the online tool

CCTop[256]. For each locus, two different gRNAs (see Table 15) were chosen and two complementary oligonucelotides with 5' BsmBI overhangs were generated. For cloning of the CRISPR/Cas9 target sites, the P7 plasmid was cut with BsmBI in the presence of alkaline phosphatase (FastAP) in a total volume of 30 µL at 37°C for 15 min. After separating the digested DNA fragments on an agarose gel, desired bands were excised, and DNA extracted from them was used for subsequent target ligation. Oligonucleotides for gRNA generation were annealed and were simultaneously phosphorylated using T4 polynucleotide kinase (PNK). Therefore, 1 µL of forward and reverse oligonucleotide (100 µM) were mixed with 1 µL T4 PNK in a total volume of 10 µL 1x T4 Ligase buffer. Mixtures were incubated for 45 min at 37°C, followed by incubation for 2.5 min at 95°C and subsequent cooling to 22°C in 0.1°C/ sec decrements. Annealed oligonucleotides were diluted 1:500 in water (DNase/ RNase free). In a T4 ligation reaction, 50 ng of the purified plasmid and 1 µL of diluted annealed oligonucleotides were mixed with 0.5 µL T4 DNA ligase in a total volume of 5 µL 1× T4 ligase buffer. Reactions were incubated for 90 min at 22°C and then transformed into competent *E.coli* (JM109 Mix&Go). The target Plasmids were designated T1-T4.

Table 15: CRISPR/Cas9 sgRNA sequences of target loci. For each locus two gRNA sequences (23 nt) were chosen as a target in the murine genome. Protospacer adjacent motifs (PAMs) adjacent to the gRNA are underlined and the locus target sequences are indicated.

Name	gRNA sequence (5'-3')	Target
ROSA1	ACTCCAGTCTTTCTAGAAGATGG	murine *Rosa26* intron 1 (XbaI site)
ROSA2	ATGTCTTTAATCTACCTCGATGG	murine *Rosa26* intron 1 (original provirus integration site)
HPRT1	CAACTAGAATGATCAGTCAACGG	murine *Hprt* exon 4
HPRT2	GAAGTACTCATTATAGTCAAGGG	murine *Hprt* exon 8

In order to clone CRISPR/Cas9 donor plasmids D1-D4 containing homologous regions (left and right homologous arm – LHA and RHA) in close proximity to the SpCas9 target site, left and right homology arms flanking the inserts were amplified. Since the two *Rosa26* target sites were chosen adjacent to each other, only one donor construct was needed for this locus, while the two target sites on the *Hprt* locus required different donor constructs. The HAs were derived by PCR amplification of genomic DNA from murine NIH/3T3 cells with conditions leading to an average amplicon size of 1 kb, using primers designed to contain Gibson overhangs complementary to the regions adjacent to the cut sites NotI and MauBI in the plasmids. The preparation of the master mix and the PCR program parameters were as presented in Table 16 and Table 17, respectively. Amplicon sizes were analyzed on an agarose gel, from which DNA was extracted and used for subsequent Gibson Assembly. Plasmid P6 was digested with NotI (LHA) and MauBI (RHA) to linearize the plasmid for insertion of both HAs.

Table 16: Master mix for Q5 PCR amplification of CRISPR/Cas9 donor homology arms. Six different reactions were set up, amplifying LHA and RHA for the constructs D1-D4. All reagents and were used in the indicated final concentrations.

	Volume [µL]	Final concentration
DNA template (50ng)	x	50 ng/25 µL
Q5 buffer (5x)	5	1x
primer_fw (10 µM)	1.25	0.5 µM
primer_rv (10 µM)	1.25	0.5 µM
dNTPs (10 µM)	0.5	10 µM
Q5 Polymerase	0.25	0.02 U/ µL
H_2O	ad 25	-
total	25	-

Table 17: PCR programm for Q5 PCR amplification of CRISPR/Cas9 donor homology arms. Annealing temperatures were adapted for the six different reactions, ranging 58 – 62°C.

Temperature [°C]	Time [mm:ss]	Cycles
98	00:30	
98	00:10	35×
58	00:30	
72	00:30	
72	02:00	
4	pause	

The Gibson Assembly was performed using the NEBuilder HiFi DNA Assembly Master Mix, following the manufacturer's recommendations for reactions with 4—6 fragments. Briefly, equimolar ratios of linearized backbone fragments and amplified HAs with Gibson Overhangs (0.05 pmol each) were mixed in three separate reactions, diluted to a final volume of 10 µL of Gibson Assembly Master Mix, reactions were incubated for 1 h at 50°C and then transformed into chemically competent NEB 5-alpha strain according to the protocol mentioned above. Transformed bacteria were plated on LB-Amp plates and incubated overnight at 37°C. Early the next day, colonies were picked, small-scale purification of DNA (miniprep) was performed and the DNA was used for subsequent sequence validation via Sanger sequencing.Next, CRISPR/Cas9 donor plasmids containing gRNAs under the control of a U6 promoter were generated (termed DG1-DG4). To this end, expression cassettes of U6-gRNAs were amplified by Q5 PCR from the previously cloned plasmids T1-T4 using with primers containing Gibson overhangs complementary to the sequences adjacent to the ApaI cut site in the corresponsing D1-D4 donor plasmids. In order to perform the Gibson assembly, D1-D4 plasmids were linearized using ApaI plus FastAP to give the dephosphorylated backbone, which was then mixed with the U6-gRNA cassettes in a molar ratio of 2:1, with a total amount of DNA equeal to 0.2 pmol. Gibson Assembly and subsequent bacterial transformation were performed as mentioned above using a shortened incubation time of 15 min. The produced constructs were isolated, verified by Sanger sequencing, and stored at -20°C.

Site-directed mutagenesis of the viral gag/pol plasmid GP1 was performed at position 64 of the coding region for integrase sequence, changing aspartate to valine and thus mutating the catalytic core domain of the integrase. Therefore, two primers in back-to-back orientation were designed, harbouring the mutation at the specified position. Next, a Q5 site-directed mutagenesis (NEB) reaction was performed according to the following scheme (see Table 18 and Table 19).

Table 18: Master mix for Q5 site-directed mutagenesis of lentiviral gag/pol plasmid. All reagents were used in the indicated concentrations.

	Volume [µL]	Final concentration
Template DNA	1	1 ng/ µL
Q5 Hot Start HF 2x MM	12.5	1x
primer_fw (10 µM)	1.25	0.5 µM
primer_rv (10 µM)	1.25	0.5 µM
H_2O	ad 25	-
total	25	-

Table 19: PCR program for Q5 site-directed mutagenesis. Annealing time (30s/ kb) was adjusted to the plasmid length (≈9 kb).

Temperature [°C]	Time [mm:ss]	Cycles
98	00:30	
98	00:10	25×
66	00:30	
72	04:30	
72	02:00	
4	pause	

After the completion of the reaction, 1 µL of the PCR product was used in a subsequent KLD reaction using 1 µL KLD Enzyme Mix in a total volume of 10 µL 1× KLD reaction buffer, supplied with the indicated kit. The reaction was incubated for 5 min at room temperature and the ligated product was used for transformation into chemically competent *E.coli* (strain Mix&Go JM109). The following day, plated colonies were used for growing a 2 mL culture in LB-Amp medium, from which DNA was isolated and Sanger-sequenced with a mutation site-specific primer. The resulting plasmid was designated GP2.

All plasmid abbreviations and the complete names are listed in Table 20, with short notifications on the used purpose. The corresponding plasmid maps of the final constructs are attached in the Appendix D-1-D-5.

Table 20: Initial and generated plasmids used for barcoding and CRISPR/Cas9 experiments. Plasmid abbreviations were chosen based on main characteristics (P – initial plasmid, LE – LeGO-EFS, LS – LeGO-SFFV, T – Target, TC – Target SpCas9 only, D – Donor, DG – Donor plus gRNA and GP – gag/pol). Plasmids available with the tSt placeholder for subsequent barcode fragments were all available as barcoded constructs, but are not listed below.

Abbreviation	Name	Purpose
P1	LeGO-SFFV-eGFP-tSt	plasmid for cloning
P2	LeGO-SFFV-BFP-tSt	plasmid for cloning
P3	LeGO-SFFV-TSapphire-tSt	plasmid for cloning
P4	LeGO-SFFV-Venus-tSt	plasmid for cloning
P5	LeGO-EFS-GFP-tSt	plasmid for cloning
P6	LeGO-SFFV-eBFP-smSt	plasmid for cloning
P7	pL40C-U6-EFS-Cas9-mNeon	plasmid for cloning
LE1	LeGO-IDa-EFS-eGFP-IDa-tSt	mouse experiments: BM-TBI, BM-CTx
LE2	LeGO-IDb-EFS-BFP-IDb-tSt	mouse experiments: BM-TBI, BM-CTx
LE3	LeGO-IDc-EFS-TSapphire-IDc-tSt	mouse experiments: BM-TBI, BM-CTx
LE4	LeGO-IDd-EFS-Venus-IDd-tSt	mouse experiments: BM-TBI, BM-CTx
LS1	LeGO-IDa-SFFV-eGFP-IDa-tSt	mouse experiments: BM-TBI^{SFFV}, PB-TBI
LS2	LeGO-IDb-SFFV-BFP-IDb-tSt	mouse experiments: BM-TBI^{SFFV}, PB-TBI
LS3	LeGO-IDc-SFFV-TSapphire-IDc-tSt	mouse experiments: BM-TBI^{SFFV}, PB-TBI
LS4	LeGO-IDd-SFFV-Venus-IDd-tSt	mouse experiments: BM-TBI^{SFFV}, PB-TBI
T1	pL40C-U6-ROSA1-EFS-Cas9-mNeon	CRISPR/Cas9 target plasmid
T2	pL40C-U6-ROSA2-EFS-Cas9-mNeon	CRISPR/Cas9 target plasmid
T3	pL40C-U6-HPRT1-EFS-Cas9-mNeon	CRISPR/Cas9 target plasmid
T4	pL40C-U6-HPRT2-EFS-Cas9-mNeon	CRISPR/Cas9 target plasmid
TC1	pL40C-EFS-Cas9-mNeon	CRISPR/Cas9 target plasmid w/o gRNA
D1/ D2	LeGO-SFFV-eBFP-smSt-ROSA26HA-G161A	CRISPR/Cas9 donor plasmid
D3	LeGO-SFFV-eBFP-smSt-HPRT1HA	CRISPR/Cas9 donor plasmid
D4	LeGO-SFFV-eBFP-smSt-HPRT2HA	CRISPR/Cas9 donor plasmid
DG1	LeGO-U6-ROSA1-SFFV-eBFP-smSt-ROSA26HA-G161A	CRISPR/Cas9 donor plasmid w/ gRNA
DG2	LeGO-U6-ROSA2-SFFV-eBFP-smSt-ROSA26HA-G161A	CRISPR/Cas9 donor plasmid w/ gRNA
DG3	LeGO-U6-HPRT1-SFFV-eBFP-smSt-HPRT1HA	CRISPR/Cas9 donor plasmid w/ gRNA
DG4	LeGO-U6-HPRT2-SFFV-eBFP-smSt-HPRT2HA	CRISPR/Cas9 donor plasmid w/ gRNA
GP1	pMDLg-pRRE	gag/pol plasmid
GP2	pMDLg-pRRE-D64V	mutated gag/pol plasmid

8.13 Generation of barcode libraries

Barcode libraries were generated according to the optimized BC32 protocol[156]. First, random double-stranded barcode oligonucleotides with the indicated architecture (Table 11) were generated in four separated batches using proofreading Q5 High-Fidelity polymerase (NEB) and specific primers in a TProfessional Standard PCR Thermocycler (Biometra). For each barcode backbone, eight separate batches were prepared, with reaction parameters adjusted to the manufacturer's instructions (Table 21 and Table 22). After the completion of the PCR, two reactions of double-stranded barcodes were pooled. DNA was purified using QIAquick PCR Purification Kit (Qiagen) according to the given protocol and eluted in 50 µL pre-warmed (37°C) EB buffer.

Table 21: Master mix for barcode library generation. All reagents were used in the indicated concentrations in a final reaction volume of 25 µL. † FP is a placeholder for the different barcode oligonucleotides assigned to the four fluorescent proteins GFP, eBFP, T-Sapphire, and Venus.

	Volume [µL]	Final concentration
Poly-MM-FP† (100ng)	x	100 ng/25 µL
Q5 buffer (5x)	5	1x
Mre32BC_fw (10 µM)	0.76	0.5 µM
MauBI32BC_rv (10 µM)	0.76	0.5 µM
dNTPs (10 mM)	0.5	10 mM
Q5 Polymerase	0.25	0.02 U/ µL
H_2O	ad 25	-
total	25	-

Table 22: Program for barcode generation. A single PCR program with ten cycles of repetition was used independently of barcode backbone.

Temperature [°C]	Time [mm:ss]	Cycles
98	00:30	
98	00:10	
66	00:20	10×
72	00:20	
72	03:00	
4	pause	

Notably, the BC32 vector constructs were designed to contain a toxic stuffer fragment (≈2.6 kb) as a placeholder for barcode sequences to be inserted. This allows for visual discrimination of restricted construct backbones on an agarose gel and also provides a positive selection system for bacteria transformed with successfully barcoded constructs, is its presence prevents most bacterial strains from growing if present. Barcodes and BC32 vectors were cut and linearized using MreI and MauBI FastDigest enzymes (Thermofisher). Specifically, 5 µg of vector was digested using 1.5 µL of each restriction enzyme in a total volume of 30 µL. Samples were incubated for 30 min at 37°C, followed by the addition of 1 U FastAP enzyme (Thermofisher) and incubation for another hour at 37°C to allow 5' dephosphorylation. Subsequently, digested products were separated on a 1.5 % agarose gel and extracted (see section 8.2). In order to estimate DNA concentration from barcodes as accurately as possible, the Qubit dsDNA HS Assay kit was used. Briefly, Qubit dsDNA Reagent was diluted 1:200 in Qubit dsDNA HS Buffer to prepare the working solution, to 199 µL of which 1 µL of barcode sample was added. After rigorous mixing and 2 min incubation at room temperature, the DNA concentration was measured in the Qubit® 3.0 Fluorometer.

After quantification, digested plasmid backbones and barcodes were ligated at a 1:3 molar ratio in a 20 µL T4 DNA Ligase (30 U/ µL, Thermofisher) reaction incubated at 16°C overnight, with the quantity of the backbone DNA kept at 500 ng. The next morning, ligase was heat-inactivated for 10 min at 65°C. As the ingredients of the ligation buffer may reduce the efficiency of the subsequent electroporation, the completed reaction was first passed through 0.025 µm VSWP dialysis filter membranes for 1.5 h. Dialyzed barcoded plasmid constructs were then transformed in electrocompetent *E.coli* (NEB 5 alpha strain) via electroporation. In brief, cuvettes were pre-cooled on ice and 40 µL of bacteria were mixed with ligation products and the mixture was pipetted in the electroporation cuvettes, which were then placed into the electroporator (Biorad). The electric pulse parameters were set at 1.8 kV, 200 Ohm, and 25 µF. Transformed bacteria were flushed with pre-warmed recovery medium and incubated at 37°C and 800 rpm for 1 h. Afterwards, 10 µL of serially diluted suspensions of transformed bacterial (1:10—1:10000) was plated on LB-Amp agar plates. The residual bacterial stock was used for midiprep and plasmid isolation (see section 8.1).

The number of possible barcodes per construct (i.e., the complexity of the library) was calculated by counting the colonies after overnight incubation of the LB-Amp agar plates using the following formula proposed by Bystrykh et al.[159]:

$$\#BCs = \frac{\#\ colonies \times transformation\ volume\ (1000\mu L)}{plated\ sample\ volume\ [\mu L]}$$

Given a number of 150 colonies for the plated sample volume of 0.01 µL would result in a theoretical complexity of 1.5×10^7 barcodes. In order to assess the quality of barcode libraries, ten colonies were selected for plasmid DNA isolation for each barcoded construct. Target sites for barcode integration were Sanger-sequenced and barcode and Illumina sequences contained in the BC32 vectors were checked for correct insertion and absence of mutations.

8.14 Next-generation sequencing and bioinformatics

After confirming that murine gDNA purified from samples collected at the last stages of the *in vivo* reconstitution dynamics experiments was adequate for deep sequencing in terms of both quantity and quality, aliquots of 200 ng dissolved in 21.8 µL underwent PCR using primer combinations designed for the BC32 backbone. By the use of up to 72 possible different primer combinations in total, simultaneous multiplexing of an equal amount of samples on an NGS flow cell was made possible. To this end, four different DUAL primers (DUAL_P5-01–DUAL_P5-04) were combined in single reactions with 18 possible MPLX primers (MPLX-35–MPLX-52), resulting in a unique, identifiable pattern per sample and PCR reaction. The preparation and the conditions of the PCR reaction are shown in Table 23 and Table 24, respectively.

Table 23: Master mix for Illumina PCR library generation. PCR reactions were set up according to the following conditions and different combinations of MLPX and DUAL primers were used. Bridging oligo was added to aid binding of the primers and EcoRI enzyme was added to fragment DNA before PCR reaction.

	Volume [µL]	Final concentration
DNA (200 ng)	21.8	40 ng/ µL
Multiplex PCR Plus 2× MM	25	1x
MLPX primer (10 µM)	1	0.2 µM
DUAL primer (10 µM)	1	0.2 µM
Illu_P2 bridging oligo (1 µM)	0.2	4 nM
EcoRI	1	-
total	50	-

Table 24: PCR program for Illumina library PCR. Before standard PCR amplification, a prior incubation step at 37°C for 15 min was preceded, to allow the restriction reaction of the template to take place.

Temperature [°C]	Time [mm:ss]	Cycles
37	15:00	
95	05:00	
95	00:30	30×
57	00:30	
72	00:30	
68	10:00	
4	pause	

Completed PCR reactions were individually purified using magnetic XP-beads and a magnetic bead separator. Samples were pipetted in DNA LoBind microcentrifuge tubes, to which beads were added in a 1.8-fold excess over sample (90 µL). Afterwards, tubes were placed in a magnetic separator for >5 min to allow beads to attach to the tube walls adjacent to the magnet. After removing the solution, beads were washed twice by resuspension in 200 µL of 70 % ethanol. After the second washing step, samples were dried for 5-10 min to ensure evaporation of residual ethanol and DNA was eluted from the beads by 15 µL of TE Buffer (pH 8.0) and a 1 µL aliquot was spent for concentration measurement using the Qubit dsDNA HS Assay as described in section 8.13. In total, 450 ng of DNA per flow cell was premixed from approximately 60 samples, i.e., ≈7.5 ng/ sample. In the case of lower DNA concentrations, the complete sample (14 µL) was added. The pooled sample, consisting of 60 differently labeled PCR reactions, was again purified using the magnetic XP-beads, according to the abovementioned protocol. The eluate containing purified DNA was transferred to a new DNA LoBind tube and its concentration was determined by the Qubit dsDNA HS assay. Quality assessment and NGS run of the samples was performed by the Dresden – Concept Genome Center. The quality level was determined by the use of a fragment analyzer (Advanced Analytical Technologies, Ankeny, USA) prior to performing MiSeq NGS (Illumina MiSeq System) with a single read sequencing length of 83 bp, containing 20 % PhiX reference genome.

Preprocessing filtering and quality control of NGS data were performed in collaboration with Technische Universität Dresden as described before[156,172] by Lars Thielecke. PCR noise was filtered by the pipeline by only using sequencing reads with a Phred Score of at least 30. Unique barcodes with sequencing reads of one were omitted from analysis, to exclude noise from low-abundant clones. Error correction was based on the hamming distance of HD=8 as a PCR error correction threshold to compensate for the combination of highly similar backbones (up to eight nucleotides), treating them as derivatives of the same original barcode. Resulting output data were barcode sequences and respective frequencies per animal and organ, total read counts per sample and total number of barcodes per sample. Further analyses and visualization of the data were performed using customized R scripts (original script kindly provided by Tim Aranyossy) and adjusted to the data and visualization method[172].

8.15 Validation of SpCas9 editing efficiency

Evaluation of CRISPR/Cas9 editing efficiency was performed in NIH/3T3 bulk cell populations transduced with viral CRISPR/Cas9 target vector supernatants. To this end, NIH/3T3 cells were transduced according to the standard lentiviral transduction protocol (see section 8.16) and three days post transduction gDNA was isolated (see section 8.1). A parallel target site-specific PCR of transduced and non-transduced control cells was performed, resulting in PCR products approximately 450-520 bp in length. Both sequences were Sanger-sequenced and the derived quantitative sequence trace data files were used for TIDE online tool[341]. The relative abundance of indels within a chosen size range was estimated and R^2 values as a goodness-of-fit measure were determined for each sample.

8.16 Lentiviral transduction

For lentiviral (co-)transduction, a total amount of 1-5 × 10^5 NIH/3T3 cells were seeded in 24-well cell culture plates and incubated for 4—6 h, to allow the cells to attach to the plate surface. Shortly before transduction, the medium was replaced by fresh, polybrene-containing, pre-warmed (37°C) medium. Dependent on viral titers, MOIs of 10—30 were used to infect cells, and transductions were performed in duplicates. Co-transductions were performed in an equal manner and indicated MOIs of both viral constructs were added to the cells consecutively. After gentle shaking, spin-infection was performed by centrifuging the plates at 1000 × *g* for 1 h at room temperature. Three days post transduction, fluorescent protein expression was measured in a BD FACS Canto II to determine ratios of fluorescent marker expressing cells.

8.17 Transient transfection

For transient transfection of CRISPR/Cas9 components to NIH/3T3 target cells, a total number of 1.5 × 10^5 cells were seeded in 6-well plates and incubated at 37°C prior to transfection. When cells attached to the plastic bottom of the well, reaction mixture was prepared, by mixing 2 µg of plasmid to be transfected with 200 µL jetOPTIMUS buffer, and 2 µL jetOPTIMUS reagent, supplied with the jetOPTIMUS DNA transfection Reagent Kit (Polyplus transfection, New York City, NY, USA). In case of co-transfection of donor and target plasmids, 1 µg of each plasmid was used. Reaction mixture was vortexed, spun down, and incubated for 10 min at room temperature. After incubation, cells were transfected by carefully adding the reaction mixture dropwise to the seeded target cells. Finally, medium was replaced 4 h after transfection and transfection rates were evaluated by FACS analysis at 24 or 48 h post transfection.

8.18 CRISPR/Cas9-mediated knock-in validation

To assess CRISPR/Cas9-mediated knock-in efficiencies after transient transfection of IDLV transduction, NIH/3T3 cells were co-transduced or –transfected with target and donor IDLVs/plasmids for the respective loci (ROSA1, ROSA2, HPRT1, or HPRT2). Delivery of the CRISPR/Cas9 components was performed according to the standard lentiviral transduction protocol (see section 8.11) using MOIs of five for each construct or co-transfected as described above (see section 8.17). Three days post transduction or 24/ 48 h post transfection, cells were sorted using BD Aria Fusion sorter based on their fluorescent marker expression and seeded in 24- or 12-well cell culture plates, depending on total cell amounts. At selected time points post delivery (3d, 7d, 10d, 21d, and 56d) FACS analysis was performed to detect stable fluorescent marker expression (eBFP) over time, and DNA was extracted on day 7 (7d) and day 21 (21d) for subsequent quantification of knock-in events. Insert-specific DreamTaq Polymerase PCR was performed in order to verify CRISPR/Cas9-mediated knock-in at the

indicated locus. Quantification of knock-in events was performed 21d post delivery by ddPCR (see section 8.9). This was performed by quantifying non-edited WT sequences of *Rosa26* and *Hprt* loci. Knock-in rates were calculated by subtracting the amount of WT sequence from total mEpo copies in the reaction.

For single-cell analyses, three days post transduction single mNeon$^+$ cells were sorted using the BD Aria Fusion sorter into four 96-well plates, allowing for the collection of cells with transient expression of SpCas9 protein. Sorted cells were cultivated and, upon reaching confluence, split and seeded in 24- or 12-well plates. Approximately 4 weeks after single-cell sorting, fluorescent protein expression (mNeon and eBFP) of clones was measured using the BD FACS Canto II. In case of presence of eBFP and absence of mNeon expression, clones were counted as harboring putative CRISPR/Cas9 knock-in events, since eBFP was stably expressed and DNA was isolated (see section 8.1). Insert- and barcode-specific DreamTaq Polymerase PCR was performed in order to verify CRISPR/Cas9-mediated knock-in at the specific site and analyze the abundance of barcode sequences. Total knock-in efficiency was calculated from the amount of eBFP$^+$ cells measured via FACS divided by the frequency of samples resulting in a knock-in- and barcode-specific band in PCR analyses.

9 Bibliography

1 Doulatov, S., Notta, F., Laurenti, E. & Dick, J. E. Hematopoiesis: a human perspective. *Cell Stem Cell* **10**, 120-136, doi:10.1016/j.stem.2012.01.006 (2012).

2 McCulloch, E. A. & Till, J. E. Perspectives on the properties of stem cells. *Nat Med* **11**, 1026-1028, doi:10.1038/nm1005-1026 (2005).

3 Kondo, M., Weissman, I. L. & Akashi, K. Identification of clonogenic common lymphoid progenitors in mouse bone marrow. *Cell* **91**, 661-672, doi:10.1016/s0092-8674(00)80453-5 (1997).

4 Morrison, S. J., Wandycz, A. M., Hemmati, H. D., Wright, D. E. & Weissman, I. L. Identification of a lineage of multipotent hematopoietic progenitors. *Development* **124**, 1929-1939 (1997).

5 Akashi, K., Traver, D., Miyamoto, T. & Weissman, I. L. A clonogenic common myeloid progenitor that gives rise to all myeloid lineages. *Nature* **404**, 193-197, doi:10.1038/35004599 (2000).

6 Manz, M. G., Miyamoto, T., Akashi, K. & Weissman, I. L. Prospective isolation of human clonogenic common myeloid progenitors. *Proc Natl Acad Sci U S A* **99**, 11872-11877, doi:10.1073/pnas.172384399 (2002).

7 Morrison, S. J., Uchida, N. & Weissman, I. L. The biology of hematopoietic stem cells. *Annu Rev Cell Dev Biol* **11**, 35-71, doi:10.1146/annurev.cb.11.110195.000343 (1995).

8 Reya, T. Regulation of hematopoietic stem cell self-renewal. *Recent Prog Horm Res* **58**, 283-295, doi:10.1210/rp.58.1.283 (2003).

9 Reya, T., Morrison, S. J., Clarke, M. F. & Weissman, I. L. Stem cells, cancer, and cancer stem cells. *Nature* **414**, 105-111, doi:10.1038/35102167 (2001).

10 Orkin, S. H. Diversification of haematopoietic stem cells to specific lineages. *Nat Rev Genet* **1**, 57-64, doi:10.1038/35049577 (2000).

11 Dick, J. E. Stem cells: Self-renewal writ in blood. *Nature* **423**, 231-233, doi:10.1038/423231a (2003).

12 Benveniste, P. *et al.* Intermediate-term hematopoietic stem cells with extended but time-limited reconstitution potential. *Cell Stem Cell* **6**, 48-58, doi:10.1016/j.stem.2009.11.014 (2010).

13 Cheshier, S. H., Morrison, S. J., Liao, X. & Weissman, I. L. In vivo proliferation and cell cycle kinetics of long-term self-renewing hematopoietic stem cells. *Proc Natl Acad Sci U S A* **96**, 3120-3125, doi:10.1073/pnas.96.6.3120 (1999).

14 Kiel, M. J. *et al.* Haematopoietic stem cells do not asymmetrically segregate chromosomes or retain BrdU. *Nature* **449**, 238-242, doi:10.1038/nature06115 (2007).

15 Osawa, M., Hanada, K., Hamada, H. & Nakauchi, H. Long-term lymphohematopoietic reconstitution by a single CD34-low/negative hematopoietic stem cell. *Science* **273**, 242-245, doi:10.1126/science.273.5272.242 (1996).

16 Morrison, S. J. & Weissman, I. L. The long-term repopulating subset of hematopoietic stem cells is deterministic and isolatable by phenotype. *Immunity* **1**, 661-673, doi:10.1016/1074-7613(94)90037-x (1994).

17 Yang, L. *et al.* Identification of Lin(-)Sca1(+)kit(+)CD34(+)Flt3- short-term hematopoietic stem cells capable of rapidly reconstituting and rescuing myeloablated transplant recipients. *Blood* **105**, 2717-2723, doi:10.1182/blood-2004-06-2159 (2005).

18 Luc, S., Buza-Vidas, N. & Jacobsen, S. E. Delineating the cellular pathways of hematopoietic lineage commitment. *Semin Immunol* **20**, 213-220, doi:10.1016/j.smim.2008.07.005 (2008).

19 Ema, H., Morita, Y. & Suda, T. Heterogeneity and hierarchy of hematopoietic stem cells. *Exp Hematol* **42**, 74-82 e72, doi:10.1016/j.exphem.2013.11.004 (2014).

20 Spangrude, G. J., Heimfeld, S. & Weissman, I. L. Purification and characterization of mouse hematopoietic stem cells. *Science* **241**, 58-62, doi:10.1126/science.2898810 (1988).

21 Challen, G. A., Boles, N., Lin, K. K. & Goodell, M. A. Mouse hematopoietic stem cell identification and analysis. *Cytometry A* **75**, 14-24, doi:10.1002/cyto.a.20674 (2009).

22 Ikuta, K. & Weissman, I. L. Evidence that hematopoietic stem cells express mouse c-kit but do not depend on steel factor for their generation. *Proc Natl Acad Sci U S A* **89**, 1502-1506, doi:10.1073/pnas.89.4.1502 (1992).

23 Okada, S. *et al.* In vivo and in vitro stem cell function of c-kit- and Sca-1-positive murine hematopoietic cells. *Blood* **80**, 3044-3050 (1992).

24 Uchida, N. & Weissman, I. L. Searching for hematopoietic stem cells: evidence that Thy-1.1lo Lin- Sca-1+ cells are the only stem cells in C57BL/Ka-Thy-1.1 bone marrow. *J Exp Med* **175**, 175-184, doi:10.1084/jem.175.1.175 (1992).

25 Bryder, D., Rossi, D. J. & Weissman, I. L. Hematopoietic stem cells: the paradigmatic tissue-specific stem cell. *Am J Pathol* **169**, 338-346, doi:10.2353/ajpath.2006.060312 (2006).

26 Mayle, A., Luo, M., Jeong, M. & Goodell, M. A. Flow cytometry analysis of murine hematopoietic stem cells. *Cytometry A* **83**, 27-37, doi:10.1002/cyto.a.22093 (2013).

27 Kiel, M. J. *et al.* SLAM family receptors distinguish hematopoietic stem and progenitor cells and reveal endothelial niches for stem cells. *Cell* **121**, 1109-1121, doi:10.1016/j.cell.2005.05.026 (2005).

28 Morita, Y., Ema, H. & Nakauchi, H. Heterogeneity and hierarchy within the most primitive hematopoietic stem cell compartment. *J Exp Med* **207**, 1173-1182, doi:10.1084/jem.20091318 (2010).

29 Balazs, A. B., Fabian, A. J., Esmon, C. T. & Mulligan, R. C. Endothelial protein C receptor (CD201) explicitly identifies hematopoietic stem cells in murine bone marrow. *Blood* **107**, 2317-2321, doi:10.1182/blood-2005-06-2249 (2006).

30 Kraus, A. B. *et al.* Early host CD8 T-cell recovery and sensitized anti-donor interleukin-2-producing and cytotoxic T-cell responses associated with marrow graft rejection following nonmyeloablative allogeneic bone marrow transplantation. *Exp Hematol* **31**, 609-621, doi:10.1016/s0301-472x(03)00082-1 (2003).

31 Zhao, Y., Swenson, K., Sergio, J. J. & Sykes, M. Pig MHC mediates positive selection of mouse CD4+ T cells with a mouse MHC-restricted TCR in pig thymus grafts. *J Immunol* **161**, 1320-1326 (1998).

32 Tomita, Y., Sachs, D. H. & Sykes, M. Myelosuppressive conditioning is required to achieve engraftment of pluripotent stem cells contained in moderate doses of syngeneic bone marrow. *Blood* **83**, 939-948 (1994).

33 Sykes, M. *et al.* Effects of T cell depletion in radiation bone marrow chimeras. III. Characterization of allogeneic bone marrow cell populations that increase allogeneic chimerism independently of graft-vs-host disease in mixed marrow recipients. *J Immunol* **143**, 3503-3511 (1989).

34 Szilvassy, S. J., Humphries, R. K., Lansdorp, P. M., Eaves, A. C. & Eaves, C. J. Quantitative assay for totipotent reconstituting hematopoietic stem cells by a competitive repopulation strategy. *Proc Natl Acad Sci U S A* **87**, 8736-8740, doi:10.1073/pnas.87.22.8736 (1990).

35 McDermott, S. P., Eppert, K., Lechman, E. R., Doedens, M. & Dick, J. E. Comparison of human cord blood engraftment between immunocompromised mouse strains. *Blood* **116**, 193-200, doi:10.1182/blood-2010-02-271841 (2010).

36 Cheng, H., Zheng, Z. & Cheng, T. New paradigms on hematopoietic stem cell differentiation. *Protein Cell* **11**, 34-44, doi:10.1007/s13238-019-0633-0 (2020).

37 Busch, K. & Rodewald, H. R. Unperturbed vs. post-transplantation hematopoiesis: both in vivo but different. *Curr Opin Hematol* **23**, 295-303, doi:10.1097/MOH.0000000000000250 (2016).

38 Velten, L. *et al.* Human haematopoietic stem cell lineage commitment is a continuous process. *Nat Cell Biol* **19**, 271-281, doi:10.1038/ncb3493 (2017).

39 Dykstra, B. *et al.* Long-term propagation of distinct hematopoietic differentiation programs in vivo. *Cell Stem Cell* **1**, 218-229, doi:10.1016/j.stem.2007.05.015 (2007).

40 Muller-Sieburg, C. E., Cho, R. H., Karlsson, L., Huang, J. F. & Sieburg, H. B. Myeloid-biased hematopoietic stem cells have extensive self-renewal capacity but generate diminished lymphoid progeny with impaired IL-7 responsiveness. *Blood* **103**, 4111-4118, doi:10.1182/blood-2003-10-3448 (2004).

41 Muller-Sieburg, C. E., Cho, R. H., Thoman, M., Adkins, B. & Sieburg, H. B. Deterministic regulation of hematopoietic stem cell self-renewal and differentiation. *Blood* **100**, 1302-1309 (2002).

42 Benz, C. *et al.* Hematopoietic stem cell subtypes expand differentially during development and display distinct lymphopoietic programs. *Cell Stem Cell* **10**, 273-283, doi:10.1016/j.stem.2012.02.007 (2012).

43 Yamamoto, R. *et al.* Clonal analysis unveils self-renewing lineage-restricted progenitors generated directly from hematopoietic stem cells. *Cell* **154**, 1112-1126, doi:10.1016/j.cell.2013.08.007 (2013).

44 Carrelha, J. *et al.* Hierarchically related lineage-restricted fates of multipotent haematopoietic stem cells. *Nature* **554**, 106-111, doi:10.1038/nature25455 (2018).

45 Adolfsson, J. *et al.* Identification of Flt3+ lympho-myeloid stem cells lacking erythro-megakaryocytic potential a revised road map for adult blood lineage commitment. *Cell* **121**, 295-306, doi:10.1016/j.cell.2005.02.013 (2005).

46 Wilson, A. *et al.* Hematopoietic stem cells reversibly switch from dormancy to self-renewal during homeostasis and repair. *Cell* **135**, 1118-1129, doi:10.1016/j.cell.2008.10.048 (2008).

47 Pietras, E. M. *et al.* Functionally Distinct Subsets of Lineage-Biased Multipotent Progenitors Control Blood Production in Normal and Regenerative Conditions. *Cell Stem Cell* **17**, 35-46, doi:10.1016/j.stem.2015.05.003 (2015).

48 REKERS, P. E., COULTER, M. P. & WARREN, S. L. EFFECT OF TRANSPLANTATION OF BONE MARROW INTO IRRADIATED ANIMALS. *Archives of Surgery* **60**, 635-667, doi:10.1001/archsurg.1950.01250010656001 (1950).

49 Barnes, D. W., Corp, M. J., Loutit, J. F. & Neal, F. E. Treatment of murine leukaemia with X rays and homologous bone marrow; preliminary communication. *Br Med J* **2**, 626-627, doi:10.1136/bmj.2.4993.626 (1956).

50 Thomas, E. D., Lochte, H. L., Jr., Lu, W. C. & Ferrebee, J. W. Intravenous infusion of bone marrow in patients receiving radiation and chemotherapy. *N Engl J Med* **257**, 491-496, doi:10.1056/NEJM195709122571102 (1957).

51 Hansen, J. A. *et al.* Transplantation of marrow from an unrelated donor to a patient with acute leukemia. *N Engl J Med* **303**, 565-567, doi:10.1056/NEJM198009043031007 (1980).

52 Appelbaum, F. R. Consolidation chemotherapy prior to hematopoietic cell transplantation for adults with acute myeloid leukemia in first remission. *Best Pract Res Clin Haematol* **29**, 365-371, doi:10.1016/j.beha.2016.10.012 (2016).

53 Daikeler, T. *et al.* Allogeneic hematopoietic SCT for patients with autoimmune diseases. *Bone Marrow Transplant* **44**, 27-33, doi:10.1038/bmt.2008.424 (2009).

54 Hirano, M. *et al.* Allogeneic stem cell transplantation corrects biochemical derangements in MNGIE. *Neurology* **67**, 1458-1460, doi:10.1212/01.wnl.0000240853.97716.24 (2006).

55 Gyurkocza, B. & Sandmaier, B. M. Conditioning regimens for hematopoietic cell transplantation: one size does not fit all. *Blood* **124**, 344-353, doi:10.1182/blood-2014-02-514778 (2014).

56 van Kempen-Harteveld, M. L. *et al.* Cataract-free interval and severity of cataract after total body irradiation and bone marrow transplantation: influence of treatment parameters. *Int J Radiat Oncol Biol Phys* **48**, 807-815, doi:10.1016/s0360-3016(00)00669-6 (2000).

57 Socie, G. *et al.* New malignant diseases after allogeneic marrow transplantation for childhood acute leukemia. *J Clin Oncol* **18**, 348-357, doi:10.1200/JCO.2000.18.2.348 (2000).

58 Gooptu, M. *et al.* A Comparison of the Myeloablative Conditioning Regimen Fludarabine/Busulfan with Cyclophosphamide/Total Body Irradiation, for Allogeneic Stem Cell Transplantation in the Modern Era: A Cohort Analysis. *Biol Blood Marrow Transplant* **24**, 1733-1740, doi:10.1016/j.bbmt.2018.03.011 (2018).

59 He, X. *et al.* Conditioning with Fludarabine-Busulfan versus Busulfan-Cyclophosphamide Is Associated with Lower aGVHD and Higher Survival but More Extensive and Long Standing Bone Marrow Damage. *Biomed Res Int* **2016**, 3071214, doi:10.1155/2016/3071214 (2016).

60 Weaver, C. H. *et al.* An analysis of engraftment kinetics as a function of the CD34 content of peripheral blood progenitor cell collections in 692 patients after the administration of myeloablative chemotherapy. *Blood* **86**, 3961-3969 (1995).

61 Saraceni, F., Shem-Tov, N., Olivieri, A. & Nagler, A. Mobilized peripheral blood grafts include more than hematopoietic stem cells: the immunological perspective. *Bone Marrow Transplant* **50**, 886-891, doi:10.1038/bmt.2014.330 (2015).

62 Duong, H. K. *et al.* Peripheral blood progenitor cell mobilization for autologous and allogeneic hematopoietic cell transplantation: guidelines from the American Society for Blood and Marrow Transplantation. *Biol Blood Marrow Transplant* **20**, 1262-1273, doi:10.1016/j.bbmt.2014.05.003 (2014).

63 Perez-Simon, J. A. *et al.* Clinical significance of CD34+ cell dose in long-term engraftment following autologous peripheral blood stem cell transplantation. *Bone Marrow Transplant* **24**, 1279-1283, doi:10.1038/sj.bmt.1702066 (1999).

64 Giralt, S. *et al.* Optimizing autologous stem cell mobilization strategies to improve patient outcomes: consensus guidelines and recommendations. *Biol Blood Marrow Transplant* **20**, 295-308, doi:10.1016/j.bbmt.2013.10.013 (2014).

65 Lisenko, K. *et al.* Comparison of Different Stem Cell Mobilization Regimens in AL Amyloidosis Patients. *Biol Blood Marrow Transplant* **23**, 1870-1878, doi:10.1016/j.bbmt.2017.07.015 (2017).

66 Mohty, M. *et al.* Autologous haematopoietic stem cell mobilisation in multiple myeloma and lymphoma patients: a position statement from the European Group for Blood and Marrow Transplantation. *Bone Marrow Transplant* **49**, 865-872, doi:10.1038/bmt.2014.39 (2014).

67 Hubel, K. in *The EBMT Handbook: Hematopoietic Stem Cell Transplantation and Cellular Therapies* (eds th *et al.*) 117-122 (2019).

68 Schmitz, N. *et al.* Filgrastim-mobilized peripheral blood progenitor cells versus bone marrow transplantation for treating leukemia: 3-year results from the EBMT randomized trial. *Haematologica* **90**, 643-648 (2005).

69 Weaver, C. H. *et al.* Syngeneic transplantation with peripheral blood mononuclear cells collected after the administration of recombinant human granulocyte colony-stimulating factor. *Blood* **82**, 1981-1984 (1993).

70 Bensinger, W. I. *et al.* Transplantation of allogeneic peripheral blood stem cells mobilized by recombinant human granulocyte colony stimulating factor. *Stem Cells* **14**, 90-105, doi:10.1002/stem.140090 (1996).

71 Shpall, E. J. *et al.* A randomized phase 3 study of peripheral blood progenitor cell mobilization with stem cell factor and filgrastim in high-risk breast cancer patients. *Blood* **93**, 2491-2501 (1999).

72 Passweg, J. R. *et al.* Hematopoietic stem cell transplantation in Europe 2014: more than 40 000 transplants annually. *Bone Marrow Transplant* **51**, 786-792, doi:10.1038/bmt.2016.20 (2016).

73 Passweg, J. R. *et al.* The EBMT activity survey: 1990-2010. *Bone Marrow Transplant* **47**, 906-923, doi:10.1038/bmt.2012.66 (2012).

74 Gertz, M. A. Current status of stem cell mobilization. *Br J Haematol* **150**, 647-662, doi:10.1111/j.1365-2141.2010.08313.x (2010).

75 Levesque, J. P. *et al.* Mobilization by either cyclophosphamide or granulocyte colony-stimulating factor transforms the bone marrow into a highly proteolytic environment. *Exp Hematol* **30**, 440-449, doi:10.1016/s0301-472x(02)00788-9 (2002).

76 Lapidot, T. & Petit, I. Current understanding of stem cell mobilization: the roles of chemokines, proteolytic enzymes, adhesion molecules, cytokines, and stromal cells. *Exp Hematol* **30**, 973-981, doi:10.1016/s0301-472x(02)00883-4 (2002).

77 Broxmeyer, H. E. *et al.* Rapid mobilization of murine and human hematopoietic stem and progenitor cells with AMD3100, a CXCR4 antagonist. *J Exp Med* **201**, 1307-1318, doi:10.1084/jem.20041385 (2005).

78 Matsuzaki, Y., Kinjo, K., Mulligan, R. C. & Okano, H. Unexpectedly efficient homing capacity of purified murine hematopoietic stem cells. *Immunity* **20**, 87-93, doi:10.1016/s1074-7613(03)00354-6 (2004).

79 Donahue, R. E. *et al.* Plerixafor (AMD3100) and granulocyte colony-stimulating factor (G-CSF) mobilize different CD34+ cell populations based on global gene and microRNA expression signatures. *Blood* **114**, 2530-2541, doi:10.1182/blood-2009-04-214403 (2009).

80 Fuchs, E. J. Haploidentical transplantation for hematologic malignancies: where do we stand? *Hematology Am Soc Hematol Educ Program* **2012**, 230-236, doi:10.1182/asheducation-2012.1.230 (2012).

81 Styczynski, J. *et al.* Death after hematopoietic stem cell transplantation: changes over calendar year time, infections and associated factors. *Bone Marrow Transplant* **55**, 126-136, doi:10.1038/s41409-019-0624-z (2020).

82 Mann, R., Mulligan, R. C. & Baltimore, D. Construction of a retrovirus packaging mutant and its use to produce helper-free defective retrovirus. *Cell* **33**, 153-159, doi:10.1016/0092-8674(83)90344-6 (1983).

83 Adamson, C. S. Protease-Mediated Maturation of HIV: Inhibitors of Protease and the Maturation Process. *Mol Biol Int* **2012**, 604261, doi:10.1155/2012/604261 (2012).

84 Uren, A. G., Kool, J., Berns, A. & van Lohuizen, M. Retroviral insertional mutagenesis: past, present and future. *Oncogene* **24**, 7656-7672, doi:10.1038/sj.onc.1209043 (2005).

85 Coffin, J. M., Hughes, S. H. & Varmus, H. E. in *Retroviruses* (eds J. M. Coffin, S. H. Hughes, & H. E. Varmus) (1997).

86 Naldini, L. *et al.* In vivo gene delivery and stable transduction of nondividing cells by a lentiviral vector. *Science* **272**, 263-267, doi:10.1126/science.272.5259.263 (1996).

87 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, doi:10.1038/nbt0997-871 (1997).

88 Kafri, T., Blomer, 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, doi:10.1038/ng1197-314 (1997).

89 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).

90 Dull, T. *et al.* A third-generation lentivirus vector with a conditional packaging system. *J Virol* **72**, 8463-8471 (1998).

91 Zufferey, R. *et al.* Self-inactivating lentivirus vector for safe and efficient in vivo gene delivery. *J Virol* **72**, 9873-9880 (1998).

92 Iwakuma, T., Cui, Y. & Chang, L. J. Self-inactivating lentiviral vectors with U3 and U5 modifications. *Virology* **261**, 120-132, doi:10.1006/viro.1999.9850 (1999).

93 Miyoshi, H., Blomer, U., Takahashi, M., Gage, F. H. & Verma, I. M. Development of a self-inactivating lentivirus vector. *J Virol* **72**, 8150-8157 (1998).

94 Bokhoven, M. *et al.* Insertional gene activation by lentiviral and gammaretroviral vectors. *J Virol* **83**, 283-294, doi:10.1128/JVI.01865-08 (2009).

95 Williams, D. A., Lemischka, I. R., Nathan, D. G. & Mulligan, R. C. Introduction of new genetic material into pluripotent haematopoietic stem cells of the mouse. *Nature* **310**, 476-480, doi:10.1038/310476a0 (1984).

96 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, doi:10.1056/NEJM199008303230904 (1990).

97 Blaese, R. M. *et al.* T lymphocyte-directed gene therapy for ADA- SCID: initial trial results after 4 years. *Science* **270**, 475-480, doi:10.1126/science.270.5235.475 (1995).

98 Candotti, F. *et al.* Gene therapy for adenosine deaminase-deficient severe combined immune deficiency: clinical comparison of retroviral vectors and treatment plans. *Blood* **120**, 3635-3646, doi:10.1182/blood-2012-02-400937 (2012).

99 Cicalese, M. P. & Aiuti, A. Clinical applications of gene therapy for primary immunodeficiencies. *Hum Gene Ther* **26**, 210-219, doi:10.1089/hum.2015.047 (2015).

100 Cicalese, M. P. *et al.* Update on the safety and efficacy of retroviral gene therapy for immunodeficiency due to adenosine deaminase deficiency. *Blood* **128**, 45-54, doi:10.1182/blood-2016-01-688226 (2016).

101 Gaspar, H. B. *et al.* Long-term persistence of a polyclonal T cell repertoire after gene therapy for X-linked severe combined immunodeficiency. *Sci Transl Med* **3**, 97ra79, doi:10.1126/scitranslmed.3002715 (2011).

102 Hacein-Bey-Abina, S. *et al.* Sustained correction of X-linked severe combined immunodeficiency by ex vivo gene therapy. *N Engl J Med* **346**, 1185-1193, doi:10.1056/NEJMoa012616 (2002).

103 Boztug, K. *et al.* Stem-cell gene therapy for the Wiskott-Aldrich syndrome. *N Engl J Med* **363**, 1918-1927, doi:10.1056/NEJMoa1003548 (2010).

104 Braun, C. J. *et al.* Gene therapy for Wiskott-Aldrich Syndrome-Long-term reconstitution and clinical benefits, but increased risk for leukemogenesis. *Rare Dis* **2**, e947749, doi:10.4161/21675511.2014.947749 (2014).

105 Bianchi, M. *et al.* Restoration of NET formation by gene therapy in CGD controls aspergillosis. *Blood* **114**, 2619-2622, doi:10.1182/blood-2009-05-221606 (2009).

106 Kang, E. M. *et al.* Retrovirus gene therapy for X-linked chronic granulomatous disease can achieve stable long-term correction of oxidase activity in peripheral blood neutrophils. *Blood* **115**, 783-791, doi:10.1182/blood-2009-05-222760 (2010).

107 Cartier, N. *et al.* Hematopoietic stem cell gene therapy with a lentiviral vector in X-linked adrenoleukodystrophy. *Science* **326**, 818-823, doi:10.1126/science.1171242 (2009).

108 Karlsson, S., Ooka, A. & Woods, N. B. Development of gene therapy for blood disorders by gene transfer into haematopoietic stem cells. *Haemophilia* **8**, 255-260, doi:10.1046/j.1365-2516.2002.00470.x (2002).

109 Ali, S. *et al.* The European Medicines Agency Review of Kymriah (Tisagenlecleucel) for the Treatment of Acute Lymphoblastic Leukemia and Diffuse Large B-Cell Lymphoma. *Oncologist* **25**, e321-e327, doi:10.1634/theoncologist.2019-0233 (2020).

110 Papadouli, I. *et al.* EMA Review of Axicabtagene Ciloleucel (Yescarta) for the Treatment of Diffuse Large B-Cell Lymphoma. *Oncologist*, doi:10.1634/theoncologist.2019-0646 (2020).

111 Hacein-Bey-Abina, S. *et al.* Insertional oncogenesis in 4 patients after retrovirus-mediated gene therapy of SCID-X1. *J Clin Invest* **118**, 3132-3142, doi:10.1172/JCI35700 (2008).

112 Hacein-Bey-Abina, S. *et al.* A serious adverse event after successful gene therapy for X-linked severe combined immunodeficiency. *N Engl J Med* **348**, 255-256, doi:10.1056/NEJM200301163480314 (2003).

113 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, doi:10.1038/nm1393 (2006).

114 Stein, S. *et al.* Genomic instability and myelodysplasia with monosomy 7 consequent to EVI1 activation after gene therapy for chronic granulomatous disease. *Nat Med* **16**, 198-204, doi:10.1038/nm.2088 (2010).

115 Cavazzana-Calvo, M. *et al.* Transfusion independence and HMGA2 activation after gene therapy of human beta-thalassaemia. *Nature* **467**, 318-322, doi:10.1038/nature09328 (2010).

116 Jonkers, J. & Berns, A. Retroviral insertional mutagenesis as a strategy to identify cancer genes. *Biochim Biophys Acta* **1287**, 29-57, doi:10.1016/0304-419x(95)00020-g (1996).

117 Notta, F. *et al.* Isolation of single human hematopoietic stem cells capable of long-term multilineage engraftment. *Science* **333**, 218-221, doi:10.1126/science.1201219 (2011).

118 Lemischka, I. R., Raulet, D. H. & Mulligan, R. C. Developmental potential and dynamic behavior of hematopoietic stem cells. *Cell* **45**, 917-927, doi:10.1016/0092-8674(86)90566-0 (1986).

119 Capel, B., Hawley, R. G. & Mintz, B. Long- and short-lived murine hematopoietic stem cell clones individually identified with retroviral integration markers. *Blood* **75**, 2267-2270 (1990).

120 Dick, J. E., Magli, M. C., Huszar, D., Phillips, R. A. & Bernstein, A. Introduction of a selectable gene into primitive stem cells capable of long-term reconstitution of the hemopoietic system of W/Wv mice. *Cell* **42**, 71-79, doi:10.1016/s0092-8674(85)80102-1 (1985).

121 Jordan, C. T. & Lemischka, I. R. Clonal and systemic analysis of long-term hematopoiesis in the mouse. *Genes Dev* **4**, 220-232, doi:10.1101/gad.4.2.220 (1990).

122 Lemischka, I. R. What we have learned from retroviral marking of hematopoietic stem cells. *Curr Top Microbiol Immunol* **177**, 59-71, doi:10.1007/978-3-642-76912-2_5 (1992).

123 Kreso, A. *et al.* Variable clonal repopulation dynamics influence chemotherapy response in colorectal cancer. *Science* **339**, 543-548, doi:10.1126/science.1227670 (2013).

124 Livet, J. *et al.* Transgenic strategies for combinatorial expression of fluorescent proteins in the nervous system. *Nature* **450**, 56-62, doi:10.1038/nature06293 (2007).

125 Schepers, A. G. *et al.* Lineage tracing reveals Lgr5+ stem cell activity in mouse intestinal adenomas. *Science* **337**, 730-735, doi:10.1126/science.1224676 (2012).

126 Rios, A. C., Fu, N. Y., Lindeman, G. J. & Visvader, J. E. In situ identification of bipotent stem cells in the mammary gland. *Nature* **506**, 322-327, doi:10.1038/nature12948 (2014).

127 Schmidt, M. *et al.* Detection and direct genomic sequencing of multiple rare unknown flanking DNA in highly complex samples. *Hum Gene Ther* **12**, 743-749, doi:10.1089/104303401750148649 (2001).

128 Keller, G., Paige, C., Gilboa, E. & Wagner, E. F. Expression of a foreign gene in myeloid and lymphoid cells derived from multipotent haematopoietic precursors. *Nature* **318**, 149-154, doi:10.1038/318149a0 (1985).

129 Rosenthal, A. & Jones, D. S. Genomic walking and sequencing by oligo-cassette mediated polymerase chain reaction. *Nucleic Acids Res* **18**, 3095-3096, doi:10.1093/nar/18.10.3095 (1990).

130 Cavrois, M., Wain-Hobson, S. & Wattel, E. Stochastic events in the amplification of HTLV-I integration sites by linker-mediated PCR. *Res Virol* **146**, 179-184, doi:10.1016/0923-2516(96)80578-4 (1995).

131 Niederer, H. A. & Bangham, C. R. Integration site and clonal expansion in human chronic retroviral infection and gene therapy. *Viruses* **6**, 4140-4164, doi:10.3390/v6114140 (2014).

132 Gabriel, R. *et al.* Comprehensive genomic access to vector integration in clinical gene therapy. *Nat Med* **15**, 1431-1436, doi:10.1038/nm.2057 (2009).

133 Busch, K. *et al.* Fundamental properties of unperturbed haematopoiesis from stem cells in vivo. *Nature* **518**, 542-546, doi:10.1038/nature14242 (2015).

134 Golden, J. A., Fields-Berry, S. C. & Cepko, C. L. Construction and characterization of a highly complex retroviral library for lineage analysis. *Proc Natl Acad Sci U S A* **92**, 5704-5708, doi:10.1073/pnas.92.12.5704 (1995).

135 Schepers, K. *et al.* Dissecting T cell lineage relationships by cellular barcoding. *J Exp Med* **205**, 2309-2318, doi:10.1084/jem.20072462 (2008).

136 Gerrits, A. *et al.* Cellular barcoding tool for clonal analysis in the hematopoietic system. *Blood* **115**, 2610-2618, doi:10.1182/blood-2009-06-229757 (2010).

137 Adair, J. E. *et al.* DNA Barcoding in Nonhuman Primates Reveals Important Limitations in Retrovirus Integration Site Analysis. *Mol Ther Methods Clin Dev* **17**, 796-809, doi:10.1016/j.omtm.2020.03.021 (2020).

138 van Heijst, J. W. *et al.* Recruitment of antigen-specific CD8+ T cells in response to infection is markedly efficient. *Science* **325**, 1265-1269, doi:10.1126/science.1175455 (2009).

139 Gerlach, C. *et al.* Heterogeneous differentiation patterns of individual CD8+ T cells. *Science* **340**, 635-639, doi:10.1126/science.1235487 (2013).

140 Schumacher, T. N., Gerlach, C. & van Heijst, J. W. Mapping the life histories of T cells. *Nat Rev Immunol* **10**, 621-631, doi:10.1038/nri2822 (2010).

141 Nguyen, L. V. *et al.* Clonal analysis via barcoding reveals diverse growth and differentiation of transplanted mouse and human mammary stem cells. *Cell Stem Cell* **14**, 253-263, doi:10.1016/j.stem.2013.12.011 (2014).

142 Wagenblast, E. *et al.* A model of breast cancer heterogeneity reveals vascular mimicry as a driver of metastasis. *Nature* **520**, 358-362, doi:10.1038/nature14403 (2015).

143 Echeverria, G. V. *et al.* High-resolution clonal mapping of multi-organ metastasis in triple negative breast cancer. *Nat Commun* **9**, 5079, doi:10.1038/s41467-018-07406-4 (2018).

144 Merino, D. *et al.* Barcoding reveals complex clonal behavior in patient-derived xenografts of metastatic triple negative breast cancer. *Nat Commun* **10**, 766, doi:10.1038/s41467-019-08595-2 (2019).

145 Roh, V. *et al.* Cellular Barcoding Identifies Clonal Substitution as a Hallmark of Local Recurrence in a Surgical Model of Head and Neck Squamous Cell Carcinoma. *Cell Rep* **25**, 2208-2222 e2207, doi:10.1016/j.celrep.2018.10.090 (2018).

146 Mathis, R. A., Sokol, E. S. & Gupta, P. B. Cancer cells exhibit clonal diversity in phenotypic plasticity. *Open Biol* **7**, doi:10.1098/rsob.160283 (2017).

147 Lan, X. *et al.* Fate mapping of human glioblastoma reveals an invariant stem cell hierarchy. *Nature* **549**, 227-232, doi:10.1038/nature23666 (2017).

148 Bhang, H. E. *et al.* Studying clonal dynamics in response to cancer therapy using high-complexity barcoding. *Nat Med* **21**, 440-448, doi:10.1038/nm.3841 (2015).

149 Hata, A. N. *et al.* Tumor cells can follow distinct evolutionary paths to become resistant to epidermal growth factor receptor inhibition. *Nat Med* **22**, 262-269, doi:10.1038/nm.4040 (2016).

150 Bell, C. C. *et al.* Targeting enhancer switching overcomes non-genetic drug resistance in acute myeloid leukaemia. *Nat Commun* **10**, 2723, doi:10.1038/s41467-019-10652-9 (2019).

151 Caiado, F. *et al.* Lineage tracing of acute myeloid leukemia reveals the impact of hypomethylating agents on chemoresistance selection. *Nat Commun* **10**, 4986, doi:10.1038/s41467-019-12983-z (2019).

152 Seth, S. *et al.* Pre-existing Functional Heterogeneity of Tumorigenic Compartment as the Origin of Chemoresistance in Pancreatic Tumors. *Cell Rep* **26**, 1518-1532 e1519, doi:10.1016/j.celrep.2019.01.048 (2019).

153 Nguyen, L. V. *et al.* Barcoding reveals complex clonal dynamics of de novo transformed human mammary cells. *Nature* **528**, 267-271, doi:10.1038/nature15742 (2015).

154 Guernet, A. *et al.* CRISPR-Barcoding for Intratumor Genetic Heterogeneity Modeling and Functional Analysis of Oncogenic Driver Mutations. *Mol Cell* **63**, 526-538, doi:10.1016/j.molcel.2016.06.017 (2016).

155 Aranyossy, T., Thielecke, L., Glauche, I., Fehse, B. & Cornils, K. Genetic Barcodes Facilitate Competitive Clonal Analyses In Vivo. *Hum Gene Ther* **28**, 926-937, doi:10.1089/hum.2017.124 (2017).

156 Thielecke, L. *et al.* Limitations and challenges of genetic barcode quantification. *Sci Rep* **7**, 43249, doi:10.1038/srep43249 (2017).

157 Weber, K., Thomaschewski, M., Benten, D. & Fehse, B. RGB marking with lentiviral vectors for multicolor clonal cell tracking. *Nat Protoc* **7**, 839-849, doi:10.1038/nprot.2012.026 (2012).

158 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, doi:10.1038/nbt.1977 (2011).

159 Bystrykh, L. V., Verovskaya, E., Zwart, E., Broekhuis, M. & de Haan, G. Counting stem cells: methodological constraints. *Nat Methods* **9**, 567-574, doi:10.1038/nmeth.2043 (2012).

160 Verovskaya, E. *et al.* Heterogeneity of young and aged murine hematopoietic stem cells revealed by quantitative clonal analysis using cellular barcoding. *Blood* **122**, 523-532, doi:10.1182/blood-2013-01-481135 (2013).

161 Cornils, K. *et al.* Multiplexing clonality: combining RGB marking and genetic barcoding. *Nucleic Acids Res* **42**, e56, doi:10.1093/nar/gku081 (2014).

162 Naik, S. H., Schumacher, T. N. & Perie, L. Cellular barcoding: a technical appraisal. *Exp Hematol* **42**, 598-608, doi:10.1016/j.exphem.2014.05.003 (2014).

163 Verovskaya, E. *et al.* Asymmetry in skeletal distribution of mouse hematopoietic stem cell clones and their equilibration by mobilizing cytokines. *J Exp Med* **211**, 487-497, doi:10.1084/jem.20131804 (2014).

164 Sun, J. *et al.* Clonal dynamics of native haematopoiesis. *Nature* **514**, 322-327, doi:10.1038/nature13824 (2014).

165 Rodriguez-Fraticelli, A. E. *et al.* Clonal analysis of lineage fate in native haematopoiesis. *Nature* **553**, 212-216, doi:10.1038/nature25168 (2018).

166 Snippert, H. J. *et al.* Intestinal crypt homeostasis results from neutral competition between symmetrically dividing Lgr5 stem cells. *Cell* **143**, 134-144, doi:10.1016/j.cell.2010.09.016 (2010).

167 Pei, W. *et al.* Polylox barcoding reveals haematopoietic stem cell fates realized in vivo. *Nature* **548**, 456-460, doi:10.1038/nature23653 (2017).

168 Pei, W. *et al.* Using Cre-recombinase-driven Polylox barcoding for in vivo fate mapping in mice. *Nat Protoc* **14**, 1820-1840, doi:10.1038/s41596-019-0163-5 (2019).

169 Hwang, B. *et al.* Lineage tracing using a Cas9-deaminase barcoding system targeting endogenous L1 elements. *Nat Commun* **10**, 1234, doi:10.1038/s41467-019-09203-z (2019).

170 Halperin, S. O. *et al.* CRISPR-guided DNA polymerases enable diversification of all nucleotides in a tunable window. *Nature* **560**, 248-252, doi:10.1038/s41586-018-0384-8 (2018).

171 McKenna, A. *et al.* Whole-organism lineage tracing by combinatorial and cumulative genome editing. *Science* **353**, aaf7907, doi:10.1126/science.aaf7907 (2016).

172 Thielecke, L., Cornils, K. & Glauche, I. genBaRcode: a comprehensive R-package for genetic barcode analysis. *Bioinformatics* **36**, 2189-2194, doi:10.1093/bioinformatics/btz872 (2020).

173 Bystrykh, L. V. & Belderbos, M. E. Clonal Analysis of Cells with Cellular Barcoding: When Numbers and Sizes Matter. *Methods Mol Biol* **1516**, 57-89, doi:10.1007/7651_2016_343 (2016).

174 Kivioja, T. *et al.* Counting absolute numbers of molecules using unique molecular identifiers. *Nat Methods* **9**, 72-74, doi:10.1038/nmeth.1778 (2011).

175 Teets, E. M., Gregory, C., Shaffer, J., Blachly, J. S. & Blaser, B. W. Quantifying Hematopoietic Stem Cell Clonal Diversity by Selecting Informative Amplicon Barcodes. *Sci Rep* **10**, 2153, doi:10.1038/s41598-020-59119-8 (2020).

176 Naik, S. H. *et al.* Diverse and heritable lineage imprinting of early haematopoietic progenitors. *Nature* **496**, 229-232, doi:10.1038/nature12013 (2013).

177 Perie, L. *et al.* Determining lineage pathways from cellular barcoding experiments. *Cell Rep* **6**, 617-624, doi:10.1016/j.celrep.2014.01.016 (2014).

178 Aiuti, A. *et al.* Lentiviral hematopoietic stem cell gene therapy in patients with Wiskott-Aldrich syndrome. *Science* **341**, 1233151, doi:10.1126/science.1233151 (2013).

179 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**, 5332-5339, doi:10.1182/blood-2010-09-306761 (2011).

180 Biasco, L. Integration Site Analysis in Gene Therapy Patients: Expectations and Reality. *Hum Gene Ther* **28**, 1122-1129, doi:10.1089/hum.2017.183 (2017).

181 Giordano, F. A. *et al.* High-throughput monitoring of integration site clonality in preclinical and clinical gene therapy studies. *Mol Ther Methods Clin Dev* **2**, 14061, doi:10.1038/mtm.2014.61 (2015).

182 Six, E. *et al.* Clonal tracking in gene therapy patients reveals a diversity of human hematopoietic differentiation programs. *Blood* **135**, 1219-1231, doi:10.1182/blood.2019002350 (2020).

183 Wang, K. *et al.* Clonal origin in normal adults of all blood lineages and circulating hematopoietic stem cells. *Exp Hematol* **83**, 25-34 e22, doi:10.1016/j.exphem.2020.01.005 (2020).

184 Demeulemeester, J., De Rijck, J., Gijsbers, R. & Debyser, Z. Retroviral integration: Site matters: Mechanisms and consequences of retroviral integration site selection. *Bioessays* **37**, 1202-1214, doi:10.1002/bies.201500051 (2015).

185 Wang, G. P. *et al.* DNA bar coding and pyrosequencing to analyze adverse events in therapeutic gene transfer. *Nucleic Acids Res* **36**, e49, doi:10.1093/nar/gkn125 (2008).

186 Papapetrou, E. P. *et al.* Genomic safe harbors permit high beta-globin transgene expression in thalassemia induced pluripotent stem cells. *Nat Biotechnol* **29**, 73-78, doi:10.1038/nbt.1717 (2011).

187 Sadelain, M., Papapetrou, E. P. & Bushman, F. D. Safe harbours for the integration of new DNA in the human genome. *Nat Rev Cancer* **12**, 51-58, doi:10.1038/nrc3179 (2011).

188 Papapetrou, E. P. & Schambach, A. Gene Insertion Into Genomic Safe Harbors for Human Gene Therapy. *Mol Ther* **24**, 678-684, doi:10.1038/mt.2016.38 (2016).

189 Jinek, M. *et al.* A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816-821, doi:10.1126/science.1225829 (2012).

190 Ishino, Y., Shinagawa, H., Makino, K., Amemura, M. & Nakata, A. Nucleotide sequence of the iap gene, responsible for alkaline phosphatase isozyme conversion in Escherichia coli, and identification of the gene product. *J Bacteriol* **169**, 5429-5433, doi:10.1128/jb.169.12.5429-5433.1987 (1987).

191 Jansen, R., Embden, J. D., Gaastra, W. & Schouls, L. M. Identification of genes that are associated with DNA repeats in prokaryotes. *Mol Microbiol* **43**, 1565-1575, doi:10.1046/j.1365-2958.2002.02839.x (2002).

192 Mojica, F. J., Diez-Villasenor, C., Garcia-Martinez, J. & Soria, E. Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. *J Mol Evol* **60**, 174-182, doi:10.1007/s00239-004-0046-3 (2005).

193 Pourcel, C., Salvignol, G. & Vergnaud, G. CRISPR elements in Yersinia pestis acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. *Microbiology* **151**, 653-663, doi:10.1099/mic.0.27437-0 (2005).

194 Bolotin, A., Quinquis, B., Sorokin, A. & Ehrlich, S. D. Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. *Microbiology* **151**, 2551-2561, doi:10.1099/mic.0.28048-0 (2005).

195 Deltcheva, E. *et al.* CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature* **471**, 602-607, doi:10.1038/nature09886 (2011).

196 Barrangou, R. *et al.* CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **315**, 1709-1712, doi:10.1126/science.1138140 (2007).

197 Yosef, I., Goren, M. G. & Qimron, U. Proteins and DNA elements essential for the CRISPR adaptation process in Escherichia coli. *Nucleic Acids Res* **40**, 5569-5576, doi:10.1093/nar/gks216 (2012).

198 Chylinski, K., Le Rhun, A. & Charpentier, E. The tracrRNA and Cas9 families of type II CRISPR-Cas immunity systems. *RNA Biol* **10**, 726-737, doi:10.4161/rna.24321 (2013).

199 Cong, L. *et al.* Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**, 819-823, doi:10.1126/science.1231143 (2013).

200 Makarova, K. S. *et al.* Evolution and classification of the CRISPR-Cas systems. *Nat Rev Microbiol* **9**, 467-477, doi:10.1038/nrmicro2577 (2011).

201 Mali, P. *et al.* RNA-guided human genome engineering via Cas9. *Science* **339**, 823-826, doi:10.1126/science.1232033 (2013).

202 Nishimasu, H. *et al.* Crystal structure of Cas9 in complex with guide RNA and target DNA. *Cell* **156**, 935-949, doi:10.1016/j.cell.2014.02.001 (2014).

203 Hefferin, M. L. & Tomkinson, A. E. Mechanism of DNA double-strand break repair by non-homologous end joining. *DNA Repair (Amst)* **4**, 639-648, doi:10.1016/j.dnarep.2004.12.005 (2005).

204 Bibikova, M., Golic, M., Golic, K. G. & Carroll, D. Targeted chromosomal cleavage and mutagenesis in Drosophila using zinc-finger nucleases. *Genetics* **161**, 1169-1175 (2002).

205 Takata, M. *et al.* Homologous recombination and non-homologous end-joining pathways of DNA double-strand break repair have overlapping roles in the maintenance of chromosomal integrity in vertebrate cells. *EMBO J* **17**, 5497-5508, doi:10.1093/emboj/17.18.5497 (1998).

206 Capecchi, M. R. Altering the genome by homologous recombination. *Science* **244**, 1288-1292, doi:10.1126/science.2660260 (1989).

207 Radecke, S., Radecke, F., Cathomen, T. & Schwarz, K. Zinc-finger nuclease-induced gene repair with oligodeoxynucleotides: wanted and unwanted target locus modifications. *Mol Ther* **18**, 743-753, doi:10.1038/mt.2009.304 (2010).

208 Soldner, F. *et al.* Generation of isogenic pluripotent stem cells differing exclusively at two early onset Parkinson point mutations. *Cell* **146**, 318-331, doi:10.1016/j.cell.2011.06.019 (2011).

209 Chen, F. *et al.* High-frequency genome editing using ssDNA oligonucleotides with zinc-finger nucleases. *Nat Methods* **8**, 753-755, doi:10.1038/nmeth.1653 (2011).

210 Rouet, P., Smih, F. & Jasin, M. Expression of a site-specific endonuclease stimulates homologous recombination in mammalian cells. *Proc Natl Acad Sci U S A* **91**, 6064-6068, doi:10.1073/pnas.91.13.6064 (1994).

211 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* **44**, e85, doi:10.1093/nar/gkw064 (2016).

212 Maruyama, T. *et al.* Increasing the efficiency of precise genome editing with CRISPR-Cas9 by inhibition of nonhomologous end joining. *Nat Biotechnol* **33**, 538-542, doi:10.1038/nbt.3190 (2015).

213 Mao, Z., Bozzella, M., Seluanov, A. & Gorbunova, V. DNA repair by nonhomologous end joining and homologous recombination during cell cycle in human cells. *Cell Cycle* **7**, 2902-2906, doi:10.4161/cc.7.18.6679 (2008).

214 Chu, V. T. *et al.* Increasing the efficiency of homology-directed repair for CRISPR-Cas9-induced precise gene editing in mammalian cells. *Nat Biotechnol* **33**, 543-548, doi:10.1038/nbt.3198 (2015).

215 Weinstock, D. M. & Jasin, M. Alternative pathways for the repair of RAG-induced DNA breaks. *Mol Cell Biol* **26**, 131-139, doi:10.1128/MCB.26.1.131-139.2006 (2006).

216 Srivastava, M. *et al.* An inhibitor of nonhomologous end-joining abrogates double-strand break repair and impedes cancer progression. *Cell* **151**, 1474-1487, doi:10.1016/j.cell.2012.11.054 (2012).

217 Tomkinson, A. E., Howes, T. R. & Wiest, N. E. DNA ligases as therapeutic targets. *Transl Cancer Res* **2** (2013).

218 Robert, F., Barbeau, M., Ethier, S., Dostie, J. & Pelletier, J. Pharmacological inhibition of DNA-PK stimulates Cas9-mediated genome editing. *Genome Med* **7**, 93, doi:10.1186/s13073-015-0215-6 (2015).

219 Yu, C. *et al.* Small molecules enhance CRISPR genome editing in pluripotent stem cells. *Cell Stem Cell* **16**, 142-147, doi:10.1016/j.stem.2015.01.003 (2015).

220 Zhang, J. P. *et al.* Efficient precise knockin with a double cut HDR donor after CRISPR/Cas9-mediated double-stranded DNA cleavage. *Genome Biol* **18**, 35, doi:10.1186/s13059-017-1164-8 (2017).

221 Yoshimi, K. *et al.* ssODN-mediated knock-in with CRISPR-Cas for large genomic regions in zygotes. *Nat Commun* **7**, 10431, doi:10.1038/ncomms10431 (2016).

222 Yao, X. *et al.* Homology-mediated end joining-based targeted integration using CRISPR/Cas9. *Cell Res* **27**, 801-814, doi:10.1038/cr.2017.76 (2017).

223 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, doi:10.1101/gr.145441.112 (2013).

224 Hoban, M. D. *et al.* Correction of the sickle cell disease mutation in human hematopoietic stem/progenitor cells. *Blood* **125**, 2597-2604, doi:10.1182/blood-2014-12-615948 (2015).

225 DeWitt, M. A. *et al.* Selection-free genome editing of the sickle mutation in human adult hematopoietic stem/progenitor cells. *Sci Transl Med* **8**, 360ra134, doi:10.1126/scitranslmed.aaf9336 (2016).

226 Genovese, P. *et al.* Targeted genome editing in human repopulating haematopoietic stem cells. *Nature* **510**, 235-240, doi:10.1038/nature13420 (2014).

227 Ortinski, P. I., O'Donovan, B., Dong, X. & Kantor, B. Integrase-Deficient Lentiviral Vector as an All-in-One Platform for Highly Efficient CRISPR/Cas9-Mediated Gene Editing. *Mol Ther Methods Clin Dev* **5**, 153-164, doi:10.1016/j.omtm.2017.04.002 (2017).

228 Lin, S., Staahl, B. T., Alla, R. K. & Doudna, J. A. Enhanced homology-directed human genome engineering by controlled timing of CRISPR/Cas9 delivery. *Elife* **3**, e04766, doi:10.7554/eLife.04766 (2014).

229 Kim, S., Kim, D., Cho, S. W., Kim, J. & Kim, J. S. Highly efficient RNA-guided genome editing in human cells via delivery of purified Cas9 ribonucleoproteins. *Genome Res* **24**, 1012-1019, doi:10.1101/gr.171322.113 (2014).

230 Cradick, T. J., Fine, E. J., Antico, C. J. & Bao, G. CRISPR/Cas9 systems targeting beta-globin and CCR5 genes have substantial off-target activity. *Nucleic Acids Res* **41**, 9584-9592, doi:10.1093/nar/gkt714 (2013).

231 Elms, S. *et al.* Insights into the arginine paradox: evidence against the importance of subcellular location of arginase and eNOS. *Am J Physiol Heart Circ Physiol* **305**, H651-666, doi:10.1152/ajpheart.00755.2012 (2013).

232 Pattanayak, V. *et al.* High-throughput profiling of off-target DNA cleavage reveals RNA-programmed Cas9 nuclease specificity. *Nat Biotechnol* **31**, 839-843, doi:10.1038/nbt.2673 (2013).

233 Frock, R. L. *et al.* Genome-wide detection of DNA double-stranded breaks induced by engineered nucleases. *Nat Biotechnol* **33**, 179-186, doi:10.1038/nbt.3101 (2015).

234 Akcakaya, P. *et al.* In vivo CRISPR editing with no detectable genome-wide off-target mutations. *Nature* **561**, 416-419, doi:10.1038/s41586-018-0500-9 (2018).

235 Tsai, S. Q. & Joung, J. K. Defining and improving the genome-wide specificities of CRISPR-Cas9 nucleases. *Nat Rev Genet* **17**, 300-312, doi:10.1038/nrg.2016.28 (2016).

236 Sternberg, S. H., Redding, S., Jinek, M., Greene, E. C. & Doudna, J. A. DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. *Nature* **507**, 62-67, doi:10.1038/nature13011 (2014).

237 Haeussler, M. *et al.* Evaluation of off-target and on-target scoring algorithms and integration into the guide RNA selection tool CRISPOR. *Genome Biol* **17**, 148, doi:10.1186/s13059-016-1012-2 (2016).

238 Ramakrishna, S. *et al.* Gene disruption by cell-penetrating peptide-mediated delivery of Cas9 protein and guide RNA. *Genome Res* **24**, 1020-1027, doi:10.1101/gr.171264.113 (2014).

239 Cho, S. W. *et al.* Analysis of off-target effects of CRISPR/Cas-derived RNA-guided endonucleases and nickases. *Genome Res* **24**, 132-141, doi:10.1101/gr.162339.113 (2014).

240 Shen, B. *et al.* Efficient genome modification by CRISPR-Cas9 nickase with minimal off-target effects. *Nat Methods* **11**, 399-402, doi:10.1038/nmeth.2857 (2014).

241 Guilinger, J. P., Thompson, D. B. & Liu, D. R. Fusion of catalytically inactive Cas9 to FokI nuclease improves the specificity of genome modification. *Nat Biotechnol* **32**, 577-582, doi:10.1038/nbt.2909 (2014).

242 Cornu, T. I., Mussolino, C. & Cathomen, T. Refining strategies to translate genome editing to the clinic. *Nat Med* **23**, 415-423, doi:10.1038/nm.4313 (2017).

243 Dunbar, C. E. *et al.* Gene therapy comes of age. *Science* **359**, doi:10.1126/science.aan4672 (2018).

244 Mandal, P. K. *et al.* Efficient ablation of genes in human hematopoietic stem and effector cells using CRISPR/Cas9. *Cell Stem Cell* **15**, 643-652, doi:10.1016/j.stem.2014.10.004 (2014).

245 Gundry, M. C. *et al.* Highly Efficient Genome Editing of Murine and Human Hematopoietic Progenitor Cells by CRISPR/Cas9. *Cell Rep* **17**, 1453-1461, doi:10.1016/j.celrep.2016.09.092 (2016).

246 Bak, R. O. *et al.* Multiplexed genetic engineering of human hematopoietic stem and progenitor cells using CRISPR/Cas9 and AAV6. *Elife* **6**, doi:10.7554/eLife.27873 (2017).

247 Dever, D. P. *et al.* CRISPR/Cas9 beta-globin gene targeting in human haematopoietic stem cells. *Nature* **539**, 384-389, doi:10.1038/nature20134 (2016).

248 Bak, R. O., Dever, D. P. & Porteus, M. H. CRISPR/Cas9 genome editing in human hematopoietic stem cells. *Nat Protoc* **13**, 358-376, doi:10.1038/nprot.2017.143 (2018).

249 De Ravin, S. S. *et al.* Targeted gene addition in human CD34(+) hematopoietic cells for correction of X-linked chronic granulomatous disease. *Nat Biotechnol* **34**, 424-429, doi:10.1038/nbt.3513 (2016).

250 Sather, B. D. *et al.* Efficient modification of CCR5 in primary human hematopoietic cells using a megaTAL nuclease and AAV donor template. *Sci Transl Med* **7**, 307ra156, doi:10.1126/scitranslmed.aac5530 (2015).

251 Kuo, C. Y. *et al.* Site-Specific Gene Editing of Human Hematopoietic Stem Cells for X-Linked Hyper-IgM Syndrome. *Cell Rep* **23**, 2606-2616, doi:10.1016/j.celrep.2018.04.103 (2018).

252 Pavel-Dinu, M. *et al.* Gene correction for SCID-X1 in long-term hematopoietic stem cells. *Nat Commun* **10**, 1634, doi:10.1038/s41467-019-09614-y (2019).

253 Tran, N. T. *et al.* Efficient CRISPR/Cas9-Mediated Gene Knockin in Mouse Hematopoietic Stem and Progenitor Cells. *Cell Rep* **28**, 3510-3522 e3515, doi:10.1016/j.celrep.2019.08.065 (2019).

254 Becker, S. *et al.* Correction of respiratory burst activity in X-linked chronic granulomatous cells to therapeutically relevant levels after gene transfer into bone marrow CD34+ cells. *Hum Gene Ther* **9**, 1561-1570, doi:10.1089/hum.1998.9.11-1561 (1998).

255 Gotzhein, F. *et al.* Age-independent influence of hematopoietic stem and progenitor cell populations during hematopoietic reconstitution. *bioRxiv*, 2020.2008.2005.237891, doi:10.1101/2020.08.05.237891 (2020).

256 Stemmer, M., Thumberger, T., Del Sol Keyer, M., Wittbrodt, J. & Mateo, J. L. CCTop: An Intuitive, Flexible and Reliable CRISPR/Cas9 Target Prediction Tool. *PLoS One* **10**, e0124633, doi:10.1371/journal.pone.0124633 (2015).

257 Concordet, J. P. & Haeussler, M. CRISPOR: intuitive guide selection for CRISPR/Cas9 genome editing experiments and screens. *Nucleic Acids Res* **46**, W242-W245, doi:10.1093/nar/gky354 (2018).

258 Chen, W. *et al.* Massively parallel profiling and predictive modeling of the outcomes of CRISPR/Cas9-mediated double-strand break repair. *Nucleic Acids Res* **47**, 7989-8003, doi:10.1093/nar/gkz487 (2019).

259 Doench, J. G. *et al.* Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nat Biotechnol* **34**, 184-191, doi:10.1038/nbt.3437 (2016).

260 Fusi, N., Smith, I., Doench, J. & Listgarten, J. In Silico Predictive Modeling of CRISPR/Cas9 guide efficiency. *bioRxiv*, 021568, doi:10.1101/021568 (2015).

261 Crooks, G. E., Hon, G., Chandonia, J. M. & Brenner, S. E. WebLogo: a sequence logo generator. *Genome Res* **14**, 1188-1190, doi:10.1101/gr.849004 (2004).

262 Mantri, S. *et al.* CD34 expression does not correlate with immunophenotypic stem cell or progenitor content in human cord blood products. *Blood Adv* **4**, 5357-5361, doi:10.1182/bloodadvances.2020002891 (2020).

263 Mendelson, A. & Frenette, P. S. Hematopoietic stem cell niche maintenance during homeostasis and regeneration. *Nat Med* **20**, 833-846, doi:10.1038/nm.3647 (2014).

264 Kumar, R., Fossati, V., Israel, M. & Snoeck, H. W. Lin-Sca1+kit- bone marrow cells contain early lymphoid-committed precursors that are distinct from common lymphoid progenitors. *J Immunol* **181**, 7507-7513, doi:10.4049/jimmunol.181.11.7507 (2008).

265 Yeoh, J. S., Ausema, A., Wierenga, P., de Haan, G. & van Os, R. Mobilized peripheral blood stem cells provide rapid reconstitution but impaired long-term engraftment in a mouse model. *Bone Marrow Transplant* **39**, 401-409, doi:10.1038/sj.bmt.1705601 (2007).

266 Roberts, A. W. *et al.* Genetic influences determining progenitor cell mobilization and leukocytosis induced by granulocyte colony-stimulating factor. *Blood* **89**, 2736-2744 (1997).

267 de Haan, G., Ausema, A., Wilkens, M., Molineux, G. & Dontje, B. Efficient mobilization of haematopoietic progenitors after a single injection of pegylated recombinant human granulocyte colony-stimulating factor in mouse strains with distinct marrow-cell pool sizes. *Br J Haematol* **110**, 638-646, doi:10.1046/j.1365-2141.2000.02252.x (2000).

268 Bonig, H. & Papayannopoulou, T. Mobilization of hematopoietic stem/progenitor cells: general principles and molecular mechanisms. *Methods Mol Biol* **904**, 1-14, doi:10.1007/978-1-61779-943-3_1 (2012).

269 Randall, T. D. & Weissman, I. L. Phenotypic and functional changes induced at the clonal level in hematopoietic stem cells after 5-fluorouracil treatment. *Blood* **89**, 3596-3606 (1997).

270 Rollini, P., Faes-Van't Hull, E., Kaiser, S., Kapp, U. & Leyvraz, S. Phenotypic and functional analysis of human fetal liver hematopoietic stem cells in culture. *Stem Cells Dev* **16**, 281-296, doi:10.1089/scd.2006.0096 (2007).

271 Bernitz, J. M., Daniel, M. G., Fstkchyan, Y. S. & Moore, K. Granulocyte colony-stimulating factor mobilizes dormant hematopoietic stem cells without proliferation in mice. *Blood* **129**, 1901-1912, doi:10.1182/blood-2016-11-752923 (2017).

272 Fry, T. J. & Mackall, C. L. Immune reconstitution following hematopoietic progenitor cell transplantation: challenges for the future. *Bone Marrow Transplant* **35 Suppl 1**, S53-57, doi:10.1038/sj.bmt.1704848 (2005).

273 Wilkinson, F. L. *et al.* Busulfan conditioning enhances engraftment of hematopoietic donor-derived cells in the brain compared with irradiation. *Mol Ther* **21**, 868-876, doi:10.1038/mt.2013.29 (2013).

274 Shemer, A. *et al.* Engrafted parenchymal brain macrophages differ from microglia in transcriptome, chromatin landscape and response to challenge. *Nat Commun* **9**, 5206, doi:10.1038/s41467-018-07548-5 (2018).

275 Capotondo, A. *et al.* Brain conditioning is instrumental for successful microglia reconstitution following hematopoietic stem cell transplantation. *Proc Natl Acad Sci U S A* **109**, 15018-15023, doi:10.1073/pnas.1205858109 (2012).

276 Hassan, M. & Ehrsson, H. Degradation of busulfan in aqueous solution. *J Pharm Biomed Anal* **4**, 95-101, doi:10.1016/0731-7085(86)80027-9 (1986).

277 Karstens, A. & Kramer, I. [Stability of busulfan injection solution (Busilvex, Busulfex) in B/Braun Injekt syringes]. *Pharmazie* **61**, 845-850 (2006).

278 Houot, M. *et al.* Physico-chemical stability of busulfan in injectable solutions in various administration packages. *Drugs R D* **13**, 87-94, doi:10.1007/s40268-013-0003-y (2013).

279 Hassan, M. *et al.* Busulfan bioavailability. *Blood* **84**, 2144-2150 (1994).

280 Grochow, L. B. *et al.* Pharmacokinetics of busulfan: correlation with veno-occlusive disease in patients undergoing bone marrow transplantation. *Cancer Chemother Pharmacol* **25**, 55-61, doi:10.1007/BF00694339 (1989).

281 Blazar, B. R. *et al.* Pretransplant conditioning with busulfan (Myleran) and cyclophosphamide for nonmalignant diseases. Assessment of engraftment following histocompatible allogeneic bone marrow transplantation. *Transplantation* **39**, 597-603, doi:10.1097/00007890-198506000-00004 (1985).

282 Shaw, P. J., Hugh-Jones, K., Hobbs, J. R., Downie, C. J. & Barnes, R. Busulphan and cyclophosphamide cause little early toxicity during displacement bone marrow transplantation in fifty children. *Bone Marrow Transplant* **1**, 193-200 (1986).

283 Morley, A. & Blake, J. An animal model of chronic aplastic marrow failure. I. Late marrow failure after busulfan. *Blood* **44**, 49-56 (1974).

284 Yeager, A. M., Shinn, C. & Pardoll, D. M. Lymphoid reconstitution after transplantation of congenic hematopoietic cells in busulfan-treated mice. *Blood* **78**, 3312-3316 (1991).

285 Yeager, A. M., Shinn, C., Shinohara, M. & Pardoll, D. M. Hematopoietic cell transplantation in the twitcher mouse. The effects of pretransplant conditioning with graded doses of busulfan. *Transplantation* **56**, 185-190, doi:10.1097/00007890-199307000-00034 (1993).

286 Yoder, M. C., Cumming, J. G., Hiatt, K., Mukherjee, P. & Williams, D. A. A novel method of myeloablation to enhance engraftment of adult bone marrow cells in newborn mice. *Biol Blood Marrow Transplant* **2**, 59-67 (1996).

287 Down, J. D., Boudewijn, A., Dillingh, J. H., Fox, B. W. & Ploemacher, R. E. Relationships between ablation of distinct haematopoietic cell subsets and the development of donor bone marrow engraftment following recipient pretreatment with different alkylating drugs. *Br J Cancer* **70**, 611-616, doi:10.1038/bjc.1994.359 (1994).

288 Down, J. D. & Ploemacher, R. E. Transient and permanent engraftment potential of murine hematopoietic stem cell subsets: differential effects of host conditioning with gamma radiation and cytotoxic drugs. *Exp Hematol* **21**, 913-921 (1993).

289 Kuramoto, K. *et al.* The impact of low-dose busulfan on clonal dynamics in nonhuman primates. *Blood* **104**, 1273-1280, doi:10.1182/blood-2003-08-2935 (2004).

290 Fehse, B., Kustikova, O. S., Bubenheim, M. & Baum, C. Pois(s)on--it's a question of dose. *Gene Ther* **11**, 879-881, doi:10.1038/sj.gt.3302270 (2004).

291 Kustikova, O. S. *et al.* Retroviral vector insertion sites associated with dominant hematopoietic clones mark "stemness" pathways. *Blood* **109**, 1897-1907, doi:10.1182/blood-2006-08-044156 (2007).

292 McKenzie, J. L., Gan, O. I., Doedens, M., Wang, J. C. & Dick, J. E. Individual stem cells with highly variable proliferation and self-renewal properties comprise the human hematopoietic stem cell compartment. *Nat Immunol* **7**, 1225-1233, doi:10.1038/ni1393 (2006).

293 Gonzalez-Murillo, A., Lozano, M. L., Montini, E., Bueren, J. A. & Guenechea, G. Unaltered repopulation properties of mouse hematopoietic stem cells transduced with lentiviral vectors. *Blood* **112**, 3138-3147, doi:10.1182/blood-2008-03-142661 (2008).

294 Montini, E. *et al.* Hematopoietic stem cell gene transfer in a tumor-prone mouse model uncovers low genotoxicity of lentiviral vector integration. *Nat Biotechnol* **24**, 687-696, doi:10.1038/nbt1216 (2006).

295 Kim, S. *et al.* Dynamics of HSPC repopulation in nonhuman primates revealed by a decade-long clonal-tracking study. *Cell Stem Cell* **14**, 473-485, doi:10.1016/j.stem.2013.12.012 (2014).

296 Brenner, S. *et al.* Concentrated RD114-pseudotyped MFGS-gp91phox vector achieves high levels of functional correction of the chronic granulomatous disease oxidase defect in NOD/SCID/beta -microglobulin-/- repopulating mobilized human peripheral blood CD34+ cells. *Blood* **102**, 2789-2797, doi:10.1182/blood-2002-05-1482 (2003).

297 Zhou, J., Lin, J., Zhou, C., Deng, X. & Xia, B. Cytotoxicity of red fluorescent protein DsRed is associated with the suppression of Bcl-xL translation. *FEBS Lett* **585**, 821-827, doi:10.1016/j.febslet.2011.02.013 (2011).

298 Snaith, H. A., Anders, A., Samejima, I. & Sawin, K. E. New and old reagents for fluorescent protein tagging of microtubules in fission yeast; experimental and critical evaluation. *Methods Cell Biol* **97**, 147-172, doi:10.1016/S0091-679X(10)97009-X (2010).

299 Fleischmann, M. *et al.* Cardiac specific expression of the green fluorescent protein during early murine embryonic development. *FEBS Lett* **440**, 370-376, doi:10.1016/s0014-5793(98)01476-8 (1998).

300 Li, J. L. *et al.* Efferent and afferent connections of GABAergic neurons in the supratrigeminal and the intertrigeminal regions. An immunohistochemical tract-tracing study in the GAD67-GFP knock-in mouse. *Neurosci Res* **51**, 81-91, doi:10.1016/j.neures.2004.10.005 (2005).

301 Yabe, I. M. *et al.* Barcoding of Macaque Hematopoietic Stem and Progenitor Cells: A Robust Platform to Assess Vector Genotoxicity. *Mol Ther Methods Clin Dev* **11**, 143-154, doi:10.1016/j.omtm.2018.10.009 (2018).

302 Weber, K., Bartsch, U., Stocking, C. & Fehse, B. A multicolor panel of novel lentiviral "gene ontology" (LeGO) vectors for functional gene analysis. *Mol Ther* **16**, 698-706, doi:10.1038/mt.2008.6 (2008).

303 Schambach, A. *et al.* Overcoming promoter competition in packaging cells improves production of self-inactivating retroviral vectors. *Gene Ther* **13**, 1524-1533, doi:10.1038/sj.gt.3302807 (2006).

304 Baum, C., Hegewisch-Becker, S., Eckert, H. G., Stocking, C. & Ostertag, W. Novel retroviral vectors for efficient expression of the multidrug resistance (mdr-1) gene in early hematopoietic cells. *J Virol* **69**, 7541-7547, doi:10.1128/JVI.69.12.7541-7547.1995 (1995).

305 Pinheiro, L. B. *et al.* Evaluation of a droplet digital polymerase chain reaction format for DNA copy number quantification. *Anal Chem* **84**, 1003-1011, doi:10.1021/ac202578x (2012).

306 Oxnard, G. R. *et al.* Noninvasive detection of response and resistance in EGFR-mutant lung cancer using quantitative next-generation genotyping of cell-free plasma DNA. *Clin Cancer Res* **20**, 1698-1705, doi:10.1158/1078-0432.CCR-13-2482 (2014).

307 Taly, V. *et al.* Multiplex picodroplet digital PCR to detect KRAS mutations in circulating DNA from the plasma of colorectal cancer patients. *Clin Chem* **59**, 1722-1731, doi:10.1373/clinchem.2013.206359 (2013).

308 Deakin, C. T. *et al.* Impact of next-generation sequencing error on analysis of barcoded plasmid libraries of known complexity and sequence. *Nucleic Acids Res* **42**, e129, doi:10.1093/nar/gku607 (2014).

309 Cheung, A. M. *et al.* Analysis of the clonal growth and differentiation dynamics of primitive barcoded human cord blood cells in NSG mice. *Blood* **122**, 3129-3137, doi:10.1182/blood-2013-06-508432 (2013).

310 Dykstra, B., Olthof, S., Schreuder, J., Ritsema, M. & de Haan, G. Clonal analysis reveals multiple functional defects of aged murine hematopoietic stem cells. *J Exp Med* **208**, 2691-2703, doi:10.1084/jem.20111490 (2011).

311 Beltman, J. B. *et al.* Reproducibility of Illumina platform deep sequencing errors allows accurate determination of DNA barcodes in cells. *BMC Bioinformatics* **17**, 151, doi:10.1186/s12859-016-0999-4 (2016).

312 Xu, Q., Schlabach, M. R., Hannon, G. J. & Elledge, S. J. Design of 240,000 orthogonal 25mer DNA barcode probes. *Proc Natl Acad Sci U S A* **106**, 2289-2294, doi:10.1073/pnas.0812506106 (2009).

313 Buschmann, T. & Bystrykh, L. V. Levenshtein error-correcting barcodes for multiplexed DNA sequencing. *BMC Bioinformatics* **14**, 272, doi:10.1186/1471-2105-14-272 (2013).

314 Miner, B. E., Stoger, R. J., Burden, A. F., Laird, C. D. & Hansen, R. S. Molecular barcodes detect redundancy and contamination in hairpin-bisulfite PCR. *Nucleic Acids Res* **32**, e135, doi:10.1093/nar/gnh132 (2004).

315 Kou, R. *et al.* Benefits and Challenges with Applying Unique Molecular Identifiers in Next Generation Sequencing to Detect Low Frequency Mutations. *PLoS One* **11**, e0146638, doi:10.1371/journal.pone.0146638 (2016).

316 Hendrikx, P. J., Martens, C. M., Hagenbeek, A., Keij, J. F. & Visser, J. W. Homing of fluorescently labeled murine hematopoietic stem cells. *Exp Hematol* **24**, 129-140 (1996).

317 Lanzkron, S. M., Collector, M. I. & Sharkis, S. J. Homing of long-term and short-term engrafting cells in vivo. *Ann N Y Acad Sci* **872**, 48-54; discussion 54-46, doi:10.1111/j.1749-6632.1999.tb08452.x (1999).

318 van der Loo, J. C. & Ploemacher, R. E. Marrow- and spleen-seeding efficiencies of all murine hematopoietic stem cell subsets are decreased by preincubation with hematopoietic growth factors. *Blood* **85**, 2598-2606 (1995).

319 Yamamoto, K. *et al.* Extramedullary hematopoiesis: Elucidating the function of the hematopoietic stem cell niche (Review). *Mol Med Rep* **13**, 587-591, doi:10.3892/mmr.2015.4621 (2016).

320 Cao, Y. A. *et al.* Shifting foci of hematopoiesis during reconstitution from single stem cells. *Proc Natl Acad Sci U S A* **101**, 221-226, doi:10.1073/pnas.2637010100 (2004).

321 Cabezas-Wallscheid, N. *et al.* Identification of regulatory networks in HSCs and their immediate progeny via integrated proteome, transcriptome, and DNA methylome analysis. *Cell Stem Cell* **15**, 507-522, doi:10.1016/j.stem.2014.07.005 (2014).

322 van den Brink, M. R., Velardi, E. & Perales, M. A. Immune reconstitution following stem cell transplantation. *Hematology Am Soc Hematol Educ Program* **2015**, 215-219, doi:10.1182/asheducation-2015.1.215 (2015).

323 Storek, J. *et al.* Reconstitution of the immune system after hematopoietic stem cell transplantation in humans. *Semin Immunopathol* **30**, 425-437, doi:10.1007/s00281-008-0132-5 (2008).

324 Richie Ehrlich, L. I., Serwold, T. & Weissman, I. L. In vitro assays misrepresent in vivo lineage potentials of murine lymphoid progenitors. *Blood* **117**, 2618-2624, doi:10.1182/blood-2010-05-287102 (2011).

325 Biasco, L. *et al.* In Vivo Tracking of Human Hematopoiesis Reveals Patterns of Clonal Dynamics during Early and Steady-State Reconstitution Phases. *Cell Stem Cell* **19**, 107-119, doi:10.1016/j.stem.2016.04.016 (2016).

326 Case, S. S. *et al.* Stable transduction of quiescent CD34(+)CD38(-) human hematopoietic cells by HIV-1-based lentiviral vectors. *Proc Natl Acad Sci U S A* **96**, 2988-2993, doi:10.1073/pnas.96.6.2988 (1999).

327 Ishikawa, F. *et al.* Development of functional human blood and immune systems in NOD/SCID/IL2 receptor {gamma} chain(null) mice. *Blood* **106**, 1565-1573, doi:10.1182/blood-2005-02-0516 (2005).

328 Scala, S. *et al.* Dynamics of genetically engineered hematopoietic stem and progenitor cells after autologous transplantation in humans. *Nat Med* **24**, 1683-1690, doi:10.1038/s41591-018-0195-3 (2018).

329 Irion, S. *et al.* Identification and targeting of the ROSA26 locus in human embryonic stem cells. *Nat Biotechnol* **25**, 1477-1482, doi:10.1038/nbt1362 (2007).

330 Liu, R. *et al.* Homozygous defect in HIV-1 coreceptor accounts for resistance of some multiply-exposed individuals to HIV-1 infection. *Cell* **86**, 367-377, doi:10.1016/s0092-8674(00)80110-5 (1996).

331 Kotin, R. M., Linden, R. M. & Berns, K. I. Characterization of a preferred site on human chromosome 19q for integration of adeno-associated virus DNA by non-homologous recombination. *EMBO J* **11**, 5071-5078 (1992).

332 Perez, E. E. *et al.* Establishment of HIV-1 resistance in CD4+ T cells by genome editing using zinc-finger nucleases. *Nat Biotechnol* **26**, 808-816, doi:10.1038/nbt1410 (2008).

333 Friedrich, G. & Soriano, P. Promoter traps in embryonic stem cells: a genetic screen to identify and mutate developmental genes in mice. *Genes Dev* **5**, 1513-1523, doi:10.1101/gad.5.9.1513 (1991).

334 Aubrecht, J., Goad, M. E. & Schiestl, R. H. Tissue specific toxicities of the anticancer drug 6-thioguanine is dependent on the Hprt status in transgenic mice. *J Pharmacol Exp Ther* **282**, 1102-1108 (1997).

335 Hacke, K. *et al.* Combined preconditioning and in vivo chemoselection with 6-thioguanine alone achieves highly efficient reconstitution of normal hematopoiesis with HPRT-deficient bone marrow. *Exp Hematol* **40**, 3-13 e13, doi:10.1016/j.exphem.2011.09.009 (2012).

336 Choudhary, R., Baturin, D., Fosmire, S., Freed, B. & Porter, C. C. Knockdown of HPRT for selection of genetically modified human hematopoietic progenitor cells. *PLoS One* **8**, e59594, doi:10.1371/journal.pone.0059594 (2013).

337 Torres, R. J. & Puig, J. G. Hypoxanthine-guanine phosophoribosyltransferase (HPRT) deficiency: Lesch-Nyhan syndrome. *Orphanet J Rare Dis* **2**, 48, doi:10.1186/1750-1172-2-48 (2007).

338 Ansell, J. D. *et al.* Hypoxanthine phosphoribosyl transferase deficiency, haematopoiesis and fertility in the mouse. *Development* **112**, 489-498 (1991).

339 Vogel, M. *et al.* HPRT and Purine Salvaging Are Critical for Hematopoietic Stem Cell Function. *Stem Cells* **37**, 1606-1614, doi:10.1002/stem.3087 (2019).

340 Liao, S., Tammaro, M. & Yan, H. Enriching CRISPR-Cas9 targeted cells by co-targeting the HPRT gene. *Nucleic Acids Res* **43**, e134, doi:10.1093/nar/gkv675 (2015).

341 Brinkman, E. K., Chen, T., Amendola, M. & van Steensel, B. Easy quantitative assessment of genome editing by sequence trace decomposition. *Nucleic Acids Res* **42**, e168, doi:10.1093/nar/gku936 (2014).

342 Bayer, M. *et al.* A large U3 deletion causes increased in vivo expression from a nonintegrating lentiviral vector. *Mol Ther* **16**, 1968-1976, doi:10.1038/mt.2008.199 (2008).
343 Lombardo, A. *et al.* Gene editing in human stem cells using zinc finger nucleases and integrase-defective lentiviral vector delivery. *Nat Biotechnol* **25**, 1298-1306, doi:10.1038/nbt1353 (2007).
344 Saenz, D. T. *et al.* Unintegrated lentivirus DNA persistence and accessibility to expression in nondividing cells: analysis with class I integrase mutants. *J Virol* **78**, 2906-2920, doi:10.1128/jvi.78.6.2906-2920.2004 (2004).
345 Yanez-Munoz, R. J. *et al.* Effective gene therapy with nonintegrating lentiviral vectors. *Nat Med* **12**, 348-353, doi:10.1038/nm1365 (2006).
346 Okada, Y. *et al.* Targeted gene modification in mouse ES cells using integrase-defective lentiviral vectors. *Genesis* **47**, 217-223, doi:10.1002/dvg.20469 (2009).
347 Vargas, J., Jr., Gusella, G. L., Najfeld, V., Klotman, M. E. & Cara, A. Novel integrase-defective lentiviral episomal vectors for gene transfer. *Hum Gene Ther* **15**, 361-372, doi:10.1089/104303404322959515 (2004).
348 Philippe, S. *et al.* Lentiviral vectors with a defective integrase allow efficient and sustained transgene expression in vitro and in vivo. *Proc Natl Acad Sci U S A* **103**, 17684-17689, doi:10.1073/pnas.0606197103 (2006).
349 Sengupta, R. *et al.* An Optimized Protocol for Packaging Pseudotyped Integrase Defective Lentivirus. *Biol Proced Online* **18**, 14, doi:10.1186/s12575-016-0044-z (2016).
350 Wright, A. V. *et al.* Rational design of a split-Cas9 enzyme complex. *Proc Natl Acad Sci U S A* **112**, 2984-2989, doi:10.1073/pnas.1501698112 (2015).
351 Zetsche, B., Volz, S. E. & Zhang, F. A split-Cas9 architecture for inducible genome editing and transcription modulation. *Nat Biotechnol* **33**, 139-142, doi:10.1038/nbt.3149 (2015).
352 Truong, D. J. *et al.* Development of an intein-mediated split-Cas9 system for gene therapy. *Nucleic Acids Res* **43**, 6450-6458, doi:10.1093/nar/gkv601 (2015).
353 Ye, L. *et al.* Genome editing using CRISPR-Cas9 to create the HPFH genotype in HSPCs: An approach for treating sickle cell disease and beta-thalassemia. *Proc Natl Acad Sci U S A* **113**, 10661-10665, doi:10.1073/pnas.1612075113 (2016).
354 Friedland, A. E. *et al.* Characterization of Staphylococcus aureus Cas9: a smaller Cas9 for all-in-one adeno-associated virus delivery and paired nickase applications. *Genome Biol* **16**, 257, doi:10.1186/s13059-015-0817-8 (2015).
355 Ran, F. A. *et al.* In vivo genome editing using Staphylococcus aureus Cas9. *Nature* **520**, 186-191, doi:10.1038/nature14299 (2015).
356 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, doi:10.1128/JVI.70.2.721-728.1996 (1996).
357 Nightingale, S. J. *et al.* Transient gene expression by nonintegrating lentiviral vectors. *Mol Ther* **13**, 1121-1132, doi:10.1016/j.ymthe.2006.01.008 (2006).
358 Kantor, B. *et al.* Notable reduction in illegitimate integration mediated by a PPT-deleted, nonintegrating lentiviral vector. *Mol Ther* **19**, 547-556, doi:10.1038/mt.2010.277 (2011).
359 Wang, Y., Wang, Y., Chang, T., Huang, H. & Yee, J. K. Integration-defective lentiviral vector mediates efficient gene editing through homology-directed repair in human embryonic stem cells. *Nucleic Acids Res* **45**, e29, doi:10.1093/nar/gkw1057 (2017).
360 Volsky, D. J. *et al.* Interference to human immunodeficiency virus type 1 infection in the absence of downmodulation of the principal virus receptor, CD4. *J Virol* **70**, 3823-3833, doi:10.1128/JVI.70.6.3823-3833.1996 (1996).
361 Bernier, R. & Tremblay, M. Homologous interference resulting from the presence of defective particles of human immunodeficiency virus type 1. *J Virol* **69**, 291-300, doi:10.1128/JVI.69.1.291-300.1995 (1995).
362 Schott, J. W. *et al.* Deciphering the impact of parameters influencing transgene expression kinetics after repeated cell transduction with integration-deficient retroviral vectors. *Cytometry A* **87**, 405-418, doi:10.1002/cyto.a.22650 (2015).
363 Frimpong, K. & Spector, S. A. Cotransduction of nondividing cells using lentiviral vectors. *Gene Ther* **7**, 1562-1569, doi:10.1038/sj.gt.3301283 (2000).
364 Aryal, N. K., Wasylishen, A. R. & Lozano, G. CRISPR/Cas9 can mediate high-efficiency off-target mutations in mice in vivo. *Cell Death Dis* **9**, 1099, doi:10.1038/s41419-018-1146-0 (2018).
365 Iyer, V. *et al.* Off-target mutations are rare in Cas9-modified mice. *Nat Methods* **12**, 479, doi:10.1038/nmeth.3408 (2015).

366 Enache, O. M. *et al.* Cas9 activates the p53 pathway and selects for p53-inactivating mutations. *Nat Genet* **52**, 662-668, doi:10.1038/s41588-020-0623-4 (2020).
367 Haapaniemi, E., Botla, S., Persson, J., Schmierer, B. & Taipale, J. CRISPR-Cas9 genome editing induces a p53-mediated DNA damage response. *Nat Med* **24**, 927-930, doi:10.1038/s41591-018-0049-z (2018).
368 Lundin, A. *et al.* Development of an ObLiGaRe Doxycycline Inducible Cas9 system for pre-clinical cancer drug discovery. *Nat Commun* **11**, 4903, doi:10.1038/s41467-020-18548-9 (2020).
369 Katigbak, A., Robert, F., Paquet, M. & Pelletier, J. Inducible Genome Editing with Conditional CRISPR/Cas9 Mice. *G3 (Bethesda)* **8**, 1627-1635, doi:10.1534/g3.117.300327 (2018).
370 Berman, J. R. *et al.* Ultra-Sensitive Quantification of Genome Editing Events Using Droplet Digital™ PCR. *Droplet Digital PCR: Bulletin 6712* (2015).
371 De Ravin, S. S. *et al.* CRISPR-Cas9 gene repair of hematopoietic stem cells from patients with X-linked chronic granulomatous disease. *Sci Transl Med* **9**, doi:10.1126/scitranslmed.aah3480 (2017).
372 Lomova, A. *et al.* Improving Gene Editing Outcomes in Human Hematopoietic Stem and Progenitor Cells by Temporal Control of DNA Repair. *Stem Cells* **37**, 284-294, doi:10.1002/stem.2935 (2019).
373 Passegue, E., Wagers, A. J., Giuriato, S., Anderson, W. C. & Weissman, I. L. Global analysis of proliferation and cell cycle gene expression in the regulation of hematopoietic stem and progenitor cell fates. *J Exp Med* **202**, 1599-1611, doi:10.1084/jem.20050967 (2005).
374 Branzei, D. & Foiani, M. Regulation of DNA repair throughout the cell cycle. *Nat Rev Mol Cell Biol* **9**, 297-308, doi:10.1038/nrm2351 (2008).
375 Sonoda, E., Hochegger, H., Saberi, A., Taniguchi, Y. & Takeda, S. Differential usage of non-homologous end-joining and homologous recombination in double strand break repair. *DNA Repair (Amst)* **5**, 1021-1029, doi:10.1016/j.dnarep.2006.05.022 (2006).
376 Lattanzi, A. *et al.* Optimization of CRISPR/Cas9 Delivery to Human Hematopoietic Stem and Progenitor Cells for Therapeutic Genomic Rearrangements. *Mol Ther* **27**, 137-150, doi:10.1016/j.ymthe.2018.10.008 (2019).
377 Aoki, T., Miyauchi, K., Urano, E., Ichikawa, R. & Komano, J. Protein transduction by pseudotyped lentivirus-like nanoparticles. *Gene Ther* **18**, 936-941, doi:10.1038/gt.2011.38 (2011).
378 Voelkel, C. *et al.* Protein transduction from retroviral Gag precursors. *Proc Natl Acad Sci U S A* **107**, 7805-7810, doi:10.1073/pnas.0914517107 (2010).
379 Cai, Y. *et al.* DNA transposition by protein transduction of the piggyBac transposase from lentiviral Gag precursors. *Nucleic Acids Res* **42**, e28, doi:10.1093/nar/gkt1163 (2014).
380 Mangeot, P. E. *et al.* Genome editing in primary cells and in vivo using viral-derived Nanoblades loaded with Cas9-sgRNA ribonucleoproteins. *Nat Commun* **10**, 45, doi:10.1038/s41467-018-07845-z (2019).
381 Gee, P. *et al.* Extracellular nanovesicles for packaging of CRISPR-Cas9 protein and sgRNA to induce therapeutic exon skipping. *Nat Commun* **11**, 1334, doi:10.1038/s41467-020-14957-y (2020).
382 Campbell, L. A. *et al.* Gesicle-Mediated Delivery of CRISPR/Cas9 Ribonucleoprotein Complex for Inactivating the HIV Provirus. *Mol Ther* **27**, 151-163, doi:10.1016/j.ymthe.2018.10.002 (2019).
383 Montagna, C. *et al.* VSV-G-Enveloped Vesicles for Traceless Delivery of CRISPR-Cas9. *Mol Ther Nucleic Acids* **12**, 453-462, doi:10.1016/j.omtn.2018.05.010 (2018).
384 Heckl, D. *et al.* Generation of mouse models of myeloid malignancy with combinatorial genetic lesions using CRISPR-Cas9 genome editing. *Nat Biotechnol* **32**, 941-946, doi:10.1038/nbt.2951 (2014).

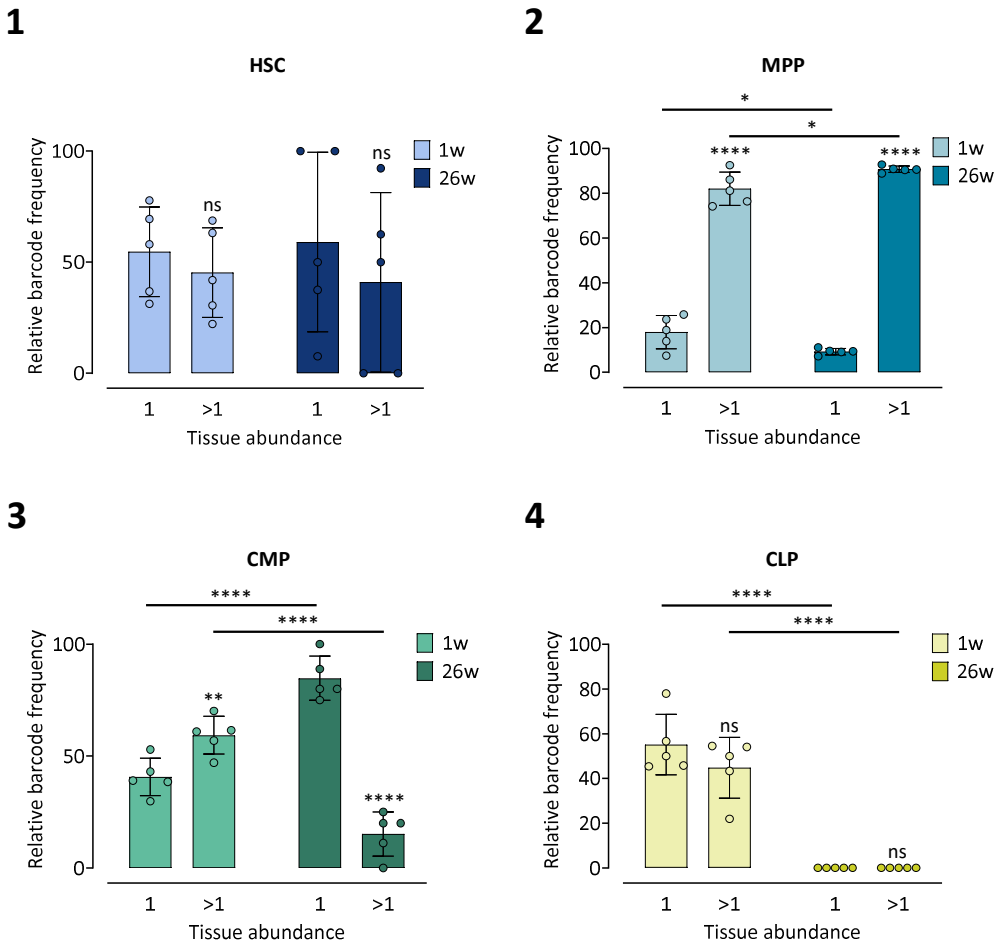
10 Appendix

10.1 Appendix A: Barcode content from NGS analyses

Table 25: Barcode content over time. In the BM-TBI group barcodes were analyzed via NGS and, after data processing and error correction, the count of different barcodes (and sum of all read counts) per sample were calculated. Final time points for analysis were 1w, 3w, 8w, 16w, and 26w post transplantation. Samples were taken from BM, PB, spleen, and thymus. HSPC-specific barcode backbones, according to the initially marked cell population were used to track cells derived from HSC, MPP, CMP, and CLP population. The total number of barcodes per week is presented in the last row. Mean values were calculated from n=5 animals per time point.
nd: not determined

Time [weeks]	**HSC** count (sum)	**MPP** count (sum)	**CMP** count (sum)	**CLP** count (sum)	**Total** count
			BM		
1	21 (5296)	542 (166014)	292 (25824)	48 (6693)	902
3	15 (1915)	171 (115733)	12 (818)	5 (572)	201
8	13 (28416)	82 (114246)	8 (158)	3 (63)	105
16	7 (14012)	59 (299920)	3 (132)	1 (33)	68
26	14 (147096)	94 (74914)	3 (118)	2 (2)	113
			PB		
1	35 (3116)	152 (32117)	3888 (45398)	25 (7529)	4099
3	nd	nd	nd	nd	nd
8	20 (9597)	164 (52149)	945 (71948)	7 (85)	1134
16	5 (2911)	38 (37201)	5 (1382)	3 (257)	50
26	7 (43383)	40 (17156)	3 (309)	3 (228)	52
			spleen		
1	36 (21907)	888 (252348)	475 (147945)	54 (53423)	1451
3	13 (4458)	166 (99513)	11 (381)	4 (297)	192
8	10 (14559)	79 (133901)	4 (273)	2 (27)	93
16	6 (4740)	76 (126012)	3 (86)	2 (119)	86
26	16 (246473)	113 (96495)	7 (313)	1 (1)	136
			thymus		
1	41 (2554)	387 (8720)	272 (10426)	28 (2351)	727
3	8 (60)	56 (2289)	9 (2836)	5 (319)	75
8	6 (93613)	42 (83423)	4 (98)	2 (20)	52
16	5 (159)	33 (75291)	2 (77)	2 (8)	41
26	16 (23958)	112 (91272)	6 (89)	3 (49)	135

10.2 Appendix B: Barcode overlap between compartments



Appendix B: Barcode overlap between tissue types. Relative barcode frequencies per barcode, abundant in one (1) ore several compartments (>1) after 1w or 26w post transplantation (mean ± SD, n=5). Analysis was performed for HSC- (A), MPP- (B), CMP- (C), and CLP-derived (D) barcode sequences and shows the contribution of the respective subpopulation to *multi-tissue* reconstitution.
CLP: common lymphoid progenitor, CMP: common myeloid progenitor, HSC: hematopoietic stem cell, MPP: multipotent progenitor
ns: not significant, * $p \leq 0.05$, ** $p \leq 0.001$, **** $p \leq 0.0001$

10.3 Appendix C: Mean titers of LV production

Table 26: Mean titers of LV productions. LV productions were performed for target and donor vectors, each as integrating (ILV) or non-integrating (IDLV) vectors. For each locus (*Rosa26* or *Hprt*), two different target vectors (ROSA1, ROSA2, HPRT1 and HPRT2) and one or two different donor vectors (ROSA1/2, HPRT1 and HPRT2) were produced. Titers [infectious particles/ mL] were calculated after serial transduction of the producer cell line with the supernatant volume leading to transduction of 5—20 % of the target cells. Mean values of infectious viral particles/ mL were calculated from several experiments

Site	ILV	IDLV
	Target vector	
ROSA1	9.98E+06	8.16E+05
ROSA2	1.24E+07	6.42E+05
HPRT1	2.33E+07	8.22E+05
HPRT2	2.87E+07	6.00E+05
	Donor vector	
ROSA1/2	5.65E+09	2.89E+07
HPRT1	2.90E+09	9.61E+06
HPRT2	1.18E+08	1.26E+07

10.4 Appendix C: Hazardous substances

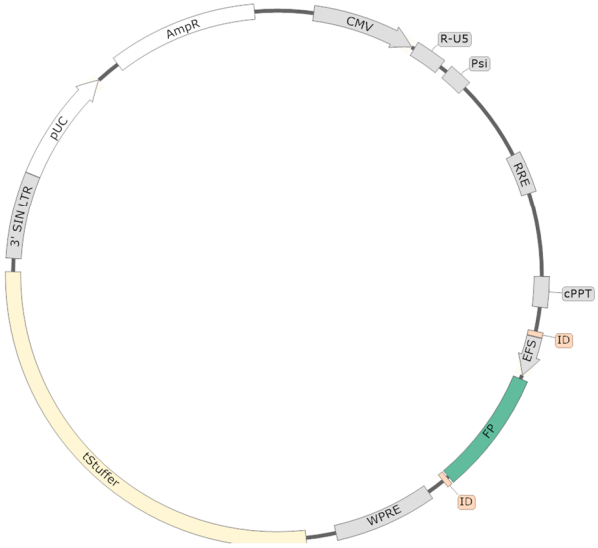
Table 27: Hazardous substances according to GHS. GHS Symbols and H and P statements are listed according to PubChem chemical information online database or the safety datasheet supplied by the manufacturer of the substance.

Substance	GHS symbol	H and P statements
AMD3100	7	H315, H319, P264, P280, P302+P352, P305+P351+P338, P321, P332+P313, P337+P313, and P362
Ammonium chloride	7	H302, H319, P264, P270, P280, P301+P312, P305+P351+P338, P330, P337+P313, and P501
Ampicillin	7, 8	H317, H334, P261, P272, P280, P285, P302+P352, P304+P341, P321, P333+P313, P342+P311, P363, and P501
Busulfan	6, 7, 8	H300, H301, H310, H315, H319, H330, H340, H350, H360 H373, P201, P202, P260, P261, P262, P264, P270, P271, P280, P281, P284, P301+P310, P302+P350, P302+P352, P304+P340, P305+P351+P338, P308+P313, P310, P312, P314, P320, P321, P322, P330, P332+P313, P337+P313, P361, P362, P363, P403+P233, P405, and P501
Calcium chloride	7	H319, P264, P280, P305+P351+P338, and P337+P3137
Chloroquine	7	H302, P264, P270, P301+P312, P330, and P501
Dimethyl sulfoxide	7	H315, H319, H335, P261, P264, P271, P280, P302+P352, P304+P340, P305+P351+P338, P312, P321, P332+P313, P337+P313, P362, P403+P233, P405, and P501
Disodium hydrogen phosphate	5, 7	H318, H319, P264, P280, P305+P351+P338, P310, and P337+P313
EDTA	7	H319, P264, P280, P305+P351+P338, and P337+P313
Enrofloxacin	7, 8, 9	H302, H317, H334, H361, H372, H400, H410, P201, P202, P260, P261, P264, P270, P272, P273, P280, P281, P285, P301+P312, P302+P352, P304+P341, P308+P313, P314, P321, P330, P333+P313, P342+P311, P363, P391, P405, and P501
Ethanol	2	H225, P210, P233, P240, P241, P242, P243, P280, P303+P361+P353, P370+P378, P403+P235, and P501
Ethidium bromide	6, 7, 8	H302, H330, H341, P201, P202, P260, P264, P270, P271, P281, P284, P301+P312, P304+P340, P308+P313, P310, P320, P330, P403+P233, P405, and P501
Formaldehyde	5, 6, 7, 8	H301, H311, H314, H317, H331, H341, H350, P201, P202, P260, P261, P264, P270, P271, P272, P280, P281, P301+P310, P301+P330+P331, P302+P352, P303+P361+P353, P304+P340, P305+P351+P338, P308+P313, P310, P311, P312, P321, P322, P330, P333+P313, P361, P363, P403+P233, P405, and P501
Guanidinium chloride	7	H302 + H332, H315, H319, and P280

Substance	GHS symbol	H and P statements
Guanidinium thiocyanat	5, 7	H302, H318, H412, P280, and P305 + P351 + P338 + P310
Isopropyl alcohol	2, 7	H225, H319, H336, P210, P233, P240, P241, P242, P243, P261, P264, P271, P280, P303+P361+P353, P304+P340, P305+P351+P338, P312, P337+P313, P370+P378, P403+P233, P403+P235, P405, and P501
Maleinsäure	7	H315, H317, H319, and P280
Penicillin/Streptomycin	7, 8	H317, H334, H360, P201, P202, P261, P280, P284, P302 + P352, P333 + P313, P342 + P311, P308 + P313, P362, and P304 + P340
Polybrene	7	H302, P264, P270, P301+P312, P330, and P501
Potassium bicarbonate	7	H315, H319, H335, P261, P264, P271, P280, P302+P352, P304+P340, P305+P351+P338, P312, P321, P332+P313, P337+P313, P362, P403+P233, P405, and P501
Proteinase K	8	H334, P261, P280, P284, P304 + P340, and P342 + P311
Sodium dodecyl sulfate	2, 5, 7	H228, H302, H315, H318, H319, H332, H335, H412, P210, P240, P241, P261, P264, P270, P271, P273, P280, P301+P312, P302+P352, P304+P312, P304+P340, P305+P351+P338, P310, P312, P321, P330, P332+P313, P337+P313, P362, P370+P378, P403+P233, P405, and P501
Sodium hydroxide solution	5	H314, P260, P264, P280, P301+P330+P331, P303+P361+P353, P304+P340, P305+P351+P338, P310, P321, P363, P405, and P501
Sodium pyruvate	7	H317. H319, P261, P264, P272, P280, P302+P352, P305+P351+P338, P321, P333+P313, P337+P313, P363, and P501
Tris-Borate-EDTA buffer	7	H315, H319, H335, P261, P264, P271, P280, P302+P352, P304+P340, P305+P351+P338, P312, P321, P332+P313, P337+P313, P362, P403+P233, P405, and P501

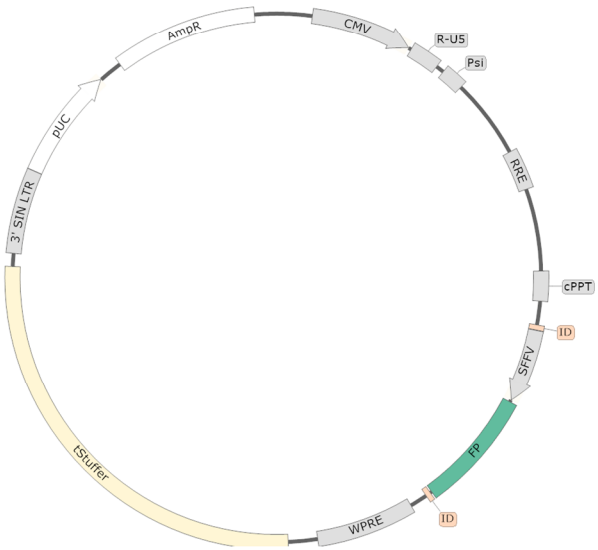
10.5 Appendix D: Plasmid maps

1



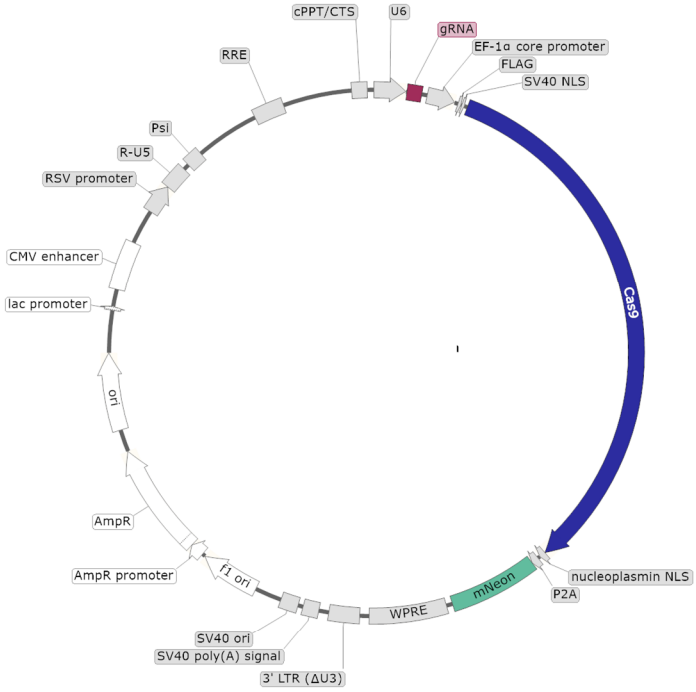
Appendix D-1: Plasmid map of lentiviral LE1-LE4 constructs. Plasmids contain ID sequences, EFS promoter-driven fluorescent marker (GFP, eBFP, T-Sapphire, or Venus) and a tStuffer. During cloning, the tStuffer is replaced by the BC32 library. Constructs were used for *in vivo* barcoding experiments (groups BM-TBI and BM-CTx). Plasmid size: approximately 9640 bp.

2



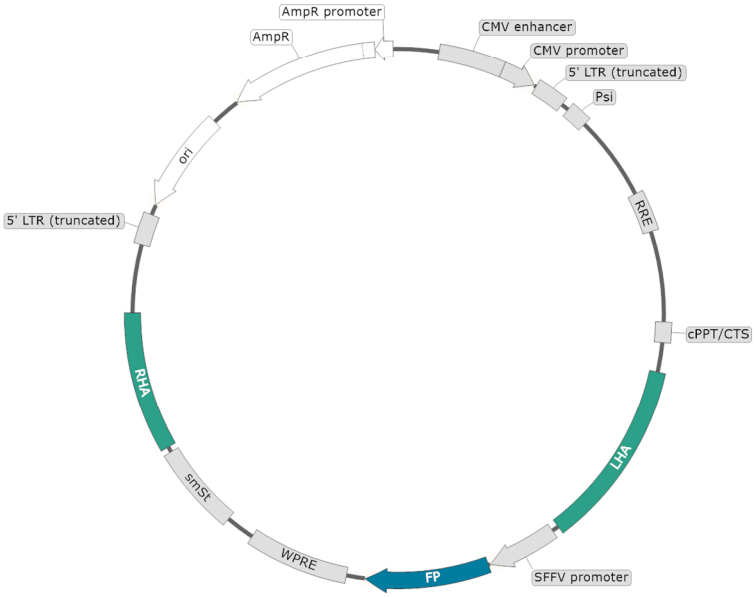
Appendix D-2: Plasmid map of lentiviral LS1-LS4 constructs. Plasmids contain ID sequences, SFFV promoter-driven fluorescent marker (GFP, eBFP, T-Sapphire, or Venus) and a tStuffer. During cloning, the tStuffer is replaced by the BC32 library. Constructs were used for *in vivo* barcoding experiments (groups BM-TBI^{SFFV} and BM-CTx). Plasmid size: approximately 9839 bp.

3



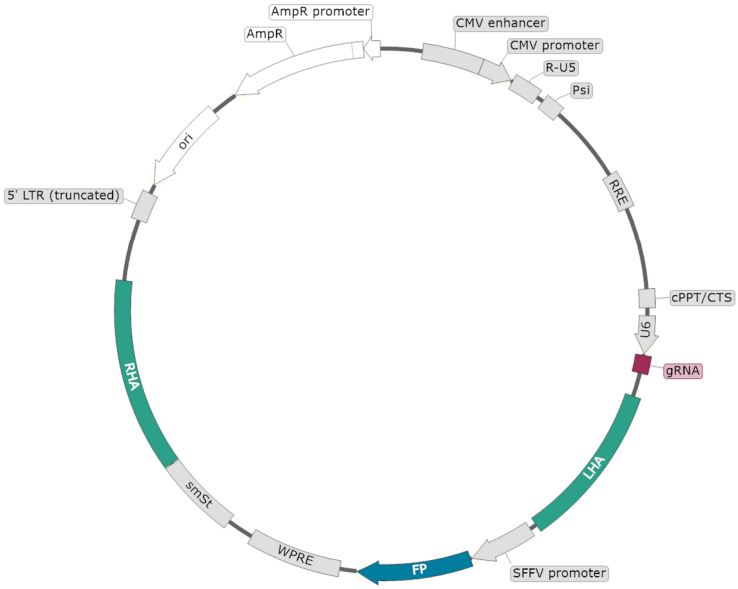
Appendix D-3: Plasmid map of lentiviral T1-T4 constructs. Plasmids contain U6 promoter-driven gRNA (ROSA1, ROSA2, HPRT1, or HPRT2) and EFS promoter driven SpCas9 sequence connected by a P2A to the fluorescent marker (mNeon). Constructs were used for CRISPR/Cas9 barcoding experiments as target vectors. Plasmid size: approximately 12480 bp.

4



Appendix D-4: Plasmid map of lentiviral D1-D4 constructs. Plasmids contain HAs (LHA and RHA), SFFV promoter-driven fluorescent marker (eBFP) and a small Stuffer (smSt). During cloning, the smSt is replaced by the BC32 library. Constructs were used for CRISPR/Cas9 barcoding experiments as donor vectors. Plasmid size: approximately 9599 bp.

5



Appendix D-5: Plasmid map of lentiviral DG1-DG4 constructs. Plasmids contain U6 promoter-driven gRNA (ROSA1, ROSA2, HPRT1, or HPRT2), HAs (LHA and RHA), SFFV promoter-driven fluorescent marker (eBFP) and a small Stuffer (smSt). During cloning, the smSt is replaced by the BC32 library. Constructs were used for the alternative approach of CRISPR/Cas9 barcoding experiments as donor vectors. Plasmid size: approximately 10311 bp.

11 Acknowledgment

First of all, I would like to acknowledge Prof. Dr. Müller, who offered me the opportunity to be part of his research group and to contribute to this interesting and challenging project. I always received great support and input for the project and my personal skills when it was needed, but I was also met with his support to establish a confident, self-organized, and independent way of working.

Special thanks go to PD Dr. Kerstin Cornils as my mentor and supervisor. Her ideas gave rise to the funding for my project and her enormous background knowledge on the topic initiated helpful discussions pushing me and thus leading to a focused scientific piece of work. You are always up for a laugh and refreshing anecdotes, while not forgetting about responsible mentoring. This gave me the strong impression of an appreciative working atmosphere with the appropriate space to find my own way. Ultimately, she encouraged me to apply for a following postdoctoral funding with a new project, which success I owe to her.

I would like to express my gratitude to Prof. Dr. Brune for agreeing to supervise my practical work and evaluate the written part of my thesis. I am aware that this additional responsibility is time-consuming, so I was impressed to receive not only the mandatory help but moreover the outstanding willingness for supporting supervision and scientific input.

Next, I would like to gratefully acknowledge the infrastructure and support of the Research Institute Children's Cancer Center Hamburg and the financial funding of the DFG and the Fördergemeinschaft Kinderkrebs-Zentrum Hamburg. Realizing this project would not have been possible without it.

For their technical support, I wish to express my thanks to Dr. Tim Aranyossy, Tanja Sonntag, Vanessa Thaden, Marianne Klokow, and Johannes Polke. Tim introduced me to the secrets of barcoding and made the beginning of my work much easier; Tanja was always available for technical questions, and by her assistance and extraordinary patience I finally managed to conduct the mice work on my own; Vanessa, you were always there, taking a lot of work off my back and confidently assisting even with laborious experiments with a heart-warming smile on your face (*don’t stop believing!*); Marianne accompanied me especially in the last part of my thesis, not only technically, but also personally; and by the help of Johannes I could many times rely on a trustworthy person to monitor the animals from the experiments.

Further, I sincerely acknowledge PD Dr. Ingmar Glauche and Dr. Lars Thielecke from the Technische Universität Dresden for their ongoing cooperation with our group, important contribution to realize the funding of my project and providing excellent support for bioinformatics analyses.

My very special thanks go to Dr. Marios Xydous and Dr. Pablo Iglesias Vázquez for proof-reading and editing my thesis. Both did an excellent job, and despite time constraints, they put extremely high effort in giving me constructive feedback, from which I learned a lot.

A pleasant, respectful, and friendly working atmosphere is an essential factor for doing good work and looking forward to it every single day – and that is exactly where Julia Strauss, Dr.

Antonia Wrzeszcz, Laura Rekowski, Marianne Klokow, Vanessa Thaden, Jaqueline Marschner, Katrin Küchler, Stefan Behrens and Bernd Hartz contributed to. I am particularly grateful of having such wonderful colleagues right next to me in the office. This of course also refers to the rest of my own working group and the (former) lab mates from the institute. You always had an open ear, we shared joyful lunchbreaks and I enjoyed being part of it.

Thanks are also due to lovely bachelors and masters students, which I was given the opportunity to accompany in the lab and co-supervise their projects and final theses. I also learned a lot from you about myself and how to support and supervise others. I wish we could have worked together for a longer period and I am grateful to have met you: Merrit Rothe, Bryan Barcelona and Jaqueline Marschner.

I also thank the staff of FACS Sorting Core Unit in UKE and HPI for their friendly collaboration and for spending uncountable hours of sorting my samples. Furthermore I would like to highly appreciate the work of Nicole Lüder from and Christiane Willmann from the FTH, who supported me in animal handling and care and provided the access to premises needed for the project.

Finally, I would like to mention and honor the strong personal support and endless love of my close friends, my family and my partner, without whom I would not be where I am right now. You might not be aware of how you participated in the success from this thesis, even (partly) from far away. I want to thank Liesa and Sarah, who were always there for me, sharing PhD feelings from different fields and together forgetting about everything with amazing dinners and the best company I can imagine after a day with a massive workload (*I do it for the glory!*). I did not only benefit from that, but also from your impressive R skills, Liesa, which helped me extremely and saved me a lot of frustrating hours. I am further very grateful to have met Kathi on my way of becoming a PhD. Thank you for sharing the important experiences we made together, and for never being tired of discussing and analyzing it. This feels great, but not as great as the moment we can both celebrate together (you know what I mean). My parents gave me the capability of always holding my head high and supported me to succeed with my personal and scientific career; always believing in me. Nils, the way you hold me in high esteem always gave me a strong background feeling. Lisa, thank you so much for endless phone calls, cheering me up in situations of doubt, and listening to all the scientific gibberish – knowing that you are there gives me all the confidence I need. And last but not least, my very special gratitude goes to Pablo, who always supported me unconditionally and opened up an entire new world for me, in every sense. *Quérote moito*.

12 Declaration on oath

I hereby declare upon oath that I have written the present dissertation independently and have not used further resources and aids than those stated. The submitted written version corresponds to the version on the electronic storage medium. I further affirm that this dissertation was not submitted in a previous doctoral procedure.

Hiermit erkläre ich an Eides statt, dass ich die vorliegende Dissertation selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Die eingereichte schriftliche Fassung entspricht der Fassung auf dem elektronischen Speichermedium. Ich versichere ferner, dass diese Dissertation nicht in einem früheren Promotionsverfahren eingereicht wurde.

Hamburg, 18.01.2021

Signature/ Unterschrift