Using barcode vectors for neutral genetic marking to study clonal dynamics of hematopoietic reconstitution

Dissertation with the aim of achieving a doctoral degree
at the Faculty of Mathematics, Informatics and Natural Sciences

Department of Biology
Universität Hamburg

Submitted by Tim Dominic Aranyossy

Hamburg 2017

Supervisor: Prof Dr Boris Fehse

Co-supervisor: Prof Dr Thomas Dobner

This thesis was successfully defended on Monday, 04.09.2017 in the presence of the following committee members:

Prof Dr Julia Kehr

Prof Dr Jonas Schmidt-Chanasit

Prof Dr Boris Fehse

Dr Jasmin Wellbrock

Dr Kerstin Cornils

1.Summary

1.1. English summary

Haematopoietic stem cell transplantation (SCT) is the only curative therapy option for a variety of malignant and non-malignant blood diseases. The contribution and importance of various cell populations to hematopoietic reconstitution have been studied extensively, but little is known about clonal dynamics within these populations. To study these dynamics, stable, inheritable marking of donor cells is required, which can be provided by integrating retroviral vectors. Unfortunately, stable genome insertion is associated with the risk of insertional mutagenesis potentially leading to malignant transformation of affected cells as documented in several gene therapy studies. Improved retroviral vector design has greatly reduced, but not removed the likelihood of insertional mutagenesis in the last decade. In parallel, development of genetic barcoding techniques has opened the possibility to analyse clonal composition and dynamics in greater detail than before. We reasoned that genetic barcoding of hematopoietic cells with state-of-the art retroviral vector system should facilitate high-resolution analysis of neutral hematopoietic reconstitution, unaffected by the marking procedure itself.

Within this thesis, I wanted to evaluate the influence of the vector type and their internal promoters on clonal dynamics of hematopoietic reconstitution after SCT. Based thereon, the vector construct best suited to study neutral reconstitution should be determined. To achieve this task, I took advantage of a genetic barcoding system with colour-coding capabilities. Alpha- and lentiviral vector constructs equipped with either a strong, intermediate or no promoter upstream of a fluorescence protein (FP) were barcoded and used to independently transduce lineage-negative cells of donor mice. I studied four groups with different competitive *in-vivo* setups by transplanting grafts, containing up to three different vector constructs, into lethally irradiated recipients. Genomic DNA was extracted from peripheral blood (PB) samples taken monthly, while selected time points were additionally analysed by flow cytometry (FC). Samples from PB, spleen, bone marrow as well as flow cytometrically sorted subsets of T cells, B cells and granulocytes were collected eight to twelve months after SCT. Barcode analysis via next-generation sequencing (NGS) created a dataset with temporal dynamics from successive PB samples, while the different populations sampled at the final analysis time point show the spatial distribution of clones.

FC analysis of chimaerism and FP expression confirmed stable long-term engraftment of the marked populations detectable over the whole observation period. The number of barcoded cells significantly contributing to haematopoiesis declined over time in most animals, although I observed big variability between individual mice.

Temporal and spatial analyses of clonal dynamics in the animals showed a diverse picture from monoclonal to polyclonal situations. In most samples, approximately 15 clones per construct contributed to more than 75% of the marked fraction. Nearly all clones measurably contributing to haematopoiesis at final analysis were already present six weeks after transplantation, and only their frequencies changed over time. Dominant or prominent clones representing a big fraction of the

marked haematopoiesis were detectable for all vector constructs tested without any bias towards any of the constructs used.

Finally, the obtained dataset was used to calculate the amount of cells contributing to haematopoiesis. These calculations indicate that around 350 cells actively supply the blood production six weeks after transplantation. Eight to twelve months after transplantation, this number decreases to around 260 cells.

In the transplantation setup investigated here, lenti- and alpharetroviral vector constructs equipped with different promoters showed comparable clonal dynamics and/or trends in all analyses. There might be some negative characteristics for the construct with the strong viral promoter, but the collected data is not sufficient for a final assessment. Thus, in principle all constructs appeared suitable for investigating undisturbed reconstitution of the haematopoietic system after transplantation. On the other hand, I also observed dominant, or at least prominent, clones marked by vector constructs without any promoters indicating that intrinsic cell features might promote clonal dominance.

In conclusion, this work demonstrates the feasibility to mark and track several distinct cell populations in parallel *in vivo* within single animals with a barcoding system. This allows competitive setups, where effects of various parameters (e.g. promoters), can be compared based on the colour-coded barcode backbones, while the individual barcodes provide further information about the clonal dynamics within a population.

1.2. German summary

Die Transplantation von blutbildenden Stammzellen (SZT) ist heutzutage die einzige kurative Therapie für eine Vielzahl von Erkrankungen des Blutes. Während Einfluss und Bedeutung der verschiedenen Zellpopulationen auf die hämatopoetische Rekonstitution intensiv erforscht worden sind, gibt es kaum Daten über die klonalen Dynamiken innerhalb dieser Populationen. Voraussetzung zur Analyse dieser Dynamiken ist eine stabile, vererbbare, Markierung der Spenderzellen, z.B. mittels integrierender retroviraler Vektoren. Durch die halbzufällige Integration dieser Vektoren ins Zellgenom besteht jedoch immer das Risiko von Insertionsmutagene die, wie leider in verschiedenen Gentherapie-Studien geschehen, zur bösartigen Transformation betroffener Zellen führen kann. Trotz großer Anstrengungen im letzten Jahrzehnt, und damit verbundenen Verbesserungen im Bereich der Vektorarchitektur, bleibt ein gewisses mutagenes Potential erhalten. Die Entwicklung von Techniken zum genetischen Barcoding in den letzten Jahren bietet die Möglichkeit, klonale Zusammensetzungen und Dynamiken im unerreichten Detail zu analysieren. Daher sollte die Kombination von genetischen Barcodes und einem modernen Vektorsystem die detaillierte, klonale, Analyse von neutraler, also durch die Markierungsprozedur unbeeinflusster, Rekonstitution des Blutsystems ermöglichen.

Innerhalb dieser Arbeit wollte ich den Einfluss von Vektorklasse sowie dem verwendetem internen Promoter auf die klonale Dynamik bei Rekonstitution des Blutsystems nach STZ untersuchen. Auf dieser Grundlage sollte das am besten für eine Analyse von neutraler Rekonstitution geeignete Vektorkonstrukt bestimmt werden. Durch die Möglichkeit der Farbcodierung ist unser neues, genetisches, Barcode-System bestens für die Beantwortung dieser Fragestellung geeignet. Alpha- und lentivirale Vektoren wurden mit einem genetischen Barcode sowie entweder einem starken, mittelstarken oder keinem internen Promoter vor einem Fluoreszenzprotein (FP) versehen. Mit diesen Vektoren wurden unabhängig voneinander *lineage* Marker negative Knochenmarkszellen von Spendermäusen transduziert. Durch verschiedene Kombinationen dieser Zellpopulationen wurden insgesamt vier kompetitive, *in vivo*, Gruppen transplantiert, in denen bis zu drei verschiedene Vektorkonstrukte miteinander konkurrierten. Den Tieren wurden monatlich Proben des peripheren Blutes (PB) entnommen, mittels Durchflusszytometrie (FC) analysiert und deren genomische DNA extrahiert. Acht bis 12 Monate nach SZT, wurden Proben von PB, Milz, Knochenmark und mittels FC sortierte Subpopulationen von T Zellen, B Zellen sowie Granulozyten aufgearbeitet. Durch die Analyse der Barcodes mittels Hochdurchsatzsequenzierung konnten die zeitlichen klonalen Dynamiken anhand der fortlaufenden PB Proben untersucht werden, während die verschiedenen Populationen zum Zeitpunkt der finalen Analyse Aufschluss über die räumliche Verteilung der Klone geben.

Die Untersuchung von Chimärismus und FP Expression mittels FC bestätigte ein stabiles Anwachsen der markierten Populationen über den kompletten Beobachtungszeitraum. In den meisten Tieren nahm die Anzahl von Barcode tragenden Klonen, mit bedeutsamen Einfluss auf die Blutrekonstitution, im Verlauf des Experimentes ab. Allerdings wurden teils deutliche Unterschiede zwischen einzelnen Tieren sichtbar.

Verschiedene Analysen der zeitlichen und räumlichen Verteilungen sowie Dynamiken der sequenzierten Barcodes zeigten variierende, mono- bis polyklonale, Situationen. In den meisten Proben rekonstituierten etwa 15 Klone je Vektorkonstrukt mehr als 75% der markierten Population. Fast alle Klone mit bedeutsamen Einfluss zum Zeitpunkt der finalen Analyse wurden bereits sechs Wochen nach Transplantation detektiert und lediglich deren Verteilung veränderte sich über die Zeit. Für alle getesteten Vektorkonstrukte wurden herausragende bzw. dominante Klone gefunden, die allein für einen Großteil der markierten Blutbildung verantwortlich waren. Dabei konnte keine Tendenz bezüglich eines bestimmten Konstruktes festgestellt werden.

Mit Hilfe der vorliegenden Daten konnte zudem die Anzahl an hämatopoetisch aktiven Zellen abgeschätzt werden. Demnach sind sechs Wochen nach SZT etwa 350 Zellen an der Blutproduktion beteiligt, während acht bis zwölf Monate nach Transplantation nur noch etwa 260 Zellen aktiv sind.

Innerhalb des hier untersuchten Transplantationsmodelles zeigten alle getesteten lenti- und alpharetroviralen Vektorkonstrukte, unabhängig vom internen Promoter, vergleichbare klonale Dynamiken und Tendenzen. Das untersuchte Konstrukt mit starkem Promoter zeigt möglicherweise Auffälligkeiten, die Datenlage ist aber nicht eindeutig genug um eine finale Aussage zu treffen. Daher scheinen prinzipiell alle Konstrukte für die Analyse von ungestörter hämatopoetischer Rekonstitution nach SZT geeignet zu sein. Innerhalb dieser Arbeit wurden jedoch selbst für Konstrukte ohne jegliche Promoteraktivität dominante, bzw. zumindest herausragende, Klone gefunden. Dieses könnte auf das Vorhandensein intrinsischer Besonderheiten in einzelnen Zellen hindeuten, welche die Entwicklung von klonaler Dominanz begünstigen.

Zusammengefasst demonstriert diese Arbeit das Potential unseres Barcode-Systems verschiedene Zellpopulationen zu markieren und deren Dynamiken innerhalb eines kompetitiven Modells parallel, also innerhalb eines einzelnen Tieres, zu verfolgen. Dadurch kann zum einen der Einfluss unterschiedliche Vektorkonstrukte, z.B. mit verschiedenen Promotoren, durch die Farbcodierung der Barcodes verglichen werden. Zum anderen können die klonalen Dynamiken innerhalb eines Vektorkonstrukts anhand der einzelnen Barcodes aufgeschlüsselt werden.

2. Acknowledgements

Numerous people contributed their part to this thesis over the last years. Therefore, I want (and need) to take some space acknowledging them:

First, I want to especially thank Prof. Dr. Boris Fehse for giving me the opportunity to work on this project and in your group. I am deeply grateful that you provided the necessary conditions, support and supervision over the course of this work!

Subsequently, I want to express my sincere gratitude to Dr. Kerstin Cornils for 4 years of excellent supervision. Thank you for your (always) open door, helpful tips, constructive criticism, answering all my (sometimes really stupid) questions, your support and guidance of my work through all ups and downs as well as the follow-up funding!

I would like to thank Prof. Dr. Thomas Dobner for your willingness to review and evaluate this work.

A very big "Thank you!" to Tanja Sonntag for revealing me the secrets of mouse work, teaching all protocols and techniques, giving guidance, answering questions, managing my (often enough last-minute) supply orders and helping me out on all the long mouse days!

Another big "Thank you!" to all current and former lab members of the Research Department Cell and Gene Therapy for providing the working atmosphere I was allowed to enjoy. Thanks for all your support, suggestions, helping hands when needed, but also all the laughs and funny situations.

I want to thank the members of the FACS Sorting Core Unit of the UKE Hamburg Eppendorf for countless hours of sorting my samples and Christian Schulze as well as Nina Kursawe at the Zentrum für Molekulare Neurobiologie Hamburg for trouble-free access to the irradiation machinery.

In addition, I am very grateful to Sabrina Noster, Ivonne Deutschmann, Silke Hauffe and Nicole Lüder working at the Forschungstierhaltung of the UKE Hamburg for quick and kindly managing all my mouse inquiries. A special "Thank you!" to Nicole Lüder for all your help, excellent caretaking and always keeping a watchful eye on my animals.

"Thank you!" to our cooperation partners at the Institute for Medical Informatics and Biometry (Prof. Dr. Ingo Roeder, Dr. Ingmar Glauche and especially Lars Thielecke) of the Technische Universität Dresden and their Deep Sequencing Group (Dr Andreas Dahl). Furthermore, I would like to acknowledge Prof. Dr. Axel Schambach and Dr. Julia Sürth (Hannover Medical School) for providing the initial alpharetroviral vector constructs and protocols.

My gratitude to the Deutsche Forschungsgemeinschaft for funding the project presented here as well as the Dr. Werner Jackstädt-Stiftung for funding a follow-up project, which was done while finishing the analyses and writing the thesis.

Finally yet importantly, there were many people outside the lab (or at least not directly related) with whom I had the pleasure meeting in the last years. Thank you! 4th floor Campus Forschung for

awesome TGIF's, the ASMB team at the ZMNH Hamburg for a highly advisable graduate program and of course all my other friends (I'm not even starting calling names here – you know who you are).

With the last words of this section, I would like to thank my family for all the support you gave me as well as your clean acceptance of all my decisions over the years.

3. List of abbreviations

The following list contains the abbreviations present in multiple parts of this work. Abbreviations used only in single passages are spelled out there.

BC32	32-wobble base barcode system, see page 29
BC	(genetic) barcode
BFP/eBFP2	(enhanced) blue fluorescent protein
BM	bone marrow
EFS	human elongation factor-1 alpha short promoter (Schambach et al., 2006)
FC	flow cytometry
FP	fluorescent protein
FACS	fluorescence-activated cell sorting
GFP/eGFP	(enhanced) green fluorescent protein
HSC	hematopoietic stem cell
LeGO (vector)	lentiviral gene ontology (vector) (Weber et al., 2008)
lin-	lineage-negative
LTR	long-terminal repeat
MG	mouse group, see page 54
NGS	next-generation sequencing
PB	peripheral blood
pd	promoter-deprived
SFFV	spleen focus-forming virus promoter
SIN	self-inactivating
TSapp	T-Sapphire, a blue-green fluorescent protein

4. Aim of this thesis

Gene therapy, the use of genes or genetic material to treat disease, holds the potential to provide a cure for many inherited diseases, currently only medicated symptomatically or with high risks (e.g. Duchenne muscular dystrophy, Severe Combined Immunodeficiency, Haemophilia and Sickle cell anaemia). Although the concept of gene therapy has been around for only 50 years (Friedmann and Roblin, 1972) with the first human trials being performed around 25 years ago (Blaese et al., 1995), remarkable success has been achieved. The first gene therapeutic drugs are already on the market[1,2], with many more in clinical trials and pipelines. However, there have been major setbacks disillusioning the first hype around the millennium, as treated patients developed leukaemias related to the vector construct used (Hacein-Bey-Abina et al., 2003a). In addition, there have been several reports about the emergence of dominant haematopoietic clones at various time points after gene-therapy treatment and stem cell transplantation (Cavazzana-Calvo et al., 2010; Ott et al., 2006; Schmidt et al., 2003). These dominant clones have often provided clinical benefit for the patients, due to high transgene expression levels. However, a monoclonal situation is always more susceptible to effects like promoter silencing or appearance of genomic instability (Stein et al., 2010), compared to a polyclonal system.

Thus, minimizing the risk of insertional mutagenesis, caused by introduction of unwanted genetic alterations by (semi)random integration of retroviral vectors, is still one of the major challenges for modern gene therapy. In the last years, great progress has been made to reduce this risk due to new vector systems, advantages in vector design, new techniques and better understanding of the mechanisms. In parallel, development of genetic barcoding techniques (Cornils et al., 2014; Gerrits et al., 2010; Schepers et al., 2008) has provided the tools necessary to reveal the clonal dynamics of haematopoiesis in far greater detail than before. However, even vectors only used for marking the haematopoietic (stem) cells before transplantation may influence cellular behaviour through insertional mutagenesis, hampering observations of presumably “neutral” reconstitution dynamics.

The main aspects investigated in this thesis tackled the following questions:

1) Is there a difference in clonal dynamics depending on the strength of the internal promoter or class of retroviral vectors used to mark haematopoiesis in a murine transplantation model?

2) What is the most neutral, retroviral vector for marking haematopoietic cells in a transplantation setting?

To answer this question, we utilize the novel BC32 system, developed during my master thesis (Thielecke et al., 2017). Using the “colour coding” capabilities of this system, we are able compare up to three different lenti- or alpharetroviral vector constructs, equipped with different promoters, within a competitive *in vivo* transplantation setup in parallel.

[1] http://blogs.nature.com/news/2012/11/gene-therapy-hits-european-market.html
[2] https://www.firstwordpharma.com/node/1386939?tsid=28&region_id=3
The dates of accession for each web source can be found in section 13.1, page 104

5. Introduction

5.1. The haematopoietic system

The blood-forming system (Greek: *haima*, blood + *poiesis*, to make[3]) belongs to the mesoderm germ layer and is primarily located in the bone marrow in adults. In humans, it represents ~5% of the total body weight and is able to produce $>10^{10}$ cells per day to replenish the turnover of the $>10^{12}$ total blood cells[4]. At the top of the haematopoietic hierarchy is the (long-term) haematopoietic stem cell (LT-HSC), which differentiates through different stages of multipotent progenitors into myeloid- or lymphoid-lineage committed precursors. These precursors further differentiate into a variety of specialized cells at the bottom of the hierarchy (Figure 1). In addition, (mostly lymphoid) progenitors may leave the bone marrow to mature at different sites of the body (e.g., T cells in the thymus, dendritic cells & macrophages inside tissues).

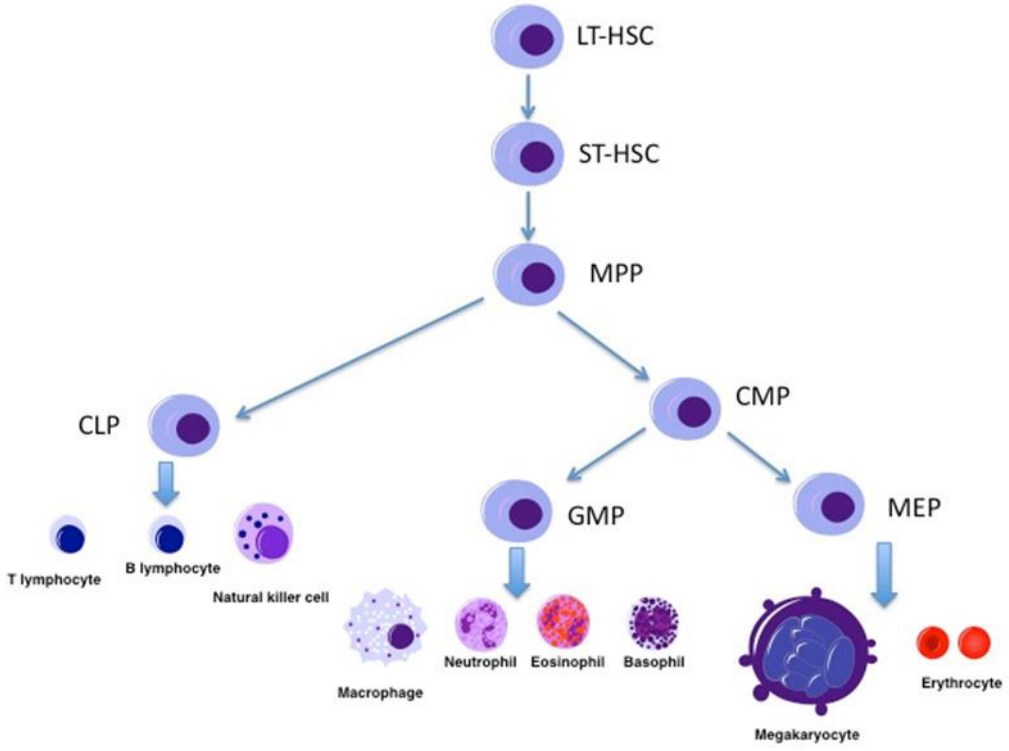


Figure 1 – Hierarchy of the haematopoietic system. Self-renewing haematopoietic stem cells (HSC) differentiate via different (multipotent) progenitor stages into myeloid- or lymphoid-lineage committed (mostly oligopotent) precursors and further in the fully differentiated cell types at the bottom of the hierarchy. HSC: haematopoietic stem cell, LT: long-term, ST: short-term, MPP: multipotent progenitor, CLP: common lymphoid progenitor, CMP: common myeloid progenitor, GMP: granulocyte-macrophage progenitor, MEP: megakaryocyte-erythroid progenitor. Figure taken from Masumi, (2013, page 66), CC BY 3.0 license, modified.

The first appearance of the term "stem cell" is documented in 1868 when the German biologist Ernst Haeckel used the German word "Stammzelle" to describe the unicellular ancestor of all evolved

[3] http://www.biology-online.org/dictionary/Hematopoiesis
[4] http://flexikon.doccheck.com/de/Erythrozyt and /Knochenmark

multicellular organisms (Haeckel, 1868). In the next decades, use of the term shifted towards the cells giving rise to germlines and the blood system (Maximow, 1909; reviewed in Ramalho-Santos and Willenbring, 2007). The importance of interactions between stem cells and the surrounding stroma cells for haematopoietic differentiation was proposed early (Maximow, 1906), but controversially discussed until proven in the 1920s. Development and testing of nuclear weapons in the 1940s boosted the interest in ionized-radiation caused damage and how it could be prevented (Jacobson et al., 1951). In the 1960s the existence of haematopoietic stem cells (HSCs) was empirically proven (Becker et al., 1963; Till and McCulloch, 1961; Till et al., 1964). Today, HSCs are characterised by their self-renewing capacity and multi-lineage engraftment potential (Kondo et al., 2003; Shizuru et al., 2005; Weissman, 2000). It has been demonstrated that a single HSC is sufficient to reconstitute haematopoiesis, and even some epithelial tissue in transplanted mice (Krause et al., 2001). Technical advances and further research led to the discovery and definition of multipotent progenitors (MPPs, multi-lineage engraftment but low or no self-renewing capacity) as well as myeloid (CMPs) and lymphoid (CLPs) committed progenitors (Akashi et al., 2000; Kondo et al., 1997; Morrison et al., 1997). In addition, some forms of inter-lineage differentiation of already committed progenitors have been described, e.g. immature lymphoid progenitors (called MLPs in man or LMPP in mouse) capable of differentiating into granulocytes or macrophages (both myeloid lineage) but showing lymphoid commitment (Adolfsson et al., 2005; Doulatov et al., 2010). In mice, the HSC population can be further divided into long-term, intermediate-term, and short-term HSCs (Doulatov et al., 2012; Seita and Weissman, 2010; Spangrude et al., 1988). These populations are defined by the duration of repopulation capacity. Short-term HSCs can sustain haematopoiesis for 4 – 6 weeks, intermediate-term clones persist for 6 – 8 months while long-term HSCs are able to reconstitute blood permanently (Benveniste et al., 2010). Each population described can be discriminated based on its specific marker expression profile, although the optimal marker combinations are constantly improved and/or redefined (Doulatov et al., 2010, 2012; Kondo et al., 2003; Seita and Weissman, 2010; Weissman and Shizuru, 2008). Recently, evolving methods of lineage tracing and genetic barcoding revealed new insights in naïve haematopoiesis as well as reconstitution of the haematopoietic system after stem cell transplantation, described in section 5.4.3, starting at page 31. Retroviral vectors are required to mark haematopoietic stem cells with an inheritable tag.

5.2. *Retroviridae*

5.2.1. Discovery of key features of retroviruses

One of the first descriptions of a retrovirus dates back to 1908 when the experimental transmission of leukaemia in chickens was shown (Ellermann and Bang, 1908). Shortly after, cell-free transmission of the - nowadays termed - Rous sarcoma virus (RSV) was described (Rous, 1910, 1911). This discovery awarded Peyton Rous a Nobel Prize in 1966. In 1936 John Bittner described vertical transmission (via germline) of - later named - Mouse mammary tumour virus (MMTV), a betaretrovirus (Bittner, 1936; Modrow et al., 2003). The provirus hypothesis, published by Howard Temin in 1964, explained how

the RNA virus genome could be used to generate new virus genomes over long periods of time. He proposed the existence of a double-stranded DNA state, the so called provirus, which is used as a template to generate new RNA for the virions (Temin, 1964). This hypothesis, in a similar fashion independently proposed by Svoboda et al., (1963), laid the basics for understanding retroviral replication. In 1970, the discovery of the reverse transcriptase, the enzyme facilitating reverse transcription of the single-stranded RNA genome into double-stranded DNA (Baltimore, 1970; Mizutani and Temin, 1970), solved the last part of the replication/transmission puzzle and was awarded with a Nobel Prize in 1975. Stehelin et al., (1976) described that the avian sarcoma virus genome contains "transforming" genes. These genes, today known as (proto-)oncogenes, e.g. the SRC gene in the Rous sarcoma virus (Suerth et al., 2014), partly explain the observed tumour-inducing capacity of some retroviruses (Modrow et al., 2003, page 390). The first description of a retrovirus causing cancer (T-cell lymphoma) in humans, the Human T-Lymphotropic Virus type 1 (HTLV-1), was published shortly after (Poiesz et al., 1980). The most relevant human retrovirus is the Human Immunodeficiency Virus 1 (HIV-1), which belongs to the lentiviral genus and causes the acquired immune deficiency syndrome (AIDS). HIV-1 was isolated and identified as AIDS causing agent in the early 1980s (Barre-Sinoussi et al., 1983; Popovic et al., 1984). This discovery was rewarded with a Nobel Prize in 2008 to Françoise Barre-Sinoussi and Luc Montagnier. Of course, there were several other important discoveries and studies not mentioned here, which ultimately have made the *Retroviridae* to one of the best-studied virus families today.

5.2.2. Characteristics and replication cycle

The family of *Retroviridae* contains seven genera, shown in Figure 2, classified based on their pathogenesis, morphological and genetic differences, as well as infection features (Modrow et al., 2003). Endogenous (= integrated provirus, which is inherited via germline and in some cases can be activated again to produce infectious particles) retroviruses from different genera have been found in mammalian species, e.g. humans, apes, felines, rodents, cows, rabbits and horses, but also birds and even some reptilian species like turtles (Hayward et al., 2013). Viruses are classified as "simple" or "complex" depending on their genome. Simple retroviruses contain only the structural gag (matrix-, capsid- and nucleocapsid-proteins), pol (reverse transcriptase, integrase and protease) and env (envelope) genes. Complex retroviruses, in contrast, contain additional, non-structural, accessory genes with varying functions, e.g. transcriptional activation of viral genes or transactivation of cellular genes (Modrow et al., 2003; Weiss, 1996).

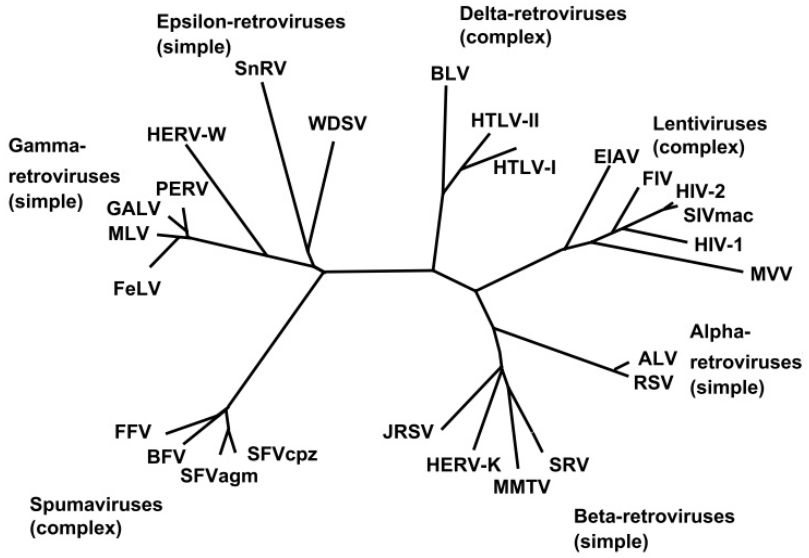


Figure 2 – Phylogeny of retroviruses. The known seven genera are shown with their members. Figure taken from Weiss, (2006), CC BY 2.0 license, modified.

Retroviral particles have a size of approximately 100 nm in diameter and contain two 5' capped and 3' polyadenylated, single-stranded (ss)RNA genomes, 7-12 kb in size (Figure 3). Genomes, as well as the proteins necessary for reverse transcription and integration, are located inside a capsid structure. Depending on the genus of the virus, different accessory proteins may be enclosed. The capsid is enveloped by a lipid bilayer containing (external and transmembrane) glycoproteins essential for virus entry into the host cell. Nomenclature of the glycoproteins normally depends on the molecular weight of the protein (e.g., the HIV-1 gp120 protein has a mass of 120 kDa). Sticking with the HIV-1 example, the external glycoprotein (gp120) is necessary for binding the cellular surface receptor (CD4) and/or co-receptors (CCR5 or CXCR4) resulting in a conformational change and exposure of a hydrophobic sequence of the associated gp41 protein. This triggers fusion of virus and cell membrane, enabling the virion to enter the host cell (Modrow et al., 2003).

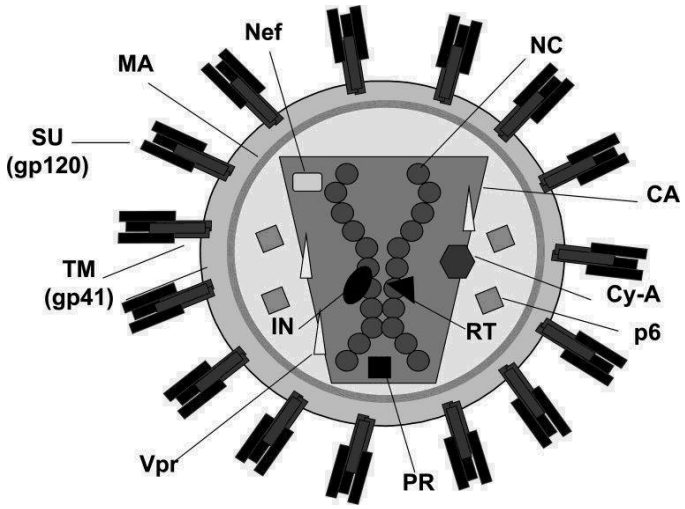


Figure 3 – Structure of HIV-1 (genus: *Lentivirus*), a complex retrovirus. Two single-stranded, 5' capped and 3' polyadenylated RNA genomes (approx. 10kb in size), the enzymes necessary for reverse transcription and integration as

well as different accessory proteins are packaged into a capsid. The viral envelope consists of a lipid bilayer and the glycoproteins, which are essential for virus entry. NC: nucleocapsid protein, CA: capsid protein, MA: matrix protein, SU: surface envelope protein, TM: transmembrane envelope protein, IN: integrase, RT: reverse transcriptase, PR: protease, Nef, p6, Vpr: accessory proteins, Cy-A: Cyclophilin A. Figure taken from Dufait et al., (2013, page 320), CC BY 3.0 license, modified.

A typical retroviral life cycle is depicted in Figure 4. After entering the cytoplasm, the capsid is uncoated. The exact time and localisation of uncoating are still matter of debate (Arhel, 2010). Afterwards, reverse transcriptase, the unique enzyme that defines the retrovirus family, converts the ssRNA genome into dsDNA in a quite complex process (summarised e.g., in Basu et al., 2008). A pre-integration complex (PIC) is formed between dsDNA, integrase and other host (as well as viral) proteins (summarised in Craigie and Bushman, 2014). The PIC import into the nucleus is a critical step for the viral replication but varies between genera of *Retroviruses*. Some genera (e.g., *Gammaretroviruses*) require nuclear envelope breakdown during mitosis (Roe et al., 1993), whereas others, like *Alpharetroviruses* and *Lentiviruses* are also able to infect non-dividing cells (Katz et al., 2002; Yamashita and Emerman, 2006). Although not requiring mitoses for infection, some viruses require cell cycle progression for different steps in their life cycle (Humphries and Temin, 1974; Humphries et al., 1981). There has been some discussion that late G1 or S-phase promotes efficient reverse transcription in different retroviruses, such as avian sarcoma virus and Moloney murine leukaemia virus (Humphries et al., 1981; Katz et al., 2002; Piéroni et al., 1999). Looking at the big picture, gammaretroviruses are most dependent on cell division (less than 1% transduction in non-dividing cells), while alpharetroviruses are able to transduce 3-30% non-dividing cells and lentivirus infection seems almost independent of cell division (Hatziioannou and Goff, 2001; Katz et al., 2002; Lewis et al., 1992; summarised in Yamashita and Emerman, 2006). Irrespective of the nuclear entry mechanism, the final step of the early retroviral life cycle is the integration into the host genome. This process and the consequential differences between retroviral genera regarding the integration patterns will be discussed in more detail in section 5.3.1, starting on page 19. The second part of the retroviral life cycle consists of transcription of full-length RNA genomes, as well as (spliced) variants, the latter ones used for translation of the viral proteins. Finally, new retroviral particles are assembled and released to infect other cells. Due to the lack of replication potential in retroviral vectors, this late phase of the life cycle is not further discussed here and interested readers may be relegated to the variety of reviews available, e.g. Bush and Vogt, (2014); Freed, (2015).

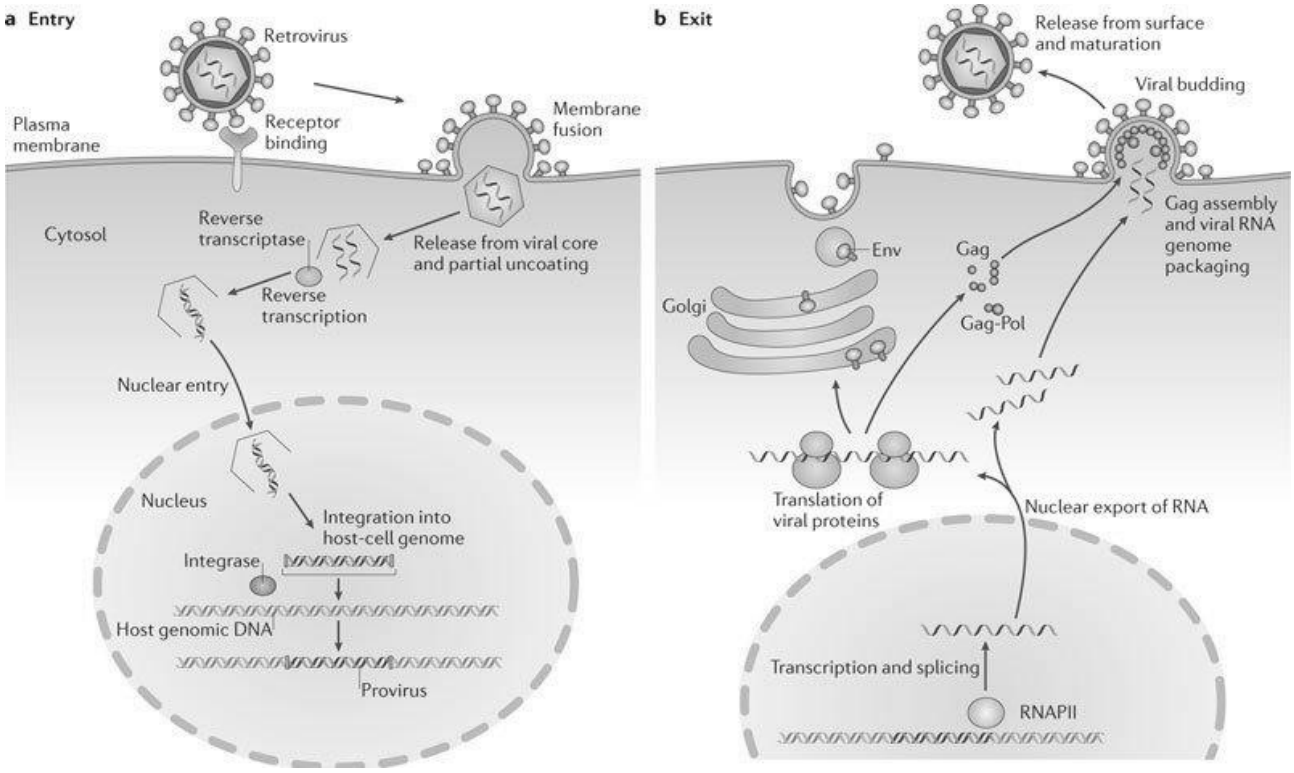


Figure 4 – A typical retroviral life cycle. Receptor binding of the host cell is facilitated by the glycoproteins on the viral surface. After binding the membranes fuse, releasing the viral capsid into the cytoplasm where reverse transcription of the ssRNA genome into dsDNA takes place and a pre-integration complex (PIC) is formed. Import of the PIC into the nucleus can be either passive during cell division (e.g., *gammaretroviruses*) or active, allowing certain retroviruses (e.g., *lentiviruses*) to infect non-dividing cells, too. The PIC then mediates semi-random integration of the provirus in the host cell genome. During the late stages of viral replication, viral genes are transcribed and spliced, while full-length virus genomes are exported for assembly of new viral particles. Figure taken from Stoye, (2012) with permission, modified.

5.3. Gene therapy

Given the unique ability to stably integrate DNA into the host genome, the potential in using retrovirus-based vectors as gene transporters for the treatment of inherited diseases (= gene therapy) has been investigated for long. Friedmann and Roblin, (1972) and Tatum (reprint 2009, original article 1966) were among the first to propose the concept of gene therapy but also raising ethical concerns and mentioning potential risks, resulting from the lack of scientific knowledge. About 10 years later, the first retroviral vector systems for use in humans were published (Mann et al., 1983; Sorge et al., 1984). The first therapeutic clinical trial took place in 1990. Circulating T cells from two patients with severe combined immunodeficiency (ADA-SCID) were *ex vivo* transduced with adenosine deaminase (ADA) carrying retroviral vectors and reinfused. Clinical benefit for both patients was reported, although they continued ADA replacement therapy (Blaese et al., 1995). App. 10 years later, clinical efficacy and significant benefit were reported 10 months after treating haematopoietic stem cells of patients with the X-linked form of SCID (SCID-X1), a (lethal) disease in which T and NK cell differentiation is blocked due to deficiency in the common cytokine receptor gamma chain (Cavazzana-Calvo et al., 2000). Unfortunately, about 3 years after gene therapy, the first severe adverse event, namely T-cell leukaemia effect was diagnosed, shortly followed by leukaemia development in a second patient

(Hacein-Bey-Abina et al., 2003a, 2003b). In the end, four of nine patients developed leukaemia 3 – 6 years after the gene therapy. Two cases could be tracked back to retroviral vector insertion near the LIM domain–only 2 (*LMO2*) proto-oncogene in combination with additional genetic alterations not related to the vector (Hacein-Bey-Abina et al., 2008). Importantly, all but one patient were successfully treated from their leukaemia and remain in complete remission without losing clinical benefit from the gene therapy since. Only one patient who received an allogeneic stem cell transplant as a rescue treatment for his leukaemia unfortunately died from severe side effects of the transplantation. Comparable results were reported from a gene therapy trial against Wiskott-Aldrich Syndrome (WAS), another x-linked immunodeficiency. Initially, clinical benefit and engraftment of cells expressing the WAS protein introduced into stem cells via gammaretroviral vector between 2006 and 2009 was observed for nine out of ten patients. Later on, seven of these patients developed acute leukaemias due to vector integrations near the known oncogenes *LMO2*, *MDS1* and *MN1* (Braun et al., 2014).

In a third gene therapy trial, stem cells from patients with chronic granulomatous disease were treated with a gammaretroviral vector introducing a non-mutated version of the $gp91^{phox}$ gene (Ott et al., 2006). Patients with this disease exhibit a defect in the NADPH oxidase enzyme and lack superoxide radicals to combat pathogens. Contrary to the two diseases discussed beforehand, the $gp91^{phox}$ gene was not known to provide a survival or growth advantage to transduced cells. Both treated patients initially experienced clinical benefit and could discontinue antimicrobial prophylaxis. However, one patient developed a myelodysplastic syndrome with monosomy of chromosome 7 and died 27 months after gene therapy. The second patient underwent allogenic stem cell transplantation due to detected dysplasia (Ott et al., 2006; Stein et al., 2010). Again, retroviral insertion into the *MDS1-EVI1* gene locus played an important role in malignant transformation.

Since these "early" studies, a lot of research has been dedicated to improve trial designs, protocols, vector architecture and safety. In the last years, gene therapy has had some remarkable success. Clinical benefit has been achieved for patients with WAS (lentiviral vectors, Aiuti et al., 2013), Adrenoleukodystrophy, SCID-X1 and, amongst others, retinal diseases (reviewed in Naldini, 2015). Furthermore, Strimvelis, a gene therapy against ADA-SCID, got approval by the European Medicines Agency in May 2016[5].

Combined, there are currently over 2300 gene therapy clinical trials approved or initiated for various indications, 89 of them in the "late" phases III/IV (Figure 5). Despite all success, insertional mutagenesis is still a concern when using retroviral vectors.

[5] http://www.ema.europa.eu/ema/index.jsp?curl=pages/medicines/human/medicines/003854/human_med_001985.jsp&mid=WC0b01ac058001d124

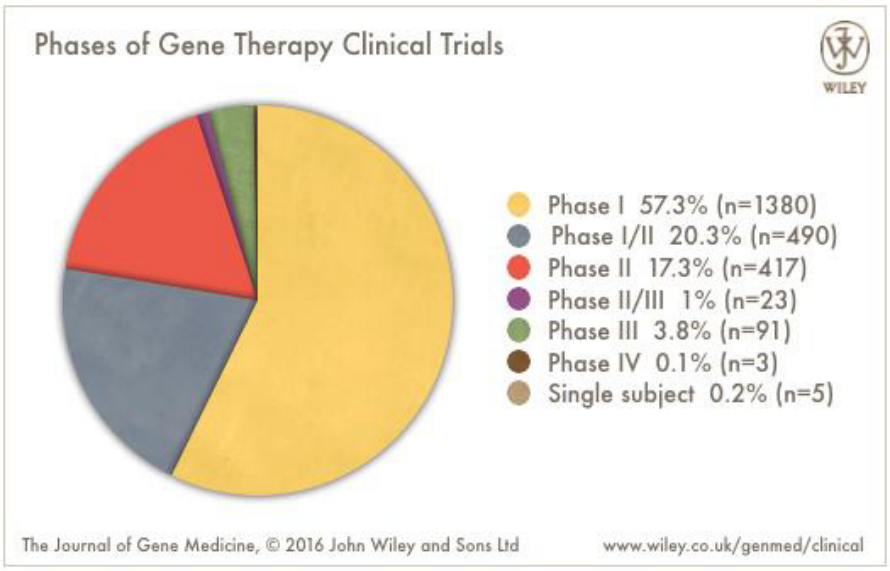


Figure 5 – Phases of gene therapy clinical trials. Clinical phase and number of registered trials (August 2016) are shown. Source: http://www.wiley.com/legacy/wileychi/genmed/clinical/, with permission

5.3.1. Insertional mutagenesis and insertion patterns

Insertional mutagenesis is defined as "*A mutation caused by the addition of DNA to effectively disrupt or alter the function of a given gene*" (Carlson and Largaespada, 2005). In the context of retroviral gene therapy, insertional mutagenesis is caused by the (semi-)random integration of a provirus into the genome. In general, several mechanisms altering gene function are possible: activating or inactivating cellular promoters, changing transcript stability and disruption or extension of an open reading frame (Baum, 2011). Depending on the architecture of the vector used, e.g. self-inactivation design or long-terminal repeat-driven transgene, genera of the vector, strength of internal promoter, splice signals and poly-A signal, some of those effects are more/less likely to occur. The most problematic mechanism in the early stages of gene therapy was the upregulation of proto-oncogenes by promoter activation (Suerth et al., 2014). Several cases of leukaemic development after transduction with clinically relevant retroviral vectors had been reported in mouse models (Li et al., 2002) and, as previously mentioned, gene therapy patients (Hacein-Bey-Abina et al., 2003a, 2008). Based on the animal and clinical trial data at that time, the frequency for such oncogenic events was estimated to be quite low (Kohn et al., 2003). Later on, it was calculated that it is all about the numbers (Baum et al., 2004). With over 200 known problematic proto-oncogenes and a problematic distance of 10 kb around them, an insertion event is likely to occur with a frequency of 10^{-3} to 10^{-2}. Given 1 million kilobases accessible for vector integration, the frequency of a risky insertion was pinned down to about 1 in 100.000 insertion events (Baum et al., 2003, 2004). However, not every single of those events leads to a malignancy. First, only a small fraction of haematopoietic cells really engrafts long-term. Second, oncogene-related signals may simply lead to apoptosis of the affected cell and/or further genetic hits or external stimuli may be necessary to induce oncogenesis (Hahn and Weinberg, 2002). In addition, the immune system may interfere. However, patients may be immunocompromised in T cells for years following stem cell transplantation (Williams and Gress, 2008).

In recent years, a lot of progress in understanding and minimising the effects of insertional mutagenesis has been made. I will focus in this section on the different insertion profiles of the retroviral vectors and discuss the vector architecture aspects in the next sections (5.3.2.1 and 5.3.2.2). Crucial to understand insertional mutagenesis is the ability to monitor and trace insertion sites of retroviral vectors. Laborious studies could show, that gammaretroviruses do not integrate randomly into the genome, but rather prefer promoter regions (Mooslehner et al., 1990; Rohdewohld et al., 1987). With development of clonal tracking methods (see section 5.4), better sequencing techniques and the availability of the human genome sequence (International Human Genome Sequencing Consortium, 2001; Venter et al., 2001) large scale mapping of retroviral integration sites became feasible. The first large scale analysis of 524 insertions sites from a HIV-based vector in a human T cell line revealed an integration bias towards active genes and regional hotspots instead of the previously expected random integration (Schröder et al., 2002). Further studies of different viral vectors and increasing amounts of mapped integration sites in different systems (e.g. Biffi et al., 2011; Bushman et al., 2005; Derse et al., 2007; Mitchell et al., 2004; Moiani et al., 2014; Narezkina et al., 2004; Wang et al., 2009; Wu et al., 2003) revealed the unique integration patterns of different retroviral vector genera. Concentrating on gammaretroviral vectors (the most commonly used in clinical trials), lentiviral vectors (rising utilisation in clinical trials) and alpharetroviral vectors (a new vector system) the unique integration patterns of these vectors are summarised in Figure 6.

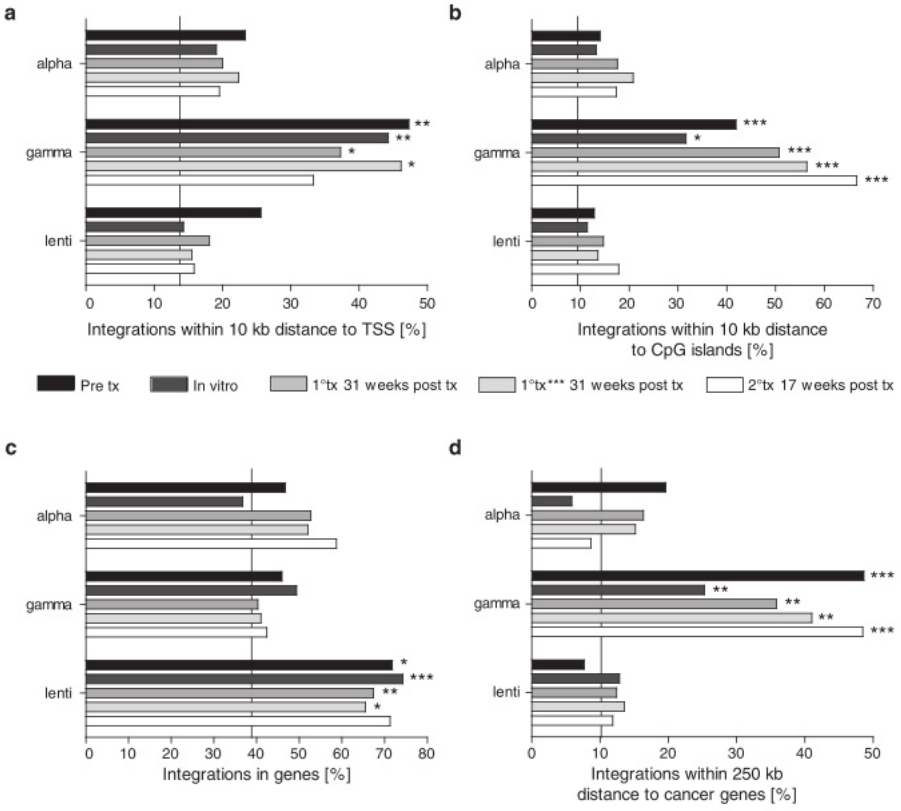


Figure 6 – Target site preferences of different retroviral vectors for specific genomic features recovered *in vitro* and from an *in vivo* transplantation model at different time points. Gammaretroviral vectors show a clear preference for integration near transcription start sites (TSS) (a), CpG islands (b) and into the proximity of cancer genes (d) compared to random (vertical line). Lentiviral vectors, in contrast favour integration into active genes (c). The integration pattern of

alpharetroviral vectors shows no specific target site preferences in comparison to the lenti- or gammaretroviral vectors. Figure taken from Suerth et al., (2012), CC BY-NC-ND 4.0 license .

Gammaretroviral vectors (primarily the Moloney murine leukaemia virus (MLV)-derived vectors) show site preferences towards integration within 10 kb of transcription start sites (TSS) and CpG islands (Mitchell et al., 2004; Wu et al., 2003). However, later studies showed that this may just be a consequence for the general preference for the MLV pre-integration complex integrating in genomic regions transcriptionally regulated by RNA polymerase II and/or associated with histone modifications indicating active transcription (Cattoglio et al., 2010; Cavazza et al., 2013; Felice et al., 2009; Lewinski et al., 2006). The underlying mechanism for this phenomenon was discovered to be the interaction between the gammaretroviral integrase and bromodomain and extraterminal domain (BET) proteins tethering the pre-integration complex to acetylated Histone3 and Histone4 tails which are found at TSS (Sharma et al., 2013).

In contrast, HIV-1 derived lentiviral vectors show a preference for integration into transcribed gene regions (Cavazza et al., 2013; Mitchell et al., 2004; Schröder et al., 2002). The main tethering factor for lentiviral pre-integration complexes was found to be lens epithelium-derived growth factor (LEDGF/p75) (Cherepanov et al., 2003; Llano et al., 2004; Maertens et al., 2003). Although LEDGF/p75 plays an important role, it is not strictly essential for HIV integration (Engelman and Cherepanov, 2008) while cofactors seem to play important roles as well (Lewinski et al., 2006; Matreyek and Engelman, 2011; Ocwieja et al., 2011).

The integration profile of alpharetroviral vectors only shows weak preferences towards integration into active genes and associated genomic features, thus displaying a more neutral integration pattern compared to the other vector classes (Mitchell et al., 2004; Suerth et al., 2012). That fact, the availability of a self-inactivating vector system (Suerth et al., 2010), as well as the ability to transduce non-dividing cells, although some cycle progression may be required (Humphries and Temin, 1974; Humphries et al., 1981), makes them an interesting alternative to the commonly used gamma- and lentiviral vectors.

5.3.2. Vectors for gene therapy

Nowadays, there are several methods to insert genes/DNA into cells. In clinical trials for gene therapy, adenoviral and retroviral vectors are the dominant choice of delivery followed by naked or plasmid DNA (see Figure 7). Due to the high turnover rate in the hematopoietic compartment, the use of integrating vectors is essential to ensure permanent marking. This work will focus on retroviral, especially lentiviral and alpharetroviral, vectors.

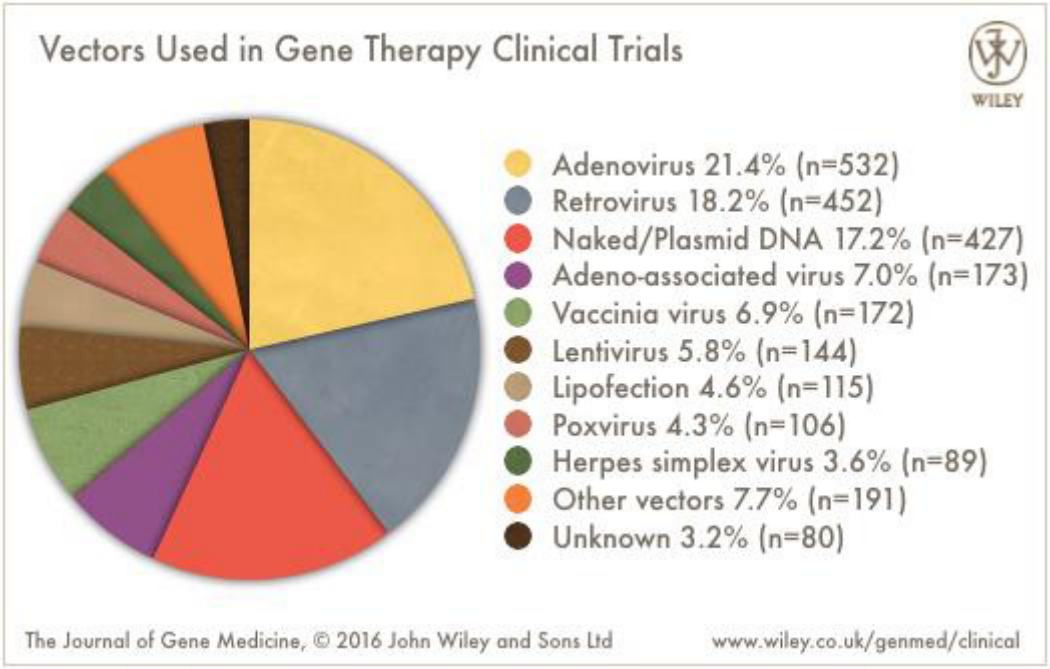


Figure 7 – Vectors used in gene therapy clinical trials (August 2016). Cumulative numbers of trials for the different vectors since 1989. Source: http://www.wiley.com/legacy/wileychi/genmed/clinical/, with permission

5.3.2.1. Lentiviral vectors

Vector systems based on lentiviruses were developed in the 1990s (Naldini et al., 1996) when basically all, haematopoiesis related, gene therapy trials were done with gammaretroviral (MLV based) vectors. As already mentioned above, lentiviruses are able to transduce non-dividing cells (Lewis and Emerman, 1994) and show a potentially more favourable integration pattern compared to gammaretroviral vectors regarding genotoxicity. Today, retroviral vector systems are divided into generations, improving from the initial first generation to the nowadays-used third generation. The classification of the generations is based on the packaging plasmids used for production of viral particles (reviewed in Escors and Breckpot, 2010). Lentiviral first-generation systems provided the structural gag and pol sequences, tat and rev as well as four accessory genes (vif, vpr, vpu and nef). In the second generation the accessory genes could be removed with no negative effect (Zufferey et al., 1997). Shortly thereafter third-generation systems were published using a so called split packaging design (Dull et al., 1998) requiring 4 plasmids for vector production – a vector plasmid with a packaging signal, a gag/pol plasmid, a plasmid containing the rev gene as well as an envelope plasmid. Consequently, coding sequences for the structural viral genes, required for vector assembly are only present during vector production and not in the packaged particles. This segregation of the viral genes strongly reduces the likelihood of recombination events that might result in the generation of replication competent viral particles, a major safety concern. Safety was further improved by creating self-inactivating (SIN) vectors (Dull et al., 1998; Miyoshi et al., 1998). In these vectors, parts of the U3 region, containing TATA box as well as transcription factor binding sites, of the 3' long-terminal repeat (LTR) were removed from the vector plasmid. During reverse transcription, the SIN deletion is

passed to the 5'LTR (readers interested into the detailed process are relegated to e.g., Hu and Hughes, (2012)), resulting in the transcriptional inactivation of the proviral LTRs and thus reducing the mobilisation- and genotoxic potential of these vectors (Modlich et al., 2009; Montini et al., 2009). The transcriptional inactivation of the LTR sequences requires use of an internal promoter for transgene expression, enabling the use of cell or tissue specific promoters and giving more control over expression levels and tissue specificity (summarized in Escors and Breckpot, 2010). Inducible promoters, e.g. the Tet-based system (Kafri et al., 2000; Reiser et al., 2000), allow for tight control of transgene expression, creating useful research tools. Aditionally, this allows the used of genes that kill the host cell if activated (suicide genes) and maybe even surrounding cells via bystander effect as safety mechanism or strategy of anti-tumour therapy (reviewed in Rama et al., 2014).

The tropism (= specificity of a virus for a particular host tissue[6]) of lentiviral vectors is determined by the envelope protein used during vector production. Changing the tropism, a process called pseudotyping, by using envelope proteins from different other viruses allows customisation of the vectors to the needs of the application (e.g., targeting the central nervous system using rabies glycoproteins (Mazarakis et al., 2001)). Today, one of the most widely used env proteins for lentiviral vectors is the one from vesicular stomatitis virus (VSV-G) based on its broad host range, good virus titres and high particle stability (Bartz and Vodicka, 1997; Burns et al., 1993). Readers interested in this topic are relegated to additional literature, e.g. (Breckpot et al., 2007; Cronin et al., 2005).

The LeGO vectors used in this work are a 3rd-generation lentiviral vector system (Weber et al., 2008). As described in the Methods section (page 38), a genetic barcode has been added after the woodchuck hepatitis virus post-transcriptional regulatory element (wPRE) resulting in the general vector structure depicted in Figure 8.

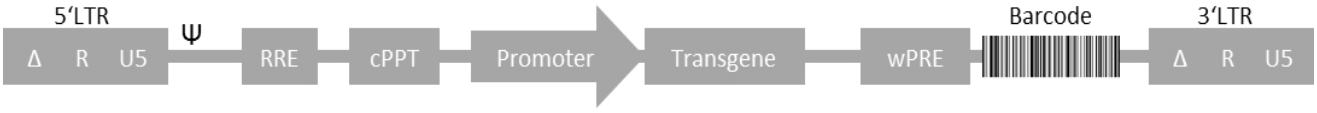


Figure 8 – Schematic overview of the LeGO constructs used in this work (as integrated provirus). The functions of the different features shown are mentioned in the text. Promoter and transgene varied depending on the specific construct. LTR: self-inactivating long-terminal repeat, Δ: partially deleted U3 (SIN deletion), Ψ: packaging signal, RRE: rev-responsive element, cPPT: central polypurine tract, wPRE: woodchuck hepatitis virus post-transcriptional regulatory element. Not to scale.

The integrated provirus consists of the flanking self-inactivating long-terminal repeats (LTRs), necessary for integration and reverse transcription. Ψ (psi) is the **p**ackaging **si**gnal, required for packaging the viral genome into the viral particles (Lever et al., 1989). The RRE (rev-responsive element) facilitates nuclear export of the full-length vector genome during production in a rev-dependent way (summarised in Pollard and Malim, 1998). The central polypurine tract (cPPT) enhances transduction of non-dividing cells (VandenDriessche et al., 2002) and can, alone or in combination with the PRE (in the case of LeGO vectors a wPRE from Woodchuck hepatitis virus is

[6] http://medical-dictionary.thefreedictionary.com/viral+tropism

used), enhance transgene expression (Barry et al., 2001; Zufferey et al., 1999). Details about the barcode can be found in section 5.4.2, starting page 29.

5.3.2.2. Alpharetroviral vectors

Alpharetroviruses were among the first retroviruses to be discovered (Ellermann and Bang, 1908; Rous, 1911) but, in contrast to gamma- and lentiviruses, a clinically applicable alpharetroviral vector system was developed late (Suerth et al., 2010). Of note, there was an alpharetroviral system described earlier, but it was replication-competent in avian cells, bearing a potential risk for use in humans (Hughes, 2004). As discussed earlier, alpharetroviruses show a more "neutral" integration pattern, compared to gammaretroviral and lentiviral vectors, making them a potential safer alternative (Suerth et al., 2012). As with the other vectors, a self-inactivating mutation has been introduced into the U3 region of the LTR, eradicating the transcriptional capacity. A split-packaging system with three different plasmids (vector plasmid, gag/pol and a pseudotypeable env) is available (Suerth et al., 2010), allowing production of high titres in mammalian cells. The general structure of the alpharetroviral provirus used in this work is shown in Figure 9.

Figure 9 – Schematic overview of the alpharetroviral constructs used in this work (integrated provirus). The functions of the different features shown are mentioned in the text. Promoter and transgene varied depending on the specific construct. LTR: self-inactivating long-terminal repeat, Δ: partially deleted U3 (SIN deletion), Ψ: packaging signal, wPRE: woodchuck hepatitis virus post-transcriptional regulatory element, DRE: direct repeat element. Not to scale.

The common elements for retroviral vectors have been described earlier (page 23). In addition to these features, alpharetroviral vectors harbour a direct repeat element (DRE) near the 3'LTR that promotes cytoplasmic accumulation of full length RNA as well as enhanced packaging (Ogert et al., 1996; Sorge et al., 1983). The barcode has been inserted between the wPRE and DRE element.

5.4. Genetic barcoding

A barcode is defined as *"a small rectangular pattern of thick and thin black lines printed on a product, or on its container, so that the details of the product can be read by and recorded on a computer system"*[7]. Barcodes in general have become an essential part of our modern world. Customers directly encounter barcodes when shopping products, validating online bought tickets, scanning Quick Response Codes with their smartphones or receiving parcels. In addition, modern logistic systems would be impossible without these unimposing labels.

[7] http://dictionary.cambridge.org/dictionary/english/bar-code

A completely different type of barcodes has been used in biological sciences. Phylogenetic barcoding allows discrimination/identification of species relationships based on common genetic sequences, (mainly) ribosomal or mitochondrial. For example, there have been phylogenetic barcoding studies using 16S or 23S rRNA, Cytochrome oxidase I/II or Cytochrome-b (reviewed in Patwardhan et al., 2014).

Sequence-based barcodes are also used in high throughput PCR applications, e.g., next-generation sequencing (NGS) for, so-called, multiplexing. In this process, amplicons from different samples are amplified by specific primers, adding a short identifier tag (= barcode in this case), which allows mixing of different samples (e.g., on one flow cell) and subsequent bioinformatic assignment back to the original sample.

Genetic barcoding, used for analysis of clonal dynamics (e.g., of the haematopoietic system) utilises the unique ability of retroviral vectors to integrate semi-randomly into the host cell. Originally, the lineage relationships and hierarchy of the haematopoietic system were investigated using the unique integration site of an individual vector, identified by southern blot analysis (Capel et al., 1990; Dick et al., 1985; Jordan and Lemischka, 1990). Technical advantages like ligation-mediated (LM-) PCR (Kustikova et al., 2005; Modlich et al., 2005; Mueller and Wold, 1990), linear-amplification-mediated (LAM-) PCR (Schmidt et al., 2002, 2007) and improved sequencing protocols streamlined the process and allowed for higher resolution by increasing numbers of clones that could be tracked. However, there are relevant limitations of LM- and LAM-PCR mainly attributed to the combinations of used enzymes and the distance of their recognition sequence to the integrated provirus. At the beginning of the LAM-PCR protocol, genomic DNA is digested using restriction enzymes with 4-bp recognition sites. Genomic regions with rare distribution of this motifs as well as very short produced fragments will be missed during subsequent analysis. Several groups showed/calculated that, even in laborious experiments with multiple combinations of enzymes (therefore also requiring a lot of sample material), only 85-90% of the genome would be accessible for integration site analysis (Bystrykh et al., 2012; Gabriel et al., 2009). In addition, integrations into repetitive genomic regions may produce ambiguous integration sites.

One way to overcome these limitations are genetic barcodes. These unique sequence tags are inserted into a defined position within the retroviral vector plasmid. Barcodes are created using oligonucleotides with several variable positions (“N”), randomly determined during the production process. The first systems were already developed in the 1990s (Golden et al., 1995), but technical limitations narrowed possible applications at that time. In 2008, the group of Ton Schumacher published their design of an artificial barcoding system (Schepers et al., 2008). They created a library of 4743 individual plasmids, containing a 98 bp semi-random stretch of DNA, which was then detected via microarray. Another approach, combining genomic barcoding and sequencing appeared only two years later (Gerrits et al., 2010). Their structure consisted of 12 semi-variable positions alternating with fixed triplets, which can theoretically give rise to over 4 million different barcodes. The emergence of NGS techniques around 2005 further expanded the possibilities of genetic barcoding. Lu et al., (2011) were able to track proliferation and development of hundreds of cells after

transplantation with, at that time point, unachieved precision by using a 27-bp random stretch with a 6-bp library identifier sequence for barcode identification. In contrast to all previous methods, deep sequencing also enabled high throughput and, given enough sequencing depth, detection of rare events. Other published barcode systems mostly were based on one of the described designs. Findings of those studies are further discussed in section 5.4.3, starting page 31.

Development of the barcoding system (Figure 10) preceding the one used in this work started in 2010 (Cornils et al., 2014). The system utilises a comparable barcode structure as described by Gerrits et al., (2010), but some important parameters have been changed. First, the number of variable positions has been extended to 16, in theory allowing up to 4^{16} = ~4.3x10^9 different barcodes. Second, by varying the positions of the fixed triplets (the barcode backbone) another level of complexity could be generated, which permits to combine different barcode backbones with different vectors. This new feature facilitated the combination of RGB marking (Weber et al., 2008) with genetic barcoding (Cornils et al., 2014). Furthermore, the structured backbone sequence prevents by-chance generation of recognition sequences for the restriction enzymes used in LM-PCR. Third, by optimising cloning and transformation protocols plasmid libraries containing ~5 million different barcodes can be generated – multiple times the number published for the other systems at that time. With this barcode construct, it could be shown *in vitro* as well as in a liver regeneration *in-vivo* model that the colour observed under the microscope and the sequenced "colour coding" of the barcodes correlated. In addition, experiments showed the feasibility of using "coloured barcoding" to analyse haematopoietic reconstitution as well as clonal leukaemia development (Cornils et al., 2014).

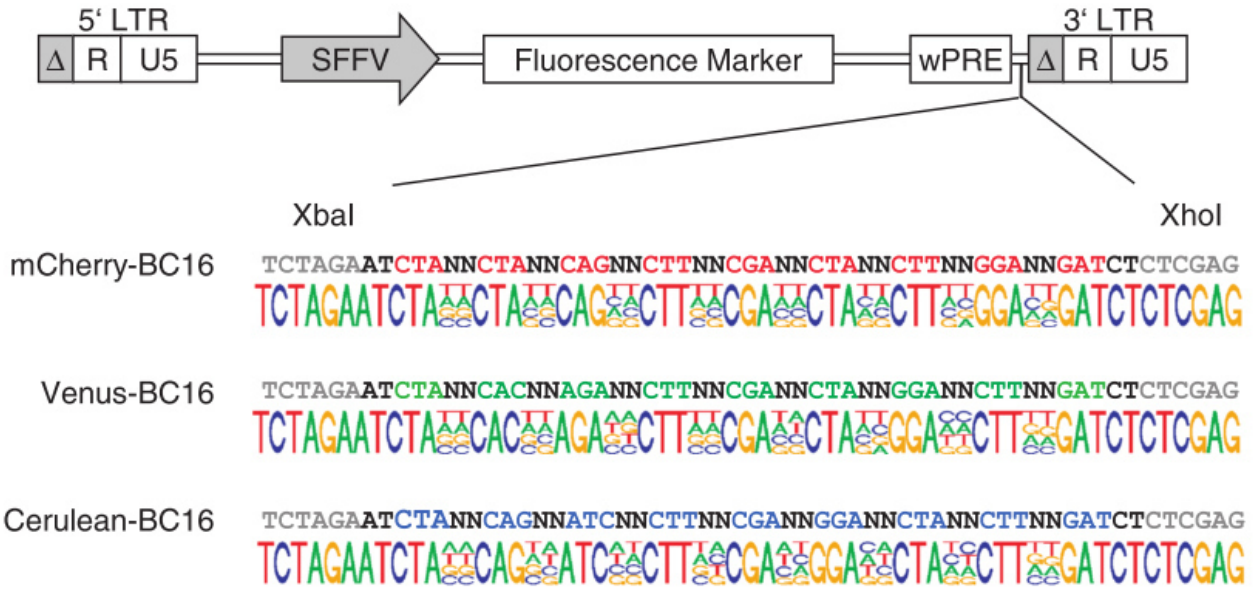


Figure 10 – The BC16 precursor of the barcoding system used in this work. The barcode is integrated near the 3'LTR of a lentiviral LeGO vector via *XbaI/XhoI*. The barcode itself consists of 16 variable positions ("N", wobble base), randomly determined when the oligo is produced (resulting in a 25% chance for each nucleotide). Wobble base pairs are interspaced by fixed triplets (barcode backbone, coloured red, green or blue, respectively), allowing identification of barcodes in the sequencing data as well as enabling the "colour coding" of barcodes (e.g., to a fluorescent protein like mCherry, Venus etc). This facilitated the use of several barcoded vectors in parallel in single animals. Figure taken from Cornils et al., (2014), CC BY 3.0 license.

In summary, genetic barcoding systems have several advantages compared to single-cell transplantations and/or LM- or LAM-PCR based integration site detection systems (some of the following points summarised in Grosselin et al., 2013). They massively reduce the number of mice needed to analyse clonal dynamics and behaviour of populations or single cells. They do not rely on accessibility for restriction enzymes, thus producing fragments of equal size and should (theoretically) enable easy and effortless quantification of clonal distributions using NGS. These systems enable competitive clonal analysis (shown in this work) and, by using either arrayed libraries or different barcode backbones, are able to link the readout to different (competing) experimental conditions or populations.

Of course, there are also some limitations of genomic barcoding (partially summarised by Grosselin et al., 2013). All systems studying haematopoietic reconstitution need to extract cells from donor animals, thereby potentially altering or at least influencing blood-cell homeostasis/regeneration. Retroviral vectors, so far used in most of the barcoding systems, *per se* may lead to insertional mutagenesis, as they require additional *ex-vivo* transduction and/or culture steps, possibly influencing cell fate (e.g., differentiation and proliferation capacities). If retroviral vectors are used, transduction rate is one critical parameter. Multiple integrations of viral vectors into one cell are hard to detect and/or correct. This may skew quantifications and frequencies in the analysis and tamper conclusions based on observed barcode numbers. It has been shown, that the expected percentage of cells with multiple integrations starts to rise with an transduction rate over 20%, following Poisson distribution (Fehse et al., 2004; Kustikova et al., 2003). Accordingly, transduction rates in barcode experiments should not exceed 20%. This is contrary to applications such as RGB marking or models, where high transduction rates are desired to express a marker gene in a high percentage of cells. Another element of uncertainty is the influence of the surrounding genome. As already discussed in section 5.3.1 (page 19), currently used retroviral vectors do not integrate randomly. Therefore, a barcoded cell developing clonal dominance may simply bear a vector integration near a (proto-)oncogene that affects the entire analysis. To determine the integration site, one has to go back to the already described, laborious, LM-PCR based techniques, which contradicts the barcoding concept. In addition, several publications demonstrated the importance and implications of barcode design, PCR-based errors or bias and strict quality control at all stages of the experiment (Blundell and Levy, 2014; Buschmann and Bystrykh, 2013; Bystrykh, 2012; Bystrykh and Belderbos, 2016; Deakin et al., 2014; Krueger et al., 2011; Thielecke et al., 2017).

With evolving techniques, it may be possible to overcome some of those limitations in the upcoming years. The concept of safe harbours, genomic regions that can be targeted for integration without the risk of genotoxic events, is highly attractive in the context of gene therapy. So far, three possible loci, namely AAVS1, CCR5 and ROSA26, have been proposed (reviewed in Sadelain et al., 2011). Targeting those sequences works (albeit at low frequencies), but the loci were shown to influence transgene expression (Lombardo et al., 2011; Rio et al., 2014). As development of safe-harbour integration techniques is fuelled by the large clinical implications, one can expect better methods in the near future. These can, and should, be adopted for barcoding too.

Besides, several systems have been proposed for *in-vivo* or *in-situ* barcoding using transgenic animals with different integrated barcode cassettes, using Cre recombinase based recombination or DNA invertases for shuffling of those cassettes (Peikon et al., 2014; Weber et al., 2016). However, these systems have only been presented as theoretical concepts *in silico*, or tested in *E. coli* so far. The first problem is getting those barcoding cassettes into the model organisms. The recent development of CRISPR/Cas gene editing may provide the necessary tools. Still, the main limitation is the length of the barcode cassettes as well as the uneven shuffling. The *in-silico* analysis of Weber et al., (2016) showed that the most frequent barcodes in their recombination system would have to be discarded. This is due to a biased recombination process, where some barcodes, generated via recombination of defined cassettes, would be more likely to appear. Thus, these highly abundant barcodes represent false-positive signals, rather than clonal influence of dominant clones. Depending on the exact recombination events and cycles, recombined barcodes have varying lengths, which (probably) leads to uneven retrieval and biased sequencing. Another problematical point is the general length of the barcode. Depending on the number of cassettes used, they may end up with final shuffling products of up to 628 bp. With the current technology, such fragments can hardly be sequenced, as the maximum read lengths in an IlluminaMiSeq instrument are 2x300 bp[8]. There are other systems coming up, however 3rd-generation sequencing machines able to generate sufficient read lengths have error rates of >10% to date, which are not acceptable for barcode analyses. Furthermore, the sequencing costs of such instruments are much higher in comparison with 2nd-generation systems (Rhoads and Au, 2015). High error rates and sequencing costs as well as the necessity to remove the majority of the reads for the previously mentioned reason of uneven recombination likelihood make the technology infeasible at the moment.

5.4.1. Hamming code and distance

The use of genetic barcoding relies on the unambiguous discrimination of "real" barcodes in the sample from "false positives", generated, through sequencing or PCR errors. In addition, an optimal barcode system should ensure that those individually small, but by high throughput accumulating, errors do not tamper the composition of the sample by converting one "real" barcode into another "real" one or changing two different "real" barcodes into two "real" and one false-positive barcode. Therefore, it is necessary to consider some error-correcting coding theories. There are several theories and principles of error-correcting codes (e.g., Reed and Solomon, 1960; Shannon, 1948). However, this work will focus on the one proposed by Hamming.

The Hamming code is a coding theory by Richard W. Hamming (Hamming, 1950). The purpose was to control the correct transfer of data in, at that time, big and complex computer systems, which were evidently error-prone. The binary hamming code consists of data-bits (db) and control-bits (p) generating code with length (n): $n = db + p$. Control-bits in combination with a checksum scheme allow for correction of a certain number of errors. This number mainly depends on the Hamming

[8] http://www.illumina.com/systems/sequencing.html

distance (d) between two (bar)codes, as displayed in the following example (Bystrykh, 2012). The Hamming distance between barcodes "AGC" and "AGT" is d = 1, because only one substitution is needed to change one into another. To change "ATT" to "AAA", two substitutions would be necessary, and converting "TAG" to "GTA" needs exchange of 3 positions. Error correction can be done for (t) errors using the formula $d_{min} = 2t + 1$. That means a minimal Hamming distance of 3 is required between 2 (bar)codes to reckon back one substitution to the original one.

In the context of genetic barcodes, the Hamming distance can be increased by increasing the number of variable positions (N), thereby enhancing the theoretical number of possible barcodes. The correlation is exponential, as the number of possible barcodes equals 4^N. The basis 4 reflects the four possible nucleotides at each position. Thus, a barcode with 16 variable positions (Cornils et al., 2014), can generate a maximum of $4^{16} = \sim 4.3x10^9$ barcodes. Computational modulations propose, that picking 10.000 barcodes randomly would result in an average Hamming distance of $d = 4.7$ (Thielecke et al., 2017, Figure S1). Using the already mentioned formula for correction of errors ($4.7 = 2t + 1$), this results in t = 1.85, which means that only one single substitution could be corrected for discrimination/identification of false positive barcodes. Even if the number of picked barcodes is lowered to 1000, a value closer to the biological situation in a transplantation model, the expected Hamming distance would be around $d = 6$ (t = 2.5), allowing correction of 2 PCR/sequencing errors. Additional variable positions lead to higher numbers of possible barcodes and higher Hamming distances, when picking a certain number of BCs. This allows for a more distinct discrimination and assignment of PCR/sequencing error generated barcodes (Thielecke et al., 2017).

5.4.2. The 32 wobble-base barcode system (BC32)

During my master thesis in 2012, we started with the development of the new BC32 system. Using the existing design of the 16 wobble base barcode (Cornils et al., 2014), we improved our system in multiple aspects (see Figure 11 for graphical representation). First, we doubled the number of variable positions from 16 to 32, achieving a maximal theoretical complexity of $4^{32} \approx 1.8x10^{19}$ different barcodes. Second, we added three additional variable positions in front of the first triplet to deal with the difficulty of cluster discrimination on the NGS flow cell for low-complexity samples (Krueger et al., 2011). During sequencing of the first nucleotides in a sequencing read, the positions of the individual clusters bound on the flow cell are defined. When sequencing genetic barcodes, the sequencing reaction begins few nucleotides upstream of the first variable position. In consequence, big parts of the flow cell light-up in the same colour because they share this consensus sequence, which would never happen with heterogeneous samples, e.g. transcriptomes. This led to some problems discriminating the individual bound clusters in the BC16 system. The three additional wobble positions upstream of the first backbone triplet resolve that issue. In theory, those nucleotides could be counted as additional wobbles, extending the barcodes to 35 wobble positions. The third improvement is the addition of the NGS adapters directly into the vector backbone. The 3-prime adaptor, a truncated version of the Illumina-Indexed-Adaptor is located directly behind the *Xho*I restriction site in our viral

vectors. The 5-prime, slightly modified, TruSeq Universal Adapter is integrated in the synthesised oligo. This offers several advantages: With the improved design of the BC32, a single PCR reaction with 30 cycles is sufficient to add the multiplexing indices and amplify barcodes in the sample. This product can directly be sequenced after PCR purification, combination of multiple samples and a final purification step. As we could show (Thielecke et al., 2017), extended PCR cycles (or multiple PCR steps) may introduce errors and, in addition, skew the distribution of barcodes in the samples, hampering quantitative conclusions. Therefore, reduction of PCR cycle numbers as well as purifications steps, during which a lot of sample is lost, is essential. The integration of the 3-prime sequencing adapter directly into the oligo also prevents the amplification of barcodes with missing/damaged adapters or empty vectors.

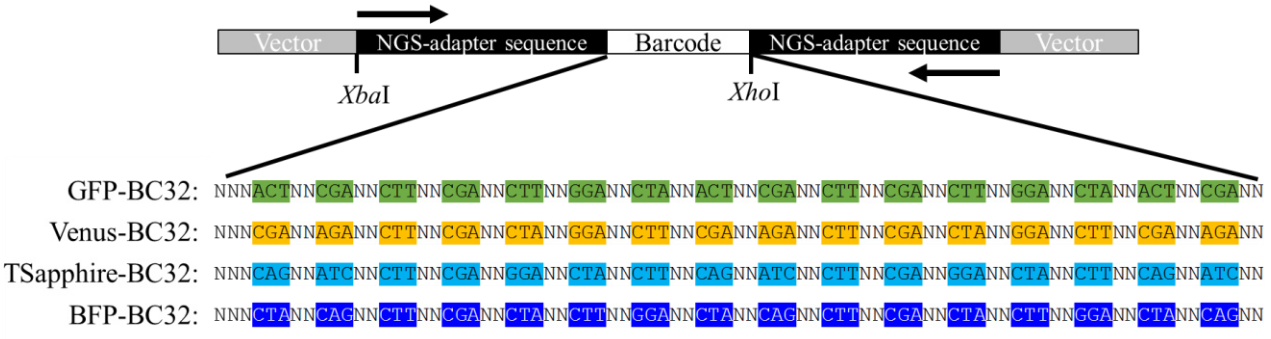


Figure 11 – Schematic overview of the improved 32 wobble barcode (BC32) design. The new barcodes consists of 35 variable positions ("N"). Each N is randomly determined during oligo synthesis and therefore represents 25% chance for each nucleotide. Pairs of those wobble bases are intercepted by fixed nucleotide triplets (coloured), defining a barcode backbone. Altering the sequence of those triplets allows generation of different barcoded backbones. Using a certain backbone in combination with another parameter, e.g., a Venus fluorescence protein, allows determination of the underlying construct by just analysing the barcode sequence. The first three variable positions are used for solving the difficulty of cluster calling on the flow cell in low complex samples and are not used in sequence evaluation (details in the text). The barcode is inserted into the barcode cloning site via *XbaI*/*Xho*I restriction sites. The 3-prime NGS-adapter is present in the cloning site, while the 5-prime one is located in the synthesised barcode oligo. This design allows for amplification as well as adding the necessary multiplexing indices for NGS in a single PCR step. Black arrows represent amplification primers. NGS: next generation sequencing.

Taken together, the new 32 wobble-base barcode system provides an, thus far, unmatched theoretical complexity of 10^{19} different barcodes, but still can be easily sequenced with the currently existing methods. The integration of the NGS adaptors into the barcode construct allows amplification as well as multiplexing in a single, 30-cycle, PCR, thus reducing the probability of PCR-induced errors and sample skewing. Additional variable nucleotides at the beginning of the barcode ensure proper cluster calling and improve the quality of the retrieved sequences. Altering the fixed triplet backbone of the BC32 allows another level of coding, e.g. by assigning a specific barcode backbone to a fluorescence protein (FP) or vector construct.

5.4.3. Genetic barcoding reveals further mechanisms of haematopoiesis

Development of lineage tracing and/or genetic barcoding systems has led to a variety of new insights about haematopoiesis and lineage differentiation in recent years. Genetic barcoding was used to study migration patterns of families of antigen-specific $CD8^+$ T cells (Schepers et al., 2008) and to show that only a small number of HSC-derived clones reconstitute the majority of the haematopoietic system after transplantation (Gerrits et al., 2010). Barcoding could also confirm the existence of two stem cell populations with long-term repopulation capacity but distinct lineage biases (Lu et al., 2011). That concept had been proposed earlier by another group after cumbersome single-cell transplantation experiments, calling those α-, β-, γ- and δ-cells (Dykstra et al., 2007). Thereby, α-cells preferentially yield myeloid cells, whereas β-cells generate output into the myeloid as well as the lymphoid lineage. γ- and δ-cells mainly contribute to the lymphoid lineage and show only limited self-renewing capacity. Arrayed (= sequence of the input barcodes is known) barcode studies confirmed the existence of those cell types and could further resolve the engraftment process, showing that few, in primary animals undetectable clones, are able to contribute to the engraftment of secondary recipients (Grosselin et al., 2013). These results imply the existence of quiescent HSC in primary transplants, becoming active after secondary transplantation. This population was undetectable with the previously used single-cell transplantation techniques. Another study revealed that the, long-term, pattern of barcoded haematopoietic reconstitution becomes stable approximately 3 months after transplantation (Verovskaya et al., 2013). In addition, the authors reported variations of unique lymphoid and myeloid clones and only a moderate correlation within the lymphoid lineage (between B and T cells), confirming the differentiation bias of HSC into certain lineages or/and cell types. They identified differences in the HSC pool of old (24 months) and young (4 months) HSCs, illustrating that multiple small clones form the old pool, whereas fewer, but larger, clones appear in the young HSC pool. Other groups used barcoding data in combination with bioinformatic modelling to refine (or redefine) the "classical" models of differentiation and lineage commitment (Höfer et al., 2016; Perié et al., 2014) and to show that imprinting of cells into lineages takes place early (Naik et al., 2013; Perié et al., 2015).

In contrast to the transplantation settings described so far, there have been some studies utilising *in situ* labelling/barcoding of cells to analyse steady-state haematopoiesis. Sun et al., (2014) used temporal restricted expression of a hyperactive Sleeping-Beauty transposase, generating barcode-like transposon tags in transgenic mice. Analysis of those tags revealed limited contribution of classical long-term HSC to blood production compared to multipotent progenitors in a steady-state system. Similar results were reported with another model using a fluorescence label (Busch et al., 2015). In contrast, the contribution of HSC to more than two-thirds of myeloid cells was reported using an inducible, fluorescence-based, labelling system (Sawai et al., 2016). The latter publication offered some possible explanations to the different results, mostly connected to leakiness of the inducible system(s) and low labelling rates in the 2015 study. Thus, the influence of HSC is still under discussion, as another group selectively depleted over 99% of the classical long-term HSC and still observed unaffected steady-state haematopoiesis (Schoedel et al., 2016). That group proposed that a large long-

term HSC pool only provides low-level input of new progenitors into the system. This results in a minimal number of divisions for each individual HSC, ensuring high genetic stability, under steady-state conditions. Still, it is unclear how far these data of steady-state naive haematopoiesis, observed under laboratory conditions, reflect the situation of a stressed, predated, pathogen-confronted mouse in the field.

Of course, barcoding has not only been used for haematopoietic lineage tracing, but also has produced considerable knowledge about other cell systems, e.g. mammary epithelial cells (Nguyen et al., 2014), cancer cells (Bhang et al., 2015), cell line propagation *in vitro* (Porter et al., 2014), thymus colonization and T cell maturation dynamics (Krueger et al., 2016; Ziętara et al., 2015), mesenchymal stromal cells (Bigildeev et al., 2016) as well as leukaemic development (Cornils et al., 2014, 2017; Klauke et al., 2015). In addition, monitoring of barcoded human cord blood $CD34^+$ cells transplanted into immunodeficient mice enabled the detailed analysis of growth and differentiation dynamics of these clinically highly attractive cells (Cheung et al., 2013). Recently, Wahlestedt et al., (2017) combined genetic barcoding with induced pluripotent stem cells to assess differences between young and aged HSCs.

Taken together, genetic barcoding has facilitated remarkable insights into lineage-fate decisions and lineage relationships.

6. Materials

The following lists contain reagents mentioned in multiple of the following methods (starting page 36). Reagents only used in one specific method are directly mentioned in that section.

6.1. Sanger sequencing

Sanger sequencing of vector constructs or barcodes was done by commercial sequencing services (Seqlab Göttingen, Germany or Eurofins Genomics Ebersberg, Germany) matching the requested parameters for DNA/Primer concentration and volume.

6.2. Enzymes

All restriction enzymes, if not otherwise specified, were purchased as “FastDigest” versions from Thermo Fisher Scientific (Waltham, USA) and used in accordance with the manufacturer’s instructions. Other enzymes are listed in Table 1.

Table 1 - List of enzymes used

Enzyme	Product number	Manufacturer
T4 DNA Ligase 5 U/µL	EL0011	Thermo Fisher Scientific
T4 DNA Ligase 30 U/µL	EL0013	Thermo Fisher Scientific
FastAP Thermosensitive Alkaline Phosphatase (1 U/µL)	EF0651	Thermo Fisher Scientific
Klenow Fragment (10 U/µL)	EP0051	Thermo Fisher Scientific

T4 Polynucleotide Kinase (10 U/µL)	EK0032	Thermo Fisher Scientific
DreamTaq DNA Polymerase (5 U/µL)	EP0702	Thermo Fisher Scientific
Q5® High-Fidelity DNA Polymerase	M0491L	New England Biolabs

dNTPs, if not included in buffer or mastermix, were mixed combining dATP (Thermo Fisher Scientific, #R0142), dTTP (#R0171), dCTP (#R0151) and dGTP (#R0161) to a final concentration of 10 µM each.

6.3. Antibodies for flow cytometry (FC)

Antibodies used for staining of mouse samples are listed directly in section 7.4, page 37.

6.4. Primers & Oligos

Barcode oligos with random nucleotides were purchased from TIB Molbio (Berlin, Germany).

Table 2 – Barcode oligo sequences

Name	Sequence (5'-3')
Poly-GFP-BC-fw	GGTGCATCTAGAACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNACTNNCGANNCTTNNCGANNCTTNN GGANNCTANNACTNNCGANNCTTNNCGANNCTTNNGGANNCTANNACTNNCGANNCTCGAGGTGCACTATG
Poly-Venus-BC-Fw	GGTGCATCTAGAACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNCGANNAGANNCTTNNCGANNCTANN GGANNCTTNNCGANNAGANNCTTNNCGANNCTANNGGANNCTTNNCGANNAGANNCTCGAGGTGCACTATG
Poly-Cerulean-BC-Fw (for TSapp. construct)	GGTGCATCTAGAACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNCAGNNATCNNCTTNNCGANNGGANN CTANNCTTNNCAGNNATCNNCTTNNCGANNGGANNCTANNCTTNNCAGNNATCNNCTCGAGGTGCACTATG
Poly-Cherry-BC-Fw (for the BFP construct)	GGTGCATCTAGAACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNCTANNCAGNNCTTNNCGANNCTANN CTTNNGGANNCTANNCAGNNCTTNNCGANNCTANNCTTNNGGANNCTANNCAGNNCTCGAGGTGCACTATG

"N" was randomly determined during oligo production, representing a 25% chance for each nucleotide.

Primers were ordered from Eurofins Genomics (Ebersberg, Germany) choosing "Salt Free" or HPLC purification method.

Table 3 – Primers used within this work

Name	Sequence (5'-3')	Use
32BarcodeMCS Fw	GTACCCAGCTGAATGATACGGCGACCACCGTCTAGATATAGCGCTAT AGCTCGAGAGATCGGAAGAGCACAAGTCTGAACTCCAGTCAC	Annealing of barcode cloning site
32BarcodeMCS Rv	GTGACTGGAGTTCAGACTTGTGCTCTTCCGATCTCTCGAGCTATAGC GCTATATCTAGACGGTGGTCGCCGTATCATTCAGCTGG	Annealing of barcode cloning site
TA1 BC-PCR-Seq	ACAGCAGCTACCAATGCTGA	Sequencing of barcodes in lentiviral constructs
32BC-Poly-fw	GGTGCATCTAGAACACTC	Create ds BC oligo, HPLC purified
32BC-Poly-rev	CATAGTGCACCTCGAG	Create ds BC oligo, HPLC purified
BC-MCS HindIII Fw	ATATAAGCTTCAGCTGAATGATACGGCGAC	Cloning barcode cloning site into alpha vectors
BC-MCS HindIII Rv	ATATAAGCTTATATGTGACTGGAGTTCAGACTTGTG	Cloning barcode cloning site into alpha vectors
TA2 Alpha-BC-Seq	GCCACGGCAGAACT	Sequencing of barcodes in alpharetroviral constructs
Sapp Fw NotI	ATATGCGGCCGCCACCATGG	TSapphire + NotI
Sapp Rv NotI	ATATGCGGCCGCTTACTTGTACAGC	TSapphire + NotI
Illu_P2 (43)	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT	Bridging oligo for sequencing
Illu_MPLX35 (164)	CAAGCAGAAGACGGCATACGAGAT TCTGAG GTGACTGGAGTTC	Index35
Illu_MPLX36 (165)	CAAGCAGAAGACGGCATACGAGAT CCTTGC GTGACTGGAGTTC	Index36
Illu_MPLX37 (166)	CAAGCAGAAGACGGCATACGAGAT TGGAGC GTGACTGGAGTTC	Index37
Illu_MPLX38 (167)	CAAGCAGAAGACGGCATACGAGAT TCGGGA GTGACTGGAGTTC	Index38
Illu_MPLX39 (168)	CAAGCAGAAGACGGCATACGAGAT AAACCT GTGACTGGAGTTC	Index39
Illu_MPLX40 (169)	CAAGCAGAAGACGGCATACGAGAT CTCTAC GTGACTGGAGTTC	Index40
Illu_MPLX41 (170)	CAAGCAGAAGACGGCATACGAGAT CGGCCT GTGACTGGAGTTC	Index41

Illu_MPLX42 (171)	CAAGCAGAAGACGGCATACGAGAT CCGGTG GTGACTGGAGTTC	Index42
Illu_MPLX43 (172)	CAAGCAGAAGACGGCATACGAGAT CAGCAG GTGACTGGAGTTC	Index43
Illu_MPLX44 (173)	CAAGCAGAAGACGGCATACGAGAT AAGTGC GTGACTGGAGTTC	Index44
Illu_MPLX45 (174)	CAAGCAGAAGACGGCATACGAGAT CAGGCC GTGACTGGAGTTC	Index45
Illu_MPLX46 (175)	CAAGCAGAAGACGGCATACGAGAT GGTAGA GTGACTGGAGTTC	Index46
Illu_MPLX47 (176)	CAAGCAGAAGACGGCATACGAGAT CCAGCA GTGACTGGAGTTC	Index47
Illu_MPLX48 (177)	CAAGCAGAAGACGGCATACGAGAT GCGCCA GTGACTGGAGTTC	Index48
Illu_MPLX49 (178)	CAAGCAGAAGACGGCATACGAGAT GGAACT GTGACTGGAGTTC	Index49
Illu_MPLX50 (179)	CAAGCAGAAGACGGCATACGAGAT GCGGAC GTGACTGGAGTTC	Index50
Illu_MPLX51 (180)	CAAGCAGAAGACGGCATACGAGAT CGAAAC GTGACTGGAGTTC	Index51
Illu_MPLX52 (181)	CAAGCAGAAGACGGCATACGAGAT CCACTC GTGACTGGAGTTC	Index52
ILL_Dual_P5-01	AATGATACGGCGACCACCGAGATCTACAC AGCTTAGT ACACTCTTTCCCTACACGACGCTCTTCCGATCxT	Dual Index01
ILL_Dual_P5-02	AATGATACGGCGACCACCGAGATCTACAC GCTACTTG ACACTCTTTCCCTACACGACGCTCTTCCGATCxT	Dual Index02
ILL_Dual_P5-03	AATGATACGGCGACCACCGAGATCTACAC CTAGGCAC ACACTCTTTCCCTACACGACGCTCTTCCGATCxT	Dual Index03
ILL_Dual_P5-04	AATGATACGGCGACCACCGAGATCTACAC TAGCAGCA ACACTCTTTCCCTACACGACGCTCTTCCGATCxT	Dual Index04
TA35 A-pd_Fw	ACCCCAGGCACGTCTTTG	ddPCR - Primer set A Alpha-pd construct
TA36 A-pd/L-pd_Rv	CGCTGAACTTGTGGCCGT	ddPCR - Primer set A Alpha-pd construct
A-pd-p	FAM-AGCGGCCGCCACCATGGT-BHQ1	ddPCR - Primer set A Alpha-pd construct
TA38 EFS_Fw	TTGAACCGGTGCCTAGAGAAG	ddPCR - Primer set A EFS construct
TA39 EFS_Rv	CGGCGACTACTGCACTTATATACG	ddPCR - Primer set A EFS construct
EFS-p	FAM-CTGGCTCCGCCTTTTTCCCGA-BHQ1	ddPCR - Primer set A EFS construct
TA40 SFFV_Fw	ACCCTGCGCCTTATTTGAATT	ddPCR - Primer set A SFFV construct
TA41 SFFV_Rv	TTATAGAGCTCGGGAAGCAGAAG	ddPCR - Primer set A SFFV construct
SFFV-p	FAM-CCAATCAGCCTGCTTCTCGCTTCTGTT-BHQ1	ddPCR - Primer set A SFFV construct
32 dPCR Fw (XbaI)	GCGACCACCGTCTAGAACAC	ddPCR - Primer set B
32dPCR P	HEX-CCCTACACGACGCTCTTCCGA-BHQ1	ddPCR - Primer set B
49164_1_Tsap	CTCTTTCCGAGTGGATGCTAG	ddPCR – BM #49164 TSapp clone 1 specific Rv
42329_1_BFP	GTAAGCCTAGATTCGCCAAGTT	ddPCR – BM #42329 BFP clone 1 specific Rv
49432_1_GFP	CCAAGGCTCGCCAAGGT	ddPCR – BM #49432 GFP clone 1 specific Rv
46579_1_GFP	CCCAAAGGGTCGTTAAGGC	ddPCR – BM #46579 GFP clone 1 specific Rv
49164_1_BFP	CCAAAGACTAGAATCGCGAAGTA	ddPCR – Spleen #49164 BFP clone 1 specific Rv
46316_3_GFP	GCTCCTAAAGAATCGGTAAGCG	ddPCR – PB #46316 GFP clone 3 specific Rv

x: phosphorothioate bond, BHQ1: Black Hole Quencher 1

6.5. Kits

All listed kits were used according to manufacturer's instructions if not specified otherwise in the Methods section.

Table 4 – Kits used for DNA extraction, purification and lineage depletion

Kit name	Product number	Manufacturer
QIAquick PCR Purification Kit	28106	Qiagen, Hilden, Germany
QIAprep Spin Miniprep Kit	27106	Qiagen
QIAGEN Plasmid *Plus* Maxi Kit	12965	Qiagen

QIAamp DNA Blood Mini Kit	51106	Qiagen
QIAamp DNA Micro Kit	56304	Qiagen
QIAquick Gel Extraction Kit	28706	Qiagen
Lineage Cell Depletion Kit, mouse	130-090-858	Miltenyi Biotec, Bergisch Gladbach, Germany

6.6. Instruments

Instruments not specifically mentioned here, e.g. pipettes and thermoblocks, matched the standard requirements of laboratory equipment.

Table 5 – Instruments used in this work

Instrument	Name	Manufacturer
Thermocycler	Biometra Professional & Biometra Professional gradient	Biometra (now Analytik Jena, Jena, Germany)
Thermocycler	Eppendorf Mastercycler & Eppendorf Mastercycler gradient	Eppendorf, Hamburg, Germany
Electroporation system	Gene Pulser Xcell	Bio-Rad, Hercules, USA
Centrifuge	Sorvall RC-5C Plus with HB-6 swinging bucket rotor	Thermo Fisher Scientific
FACS analyser	BD FACS CantoII 407, 488 and 633 nm Laser	BD Biosciences
FACS sorter	BD FACS AriaIIIu 407, 488, 561 and 633 nm Laser	BD Biosciences
FACS sorter	BD FACS AriaFusion 355, 405, 488, 561 and 643 nm Laser	BD Biosciences
NGS Sequencer	Illumina MiSeq (located at TU Dresden)	Illumina, San Diego, USA
Digital droplet PCR	QX100 system	Bio-Rad, Hercules, USA

6.7. Laboratory plastic ware

All plastic ware used in this work matched the standard requirements of laboratory equipment. If required, special products are mentioned in the text.

6.8. Buffers and growth media

Buffers and media are listed in alphabetical order:

DMEM (DMEM, high glucose, GlutaMAX™ Supplement, pyruvate, Life Technologies, Carlsbad, USA, # 31966021)

Additives

- 10% Fetal Bovine Serum (Sigma Aldrich, St. Louis, USA, #F7524)
- 1% Penicillin-Streptomycin (Life Technologies, Carlsbad, USA, # 15140122)
- 20 mM HEPES (Life Technologies, Carlsbad, USA, # 15630056)
- HEPES was normally only added for production of viral particles. However, once added the medium was also used for all other described uses.

DPBS, no calcium, no magnesium (Life Technologies, Carlsbad, USA, #14190094)

Ery-lysis buffer
- 155 mM NH_4Cl
- 10 mM $KHCO_3$
- 87 µM $C_{10}H_{12}K_2MgN_2O_8 * 2H_2O$
- pH 7.4 (adjusted via KOH) and sterilised via 0.2 µM filtration

2xHBS
- 275.8 mM $CaCl_2$
- 10.2 mM KCl
- 1.41 µM Na_2HPO_4
- 42 mM HEPES
- 1.1 mM Glucose
- pH 7.05 adjusted via NaOH and sterilised via 0.2 µM filtration

LB-agar-amp plates (Fast-Media® Amp Agar, InvivoGen, San Diego, USA, # fas-am-s)

LB medium (Lennox, Carl Roth, Karlsruhe, Germany, #X964.2)
- 0.1 mg/mL (f.c.) ampicillin added before use (Carl Roth, Karlsruhe, Germany, #K029.1)

MACS buffer
- PBS (Life Technologies, Carlsbad, USA, #14190094)
- 0.5% BSA (Sigma Aldrich, St. Louis, USA, #A8412)
- 2 mM EDTA (Sigma Aldrich, St. Louis, USA, #E5134)

EDTA Stock solution is sterilised via 0.2 µM filtration. BSA and PBS are already sterile.

StemSpan Medium SFEM (Stemcell Technologies, Vancouver, Canada #09650)
Additives:
- 1% Penicillin-Streptomycin (Life Technologies, Carlsbad, USA, # 15140122)
- 1% Sodium Pyruvate (Life Technologies, Carlsbad, USA, # 11360039)
- 1% L-Glutamine (Life Technologies, Carlsbad, USA, # 25030024)
- mTPO (PeproTech, Rocky Hill, USA, #315-14; final concentration 20 ng/mL)
- mSCF (PeproTech, Rocky Hill, USA, #250-03; final concentration 10 ng/mL)
- hFGF-a (PeproTech, Rocky Hill, USA, #100-17A; final concentration 10 ng/mL)
- mIGF-II (R&D Systems, Minneapolis, USA, #792-MG; final concentration 20 ng/mL)

Cytokine aliquots were only used for 1 week after thawing from -80°C.

7. Methods

7.1. DNA extraction

7.1.1. Plasmids and genomic DNA

Mini and maxi preps of cloned constructs were done using QIAprep Spin Miniprep Kit or QIAGEN Plasmid Plus Maxi Kit (both Qiagen, Hilden, Germany) according to manufacturer's instructions.

7.1.2. Primary murine samples

DNA extraction from primary murine samples was done using the QIAamp DNA Blood Mini Kit or QIAamp DNA Micro Kit according to manufacturer's instructions (Qiagen, Hilden, Germany). Special care was taken to avoid barcode cross contamination of the samples. All steps were done using filter tips and the workspace was cleaned regularly using DNA removing cleaning agents.

DNA from blood, bone marrow, lineage-positive and spleen samples was extracted with the Blood Mini Kit. Because of the low cell numbers in the sorted subsets (T cells, B cells and Granulocytes) and in the lineage-negative fraction, DNA of these samples was extracted with the Micro Kit. Elution volumes were based on the initial cell number ranging from 27 µL (Ly.6G samples) to 107 µL (most lin- samples). Elution times at the end of the procedure were extended up to 30 min after addition of elution buffer to achieve maximum DNA yield. DNA concentration of the samples was determined afterwards using a Qubit 2.0 spectrophotometer (dsDNA BR assay, Thermo Fisher Scientific).

7.2. DNA concentration measurement

DNA concentration measurements were carried out either using a NanoDrop 1000 spectrophotometer or a Qubit2.0 System (both Thermo Fisher Scientific) according to manufacturer's instructions. Depending on the sample, different Qubit assays (Thermo Fisher Scientific) were used: Qubit ssDNA Assay Kit (Q10212), Qubit dsDNA BR Assay Kit (Q32850) or Qubit dsDNA HS Assay Kit (Q32851), all of them in combination with the recommended Qubit Assay Tubes (Q32856).

7.3. Gel electrophoresis

Electrophoresis gels were casted using UltraPure Agarose (formerly Invitrogen, now Thermo Fisher Scientific, #16500500) and TAE buffer (UltraPure DNA Typing Grade 50X TAE Buffer, Thermo Fisher Scientific, #24710030, diluted with desalted H_2O to 1x TAE). For DNA detection, peqGREEN (Peqlab, Wilmington, USA) was added using the lower limit of the manufacturer's recommendation (2 µL for 50 mL gel and 4 µL for 100 mL). Depending on DNA fragment size gels with 0.8 – 2% agarose were used. Electrophoresis itself was running with ~5 – 8.5 V/cm.

7.4. Flow cytometry and antibody staining

Flow cytometric (FC) staining was done using the following antibodies.

Table 6 – Flow cytometry antibodies and volumes used in this work

Antibody	Clone	Manufacturer	µL per 10^6 cells
Fc-Block	93	Biolegend	2
PE anti mouse CD45.1	A20	Biolegend/ BD Biosciences	1
APC anti mouse CD45.2	104	BD Biosciences	1
PE anti mouse ScaI	E13-161.7	BD Biosciences	1
APC anti mouse cKit	2B8	BD Biosciences	1
PE anti mouse Ly-6G and Ly-6C (Gr-1)	RB6-8C5	BD Biosciences	1
PE anti mouse Ly.6G	1A8	Biolegend	1

APC anti mouse CD3e	145-2C11	Biolegend	1
PE anti mouse B220	RA3-6B2	BD Biosciences	1
APC-Cy7 anti mouse B220	RA3-6B2	Biolegend	0.3

During the project, the commercial source of some antibodies was changed sometimes. However, the antibody clone was kept constant and titration for optimal staining result of the new antibody was performed. In the beginning, B220 and CD3e (PE/APC) were stained in parallel on spleen cells while Gr-1 (PE) was stained in a separate tube. Later on, this was combined into one reaction with three antibodies. Therefore, an APC-Cy7 conjugated B220 antibody was chosen. At the same time, the Gr-1 antibody, detecting both Ly-6G and Ly-6C, was replaced by a Ly-6G only antibody, due to $CD3e^+$ and $Ly\text{-}6C^+$ expressing lymphocytes.

Cells were stained in a small volume of PBS (~100 µL, for high cell numbers, e.g. the spleen sorting, the volume was increased) in FACS tubes or 15-mL falcons. Unspecific binding was blocked by adding Fc-Block and incubating 5 min at room temperature (RT) before adding the appropriate amount of antibody. After 20 – 30 min at 4°C, cells were washed twice with PBS (centrifugation: 300g, 4°C, 7 min) and resuspended in ~150 µL PBS for FC analysis (see also section 7.9.3, page 44).

FC analyses were performed at the FACS Sorting Core Unit of the University Medical Center Hamburg-Eppendorf on a FACS CantoII using BD FACSDiva software. Fluorescence activated cell sorting (FACS) was done on a BD FACS AriaIIIu or BD FACS AriaFusion.

7.5. Cloning of barcode vector constructs

7.5.1. Lentiviral vectors

The initial cloning of the barcode multiple cloning site was part of my master thesis in 2012. Briefly, the barcode multiple-cloning site (Barcode-MCS) containing the *Xba*I/*Xho*I restriction sites for barcode insertion and the 3' Illumina sequencing adaptor was generated by annealing the oligos 32BarcodeMCS Fw and 32BarcodeMCS Rv. Therefore, 20 µL of each oligo (200 µM) were added to 40 µL H_2O and incubated for 5 min at 85°C in a water bath. 20 µL 5x annealing buffer (0.5 M Tris (pH 7.4), 0.35 M $MgCl_2$) were added before switching off the heat. The mixture was removed from the water bath the next morning and cloned into an LeGO pV2 vector (described in Weber et al., 2008) via *Acc65*I/*Pvu*II. Therefore, LeGO pV2 was digested with the mentioned enzymes for 30 min and purified on an agarose gel. After gel extraction 200 ng digested vector and 5 µL annealed oligo were ligated in 10 µL (1 µL T4 Ligase 5 U/µL) overnight at 16°C. The next afternoon, transformation into chemically competent TOP10F (self-made, competence > 10^6 cfu/µg) was conducted and bacteria were plated onto LB-agar-amp plates. Single colonies were picked after incubating the plate over night at 37°C, and cultured in 13 mL tubes containing 2 mL LB-Amp medium over night at 37°C at 180 rpm. The next day, DNA was extracted using the QIAprep Spin Miniprep Kit according to manufacturer's instructions. Mini preps were digested with *Xba*I/*BspE*I and *Xho*I/*BspE*I, and the pattern was checked on an agarose gel. Correct sequences of clones with the expected patterns were confirmed by sequencing with TA1 BC-PCR-Seq. Using the clone with the correct sequence as matrix, different

promoters were cloned in combination with fluorescent proteins to achieve the desired vector constructs for this thesis (Figure 15, page51). Most promoters and fluorescence proteins were already available in the LeGO system (http://www.lentigo-vectors.de). Fluorescent proteins were exchanged using the flanking *BamH*I and *EcoR*I restriction sites while promoters were exchanged via *Not*I/*BamH*I. The EFS promoter was excited from a LeGO EFS-G2 plasmid (provided by K. Riecken). The promoter-deprived lentiviral construct was generated using the already mentioned pV2-BC-MCS plasmid and removing the SFFV promoter via *Not*I/*BamH*I. Afterwards, sticky ends were blunted using Klenow fragment and the plasmid was self-ligated. All cloned constructs were checked for correctness by sequencing and plasmid maxi preps were prepared using the QIAGEN Plasmid Plus Maxi Kit according to manufacturer's instructions. During the course of this work, a random sequence stuffer fragment (initially ~3.7 kb) was inserted between the *Xba*I/*Xho*I restrictions sites. This allows for a visual confirmation on the agarose gel if the restriction of the backbone was successful before inserting the barcode fragment. A small fraction of plasmids still was observed to contain the stuffer fragment after barcode ligation. Hence, the stuffer used beforehand was exchanged with a new, "toxic" stuffer (tSt, ~2.6 kb) containing a ccdB suicide gene under control of a lac promoter. This suicide gene reduces the likelihood of non-barcoded constructs, as it kills most commonly used bacteria strains. Plasmids containing a tStuffer have to be grown in ccdB resistant bacteria (founders: One Shot ccdB Survival 2 T1R, Thermo Fisher Scientific, self made batch > $5x10^6$ cfu/µg). Maps of the final vector constructs used for this word are provided in the appendix (section 14, starting page 105).

7.5.2. Alpharetroviral vectors

The initial alpharetroviral plasmid (pAlpha.SIN(noTATA)EFS.EGFP.wPRE) was kindly provided by A. Schambach and J. Suerth from Hannover Medical School and is similar to the one described in Suerth et al., 2012. Contrary to the published construct, it carries an EFS instead of an SFFV promoter. The first step was to remove the *Xho*I restriction site located 3' of the vector-coding region, as *Xho*I is needed for barcode insertion. The plasmid was digested with *Xho*I and sticky ends were blunted using Klenow fragment. After self-ligation the plasmid was transformed into TOP10F bacteria and plated onto LB-Amp-agar plates. DNA was extracted from picked single colonies after overnight incubation at 37°C and digestion with *Xho*I/*Sal*I confirmed the deletion of the *Xho*I site. The next step was the insertion of the barcode MCS, already mentioned in the previous section, 3' of WPRE- 5' of the DRE-element. To achieve this, a PCR was conducted using Q5 polymerase and primer BC-MCS HindIII Fw and BC-MCS HindIII Rv on a lentiviral LeGO pC2-BC mini. These primers added/exchanged *Hind*III restriction sites at both ends of the barcode MCS. The PCR fragment as well as the alpharetroviral plasmid were digested with *Hind*III and the latter one dephosphorylated with FastAP to prevent self-ligation. After clean-up of the reactions, sticky-end ligation of the two fragments was performed. The correct sequence of the barcode cloning site was confirmed by sequencing with primer TA2 Alpha-BC-Seq. To create the promoter-deprived, TSapphire-carrying alpharetroviral, construct, the EFS-GFP cassette of the A-EFS-GFP-vMCS plasmid was removed by *NotI* digestion, FastAP dephosphorylation and purification via agarose gel. In parallel, a PCR was performed using the lentiviral LeGO-S vector (provided by K. Riecken), Q5 polymerase and primers Sapp Fw NotI and

Sapp Rv NotI. The amplified TSapphire fragment contained NotI sites at both ends and was, after digestion with *NotI* and purification via agarose gel, ligated into the digested backbone. Finally, the toxic stuffer (see page38) was inserted into the alpharetroviral constructs, too. Maps of the final vector constructs used for this word are provided in the appendix (section 14, starting page 105).

7.6. Generation of barcode plasmid libraries

For barcoding vector constructs, barcode oligos (page 33) were diluted to 100 ng/µL with H_2O (50 µM oligo stock has approx. 1500 ng/µL, Qubit ssDNA assay). Double-stranded barcodes were generated using the following reaction and PCR program (Lid 99°C, preheat on).

Table 7 – PCR setup and program for barcode double-strand generation

[µL]	
1	Barcode oligo (100 ng/µL)
1.25	32BC-Poly-Fw (10 mM)
1.25	32BC-Poly-Rv (10 mM)
0.5	dNTP (10 mM)
5	Q5 reaction buffer
0.25	Q5 DNA Polymerase (NEB)
15.75	H_2O

8 reactions in parallel

Time [s]	Temp. [°C]
30	98
10	98
20	59
20	72
Go to step 2	Rep 9
180	72
Hold	4

After PCR clean-up (QIAquick PCR Purification Kit according to manufacturer's instructions), two PCR reactions were pooled into one clean-up reaction and each DNA was eluted from the column with 50 µL 37°C prewarmed elution buffer for 10 min. All four cleaned reactions were pooled (~200 µL total) and 24 µL Fastdigest buffer as well as 6 µL *Xba*I + 6 µL *Xho*I were added. The reaction then was split again into four tubes, containing 60 µL each and incubated for 1.5 h at 37°C. In parallel, 5 µg DNA of the vector plasmid were digested with 2.5 µL *Xba*I and *Xho*I in a total volume of 60 µL for 2.5 h at 37°C. Afterwards, dephosphorylation was performed by adding 1 µL of FastAP and incubating for one additional hour at 37°C, to reduce backbone self-ligation capacity. Barcode (~1.5% agarose) and backbone (~0.8% agarose) were then purified via agarose gel. UV exposition was kept to a minimum to prevent DNA damage. Fragments of the desired size (backbone: ~7kb, barcode 128 bp) were cut tightly out of the gel using a sharp scalpel. For gel extraction (QIAquick Gel Extraction Kit) fragments from 2 lanes were pooled into one tube and at the end eluted using 30 µL of elution buffer prewarmed to 50°C in a 10 min incubation step. Both barcode extractions were pooled and DNA concentration was determined using a Qubit spectrophotometer (dsDNA BR Assay, barcode concentration range: 5 – 20 ng/µL).The gel-extracted vector construct was eluted in 50 µL elution buffer. In the next step, the digested barcode was ligated into the vector backbone. To do so, 500 ng vector were mixed with 3x molar excess of barcode, calculated with the formula below.

$$ng_{barcode} = molar\ excess_{barcode} * \frac{bp_{\ barcode}}{bp_{\ backbone}} * ng_{backbone}$$

The reaction volume of the ligation varied, depending on the barcode concentrations recovered. The maximum volume was 60 µL in a 0.2-mL PCR tube. 10x T4 DNA ligation buffer was added to the

reaction to a final concentration of 1x, and 2.5 µL high-concentrated T4 DNA Ligase (30 U/µL) completed the reaction mix. Ligation was performed at 16°C for 12 h followed by 65°C for 10 min to inactivate the enzyme before cooling to 4°C. Prior to the electroporation of barcode plasmid the ligation reaction had to be desalted. Therefore, the ligation reaction was pipetted onto a membrane filter (MF-Millipore Membrane Filter 0.025 µm, VSWP01300, Merck Millipore, Billerica, USA) floating in a 10 cm dish filled with ultrapure H_2O. After approx. 30 min of floating, the remaining reaction was aspirated carefully and pipetted into a new tube. To achieve high transformation efficiencies for the electroporation of barcode plasmids into MegaX DH10B T1R Electrocomp. Cells (C640003, Thermo Fisher Scientific), electroporation cuvettes (Gene Pulser/MicroPulser Electroporation Cuvettes, 0.1 cm gap, 1652089, Bio-Rad) as well as the dialysed ligation were chilled on ice. 1 mL recovery medium (Thermo Fisher Scientific, included with DH10B) was prewarmed to 37°C. DH10B bacteria were gently thawed from - 80°C and 40 µL were carefully mixed with the ligation avoiding air bubbles. The mixture was then pipetted into the cuvette and the electroporation pulse (1.8 kV, 200 Ω, 25 µF) triggered. Afterwards, recovery medium was used to flush the cuvette and transfer the bacteria back into the 1.5 mL vial. Transformed bacteria were incubated at 37°C for 1 h (shaking at approx. 600 rpm). After 1 hour, 20 µL of the bacteria suspension were transferred into 180 µL LB medium (1st dilution) and further diluted (1:10) 3 times (2nd, 3rd, and 4th dilution). 100 µL of the dilution containing 10 µL of the original transformed reaction (1st dilution) as well as 100 µL of dilutions 3 and 4 (containing 0.1 and 0.01 µL of the transformed reaction) were plated onto LB-Agar-amp plates and incubated at 37°C overnight. These dilutions were used to roughly calculate the number of transformants, and therefore barcodes, in the preparation. The remaining ~1000 µL of the transformed reaction were pipetted into an Erlenmeyer flask containing 120 mL LB-Amp medium and incubated at 37°C and 180-200 rpm overnight. The next morning, colonies grown on the 0.01 and 0.1 µL plates were counted and an estimation of transformants was calculated using the formula below.

$$Number\ of\ barcodes = colonies_{counted} * \frac{Electroporation\ recovery\ volume\ (= 1000\ \mu L)}{recovery\ volume\ plated\ on\ dish\ [\mu L]}$$

Our barcode libraries were calculated to contain between $7x10^5$ to $4.3x10^6$ barcodes. DNA from the 120 mL bacteria culture was isolated using the QIAGEN Plasmid Plus Maxi Kit according to manufacturer's instructions. For quality control, 15-20 colonies from each transformation were picked from the dishes, DNA extracted and their barcodes were Sanger sequenced.

7.7. Production of lenti- and alpharetroviral vectors

Viral vectors used in this work were pseudotyped with vesicular stomatitis virus glycoprotein (VSV-G) envelope. Thus, they were potentially able to transduce human cells and therefore assigned to biosafety level 2 with all resulting guidelines and requirements.

The barcoded plasmid libraries produced were used for production of lenti- or alpharetroviral vectors. The protocol for lentiviral vector production is available online[9]. 5 million 293T cells per 10 cm dish

[9] http://www.lentigo-vectors.de

were seeded in 10 mL DMEM + additives (page 35) in the afternoon of day 1. Most times, 12 dishes were produced per construct. Early on day 2, required amounts of plasmids (Table 8) were mixed and H_2O was used to increase the final volume to 450 µL before adding 50 µL $CaCl_2$ (2.5 M). The whole mixture was pipetted drop wise into 500 µL 2x HBS buffer, prepared in a 15 mL falcon tube, while blowing air through the HBS with a Pasteur pipette. While incubating the mixture 15 – 20 min at RT, the old medium from the cells was replaced by 10 mL DMEM + additives, containing 25 µM chloroquine (Sigma Aldrich). 1 mL of the HBS/DNA solution was added drop-wise to the cells and swirled gently. Dishes were then incubated for ~6 – 8 h at 37°C, 5% CO_2 and the old medium was replaced by 10 mL fresh DMEM + additives but without chloroquine. 24 hours later, viral supernatant was harvested and filtered through a 0.45 µm filter to remove cellular debris (FP 30 mm Cellulose Acetate Syringe Filter, 0.45 µm, GE Healthcare, Little Chalfont, U.K, #10462100). If no further concentration of the virus was planned, the supernatant was aliquoted and stored at –80°C. If further concentration was required, filtered supernatant was put into 30 mL centrifuge tubes (Kimble Chase, Vineland, USA, #45500-30) sealed with parafilm and centrifuged overnight (~15 h, 4°C, 8000g, Sorvall RC-5C Plus centrifuge (Thermo Scientific Fisher) with HB-6 swinging bucket rotor and rubber adaptors for 30 mL tubes). The next morning, supernatant was carefully aspirated until only the desired volume of 300 – 600 µL remained. If multiple tubes had been used for centrifugation, resuspended viral particles from the same construct were pooled and then divided into aliquots to ensure equal distribution. Viral particles were frozen at -80°C. Aliquots used for transplantation experiments were only thawed once.

Table 8 – Quantity of plasmids used for production of lenti- and alpharetroviral vectors

Lentiviral vectors	
Plasmid	µg per dish
vector	10
gag/pol	10
VSV-G	2
rev	5

Alpharetroviral vectors	
Plasmid	µg per dish
vector	5
gag/pol	2.5
VSV-G	1.5

The alpharetroviral gag/pol plasmid (pcDNA3.Alpha.gagpol.co) kindly provided by A. Schambach and J. Suerth (Hannover Medical School) was described previously (Suerth et al., 2010). Lentiviral gag/pol (pMDLg/pRRE), rev (pRSV-Rev) and the VSV-G envelope (phCMV-VSV-G) were described by Weber et al., (2008).

7.8. Titre determination of lenti- and alpharetroviral vectors

Titre of viral supernatant was determined on 293T cells, the protocol used is available online[10]. 50,000 cells were seeded into a 24-well plate in 500 µL DMEM + additives and with 8 µg/mL Polybrene (Sigma Aldrich) per well. After attachment of the cells (2 – 4 h) varying amounts of viral supernatant,

[10] http://www.lentigo-vectors.de

for concentrated virus typically 10, 1, 0.1 and 0.01 µL and for non-concentrated virus 100, 10, 1 and 0.1 µL, were added. Analysis was performed in triplicates. For the highest volume, only a single well was prepared as those cells normally are 99% positive and die due to high virus load. The plate was then centrifuged for 1 h at 1000g at RT and incubated afterwards at 37°C, 5% CO_2. After approx. 5 – 6 h, 500 µL fresh medium, without polybrene, was added to each well. On day 4, cells were analysed via flow cytometry on a BD FACS CantoII, measuring the percentage of fluorescent-protein expressing cells. Virus titre [infectious particles/mL] was calculated using the formula

$$Titre = \frac{number\ of\ plated\ cells * \frac{percentage\ of\ transduced\ cells}{100}}{volume\ of\ supernatant\ added\ [mL]}$$

Typical titres obtained were in the range of $8x10^7$ – $7x10^8$ for both types of vectors for concentrated viral supernatants.

7.9. Animal procedures

All animal procedures in this work were conducted in accordance with the regulatory guidelines and approval from the local authorities (Behörde für Gesundheit und Verbraucherschutz - Veterinärwesen/Lebensmittelsicherheit) and the University Medical Center Hamburg-Eppendorf (UKE). Animals were kept in IVC cages with 3 – 4 animals per cage. Food and water were available *ad libitum* and nesting material was provided. Water soaked food pellets were provided after daily monitoring in the first two weeks after irradiation/transplantation to support the animals. Later on, animals were visited/monitored daily and body weight was determined twice a week. Additional soaked food pellets were provided after weight monitoring, weekly cage exchange and after taking blood samples.

Animals used in the work were bred in-house at the Forschungstierhaltung of the University Medical Centre Hamburg-Eppendorf (UKE). Recipient mice were female wild type C57BL/6J. The strains were imported either 2010 or 2014 into the mouse facility of UKE from Jackson Laboratory (Bar Harbor, USA, #00664). In-house bred, male, donor mice were B6.SJL-Ptprc_aPep3_b/BoyJ (initially from Jackson Laboratory, #002014) carrying the CD45.1 point mutation that can be used for chimaerism determination given that wild type animals are CD45.2.

7.9.1. Transduction and transplantation of lineage-negative cells

Approximately 8-week old male, CD45.1 donor mice were sacrificed on day 1 via cervical dislocation after sedation with 95% CO_2 + 5% O_2. *Tibia*, *Femur* and *Ilium* were removed, cleansed from flesh and kept in DMEM medium at 4°C. Bones pooled from 2-3 donors were crushed using mortar and pestle, washed with PBS and filtered (CELLSTAR EASYstrainer, Cell Strainers, 70 µm Greiner Bio-One, Kremsmünster, Austria). After centrifugation (300g, 5 min, 4°C) the pellet was resuspended using 5 mL ery-lysis buffer (page 36) per donor animal and incubated for 5 min on ice. Cells were washed using PBS and counted subsequent to centrifugation (300g, 5 min, 4°C). After another centrifugation

step, lineage depletion was carried out according to manufacturer's instructions (Lineage Cell Depletion Kit, mouse Miltenyi Biotec, Bergisch Gladbach, Germany, #130-090-858). Enriched lineage-negative cells were counted, seeded with $1x10^6$ cells per well in 6-well plates and incubated at 37°C for three nights in StemSpan + additives (page 36).

On day 4, cells were collected out of the 6-well plates. To do so, PBS was used to wash wells 2 – 3 times to ensure maximum yield and cells were counted again, expecting around three times the day 1 input. The desired number of cells was afterwards seeded in 1 – 1.5 mL StemSpan + additives, 8 µg/mL polybrene, and transduced with the chosen construct at the desired multiplicity of infection (MOI). In the first experiments MOI 50 was used, which was later changed to MOI 30. After addition of viral supernatant, cells were centrifuged for 1 h at 1000g at RT and incubated at 37°C for app. 5 hours. Thereafter 1.5 – 2 mL StemSpan medium + additives, but no polybrene, were added and cells were incubated overnight at 37°C, 5% CO_2.

Late on day 4, or early on day 5, female, ~8 week-old, wild-type C57Bl/6 recipient animals were lethally irradiated with 9.5 Gy whole body irradiation (Cs^{138} source, Biobeam2000, Eckert & Ziegler BEBIG, Berlin, Germany). Transduced cells were harvested from the 6-well plate(s) on day 5, again washing wells 2 – 3 times with PBS to ensure maximum yield and washed 2 times with PBS. Cells transduced with different constructs were counted, mixed in a ratio of 1:1 and diluted with PBS to compose one graft. In total, 400,000 cells in a total volume of 150 µL per mouse were transplanted via intravenous injection into the tail vein. Following transplantation mice received antibacterial treatment via drinking water for 4 – 6 weeks – 800 µL Baytril per 100 mL H_2O in a light protected bottle (Baytril 2.5% oral solution, Bayer, Leverkusen, Germany).

7.9.2. Blood samples

Starting around 6 weeks after transplantation, monthly blood samples were taken by submandibular bleeding. Only small volumes of blood were taken to prevent induction of stress haematopoiesis. Therefore, 2 – 3 blood drops were collected into EDTA coated 1.5 mL tubes (GK 150 EDTA 200 µL, 077001, Kabe Labortechnik, Nümbrecht-Elsenroth, Germany) and chilled on ice. Preparation of blood for flow cytometry and/or DNA extraction is described in section 7.9.3. Genomic DNA was extracted using the QIAamp DNA Blood Mini Kit according to manufacturers instructions. At chosen time points, typically around day 40 and day 200 after transplantation, blood samples were additionally analysed for fluorescent protein expression as well as chimaerism using a CantoII flow cytometer.

7.9.3. Final analysis

Mice were sacrificed by terminal retro-bulbar bleeding (after sedation with 95% CO_2 + 5% O_2) and the blood collected into EDTA coated 1.5 mL tubes (GK 150 EDTA 200 µL, 077001, Kabe Labortechnik, Nümbrecht-Elsenroth, Germany). *Tibia*, *Femur* and *Ilium* were removed, cleansed from flesh and stored in DMEM medium at 4°C. The spleen was removed, weighted, and stored in DMEM medium at 4°C.

Preparation of bone marrow cells and subsequent ery-lysis were done as described in section 7.9.1, page 43. Cells were counted afterwards and split into aliquots for flow cytometry, DNA extraction (~$8x10^6$ cells) and lineage depletion (described earlier on page 43f). Lineage positive cells were eluted from the column without a magnetic field. Approx. $8x10^6$ lineage -positive cells were used for DNA extraction, the rest were frozen as pellet after aspiration of PBS.

The spleen was dissolved by gently mashing it through a 70-µM cell strainer with a disposable syringe stamp and subsequently washed with PBS. Ery-lysis was performed similar to bone marrow, and cells were counted after resuspension in PBS. Spleen cells were split into each one aliquots for FACS and DNA extraction (~$8x10^6$ cells), while the rest of the cells (or $2x10^7$) were used for subset staining and sorting.

Peripheral blood was mixed with ~2 mL ery-lysis buffer in two 1.5-mL reaction tubes. Centrifugation (7 min 300g RT) was performed after 5 min incubation at RT. If required, ery-lysis was repeated a second time. The remaining white blood cells were afterwards resuspended into 800 µL PBS in a clean 1.5-mL tube.

Details about the antibodies used are listed in section 7.4, page 37. The basic final analysis consisted of the following staining panels and was recorded on a BD FACSCantoII (BD Biosciences).

Table 9 – Flow-cytometric staining panels for the final analysis

Sample	Bone marrow	
#1	Unstained control	
#2	CD45.1	CD45.2
#3	ScaI	cKit
#4	Lineage positive unst.	
#5	Lineage negative unst.	

Spleen		
Unstained control		
CD45.1		CD45.2
B220	CD3e	(Ly.6G)*
		Gr-1*

Blood	
Unstained control	
CD45.1	CD45.2

*explanation for Gr-1 or Ly.6G is given in section 7.4, page 37.

In addition to antibody staining, all samples were analysed for GFP and eBFP2 fluorescence (FITC and Pacific Blue channel). If only few unstained lineage-negative cells were available, no FC analysis of this population was performed to save cells for the DNA extraction.

Sorting of T cells ($CD3e^+$), B cells ($B220^+$) and Granulocyte ($Gr\text{-}1^+$, later $Ly.6G^+$) was done either on a BD FACS AriaFusion or a BD FACS AriaIIIu (both BD Biosciences). Sorted cells were collected into FACS tubes, filled with 1 mL PBS. Depending on the initial amount of stained cells and animal, varying numbers of sorted cells ranging from few thousand cells ($Ly.6G^+$) to millions ($CD3e^+$ or $B220^+$) were obtained. Sorted cells were pelleted at 300g, 4°C, 7 min and the PBS was aspirated until only a small volume (<1 mL) remained. The (non-visible) pellet was resuspended and pipetted into a clean 1.5- mL tube for DNA extraction, described earlier (section 7.1.2, page 37).

7.9.4. Secondary transplantation experiments

Four animals of each primary cohort were chosen for a secondary transplantation with three recipients each. Lineage-negative cells of the primary animals were purified as described (section 7.9.3, page 44). 200,000 to 250,000 lineage-negative cells of the primary animals in a total volume of 150 µL were

transplanted via intravenous injection into the tail vein of lethally irradiated ~8-week old female, wildtype C57Bl/6 (9.5 Gy, whole body irradiation) secondary recipients. Following transplantation, mice received antibacterial treatment via drinking water for 4 weeks – 800 µL Baytril per 100 mL H_2O in a light-protected bottle (Baytril 2.5% oral solution, Bayer, Leverkusen, Germany).

7.10. Next-generation sequencing

Barcodes were amplified from 200 ng (or 23 µL) gDNA of the mouse samples chosen for NGS with the following reaction mix and PCR program in 0.2-mL 8-strip tubes.

Table 10 – PCR mix and program for barcode amplification and multiplexing

[µL]	
23 µL or 200 ng	DNA
25	Multiplex PCR Plus MM (Qiagen, #206152)
1	MPLX-primer (10µM)
1	DUAL-primer (10µM)
0.2	Illu_P2 (43) (1µM) [Bridging Oligo]
Fill to 50 µL	H_2O

Time [s]	Temp. [°C]
300	95
30	95
30	57
30	72
Go to step 2	Rep 29
600	68

Individual MPLX- (Illu_MPLX35 to Illu_MPLX52) and DUAL- (ILL_Dual-P5-01 to ILL_Dual-P5-04) primer (see setion 6.4) combinations were used to create a unique identifier for each sample to enable multiplexing on the NGS flow cell. The 18 MPLX primers were combined with the first DUAL primer to label the first 18 samples. For next 18 samples the MPLX primers were reused, but combined with the second DUAL primer. This pattern was repeated until the desired 57-58 samples per flow cell were uniquely marked. Samples were purified after the PCR using Agencourt AMPure XP beads (Beckman Coulter, Brea, USA, #A63880). The 50-µL reaction was mixed with 90 µL XP beads in LoBind microcentrifuge tubes (Eppendorf, Hamburg, Germany, #022431021), mixed by pipetting and incubated for ~8 min. Tubes were put on a magnet plate and supernatant was removed when the beads were completely captured by the magnetic field. To wash the DNA, 200 µL 70% freshly mixed EtOH were added and beads resuspended by pipetting. After binding the beads to the magnet again, the supernatant was removed and the washing step repeated a second time. Remaining EtOH evaporated during a short incubation (5 – 10 min) of the tubes with an open lid. DNA was eluted from the beads using 15 µL TE elution buffer. Beads were bound to the magnet again and the supernatant, containing the target DNA, was pipetted into a fresh low-binding tube. A concentration measurement via Qubit dsDNA HS Assay was done with 1 µL eluted DNA to determine the yield of the purified sample, typically 0.3 – 3 ng/µL. In the next step 57 or 58 reactions had to be combined into one flow cell preparation, containing a total of 300 ng DNA in 15 µL. Thus, 5.26 ng, or the remaining 14 µL, of each purified sample were pooled into one reaction tube. The 1.8x volume of XP beads was added, mixed and incubated for ~8 min. The reaction tube was transferred onto the magnet plate and supernatant discarded after the beads were captured by the magnetic field. Two washing steps with 1 mL 70% EtOH were carried out afterwards, as described above. The sample was dried with an open lid for some minutes before resuspension in 15 µL TE buffer and elution for 10 min. Beads were

removed via magnet and the TE buffer, containing the eluted DNA, pipetted into a new low-binding tube. Concentration was determined via Qubit dsDNA HS assay. Samples were sent to the Deep Sequencing Group SFB 655 at the Technische Universität Dresden, Dresden Germany, where the quality of the sample preparation was confirmed via Fragment Analyzer (Advanced Analytical Technologies, Ankeny, USA) prior to sequencing via NGS with single end reads of 83bps length on an Illumina MiSeq System, with 20% PhiX.

7.11. Bioinformatic processing

Initial bioinformatic processing, filtering and quality control of the sequencing reads were done by our collaboration partners at the Technische Universität Dresden as described (Thielecke et al., 2017). Error correction threshold was set to a Hamming Distance of HD = 8. Thus, barcodes with up to eight nucleotide exchanges were considered to be derived from one ancestor barcode and summed up. In addition, we omitted all barcodes with frequencies <0.5% of backbone reads due to biological irrelevance. Reads of samples from the same animal were combined in one data table, listing the nucleotide sequence in one column and the frequency of this specific barcode for the different samples in the following columns. Additional files contained total read counts per sample, total number of barcodes per sample as well as the average Hamming Distance of barcodes within one sample. In-depth analysis and graphical display of this initial data set were done by myself using Microsoft Excel or customized R scripts with the implementation of available plugins (e.g. venneuler, ggplot2, colourbrewer).

7.12. Digital droplet PCR

We used digital droplet PCR (ddPCR) to quantify obtained NGS frequencies of selected barcodes. A master mix was prepared as listed in Table 11. Primer set A with a FAM-labelled probe was specific for the promoter region of each construct, while the forward primer and HEX-labelled probe of set B were located on the forward Illumina NGS adapter (Figure 12). The reverse primer of primer set B was barcode-specific. Hence, it had to be individually designed for each barcode and was separately added to the reaction mix.

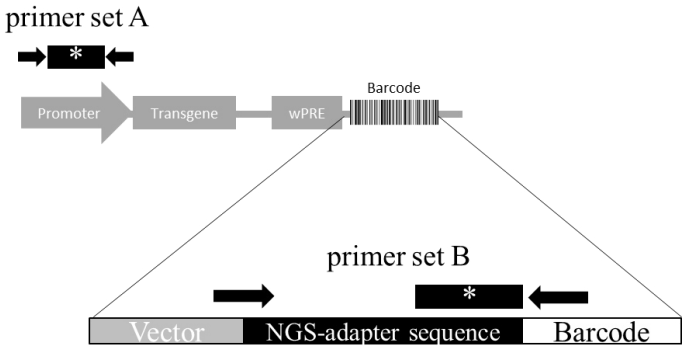


Figure 12 – Schematic localisation of primer sets A and B used for ddPCR. Primers/Probe of set A were specific for the promoter regions of the different constructs used. Forward primer and probe of set B were located on the forward Illumina adapter, while the reverse primer was positioned within the variable BC region and thus specific for one individual barcode.

The 5' end of probe A was labelled with FAM, while probe B carried a HEX fluorophore. Both probes 3' ends were conjugated to a Black Hole Quencher 1. Arrows: primers, box with *: probe

Design of the reverse primers used a general starting position at the third variable nucleotide and extended the primer sequence until an annealing temperature of ~59°C was reached, resulting in primer lengths of 17 - 23 nucleotides. The reaction mix was prepared as outlined in Table 11.

Table 11 – Mastermix and program for digital droplet PCR

[µL]	
20 – 50 ng	DNA
12.5	ddPCR Supermix for Probes (Bio-Rad, #186-3010)
0.5	Fast digest *EcoR*I
0.07	$MgCl_2$ [1 M]
f.c 900 nM	Primer set A and Fw Primer B
f.c. 250 nM	Probes A and B
Fill to 25 µL	H_2O

Time [s]	Temp. [°C]
600	95
30	94
120	60
Go to step 2	Rep 39
600	98

Genomic DNA in the reaction was digested for 15 min at 37°C and droplets generated following manufacturer's instructions in a QX100 droplet generator using droplet generation oil (1863005), gaskets (1863009), cartridges (1864008, all from Bio-Rad) as well as 96-well PCR plates (951020389, Eppendorf, Germany). Plates were sealed and the PCR run with the thermal profile indicated above. Droplets were readout on a QX100 Reader (Bio-Rad) after manufacturer's instructions and analysed using the provided software (Quantasoft). Primer and probe sequences are listed in Table 3, page 33.

7.13. Calculating the number of clones contributing to haematopoiesis

The generated dataset was used to calculate the number of engraftment capable cells (ECCs), contributing to haematopoiesis at various time points after transplantation. The general scheme for the calculation is outlined in Figure 13. As we did not see differences in clonal dynamics between vector constructs, data from all groups and animals was pooled within this analysis and means were used for the calculation. The calculation itself can be found in the appendix, page 108. Combining transduction rates measured by FC with the frequencies of LSK $CD150^+$ cells (long-term engraftment capability) allows calculation of the expected number of barcoded ECCs within the graft. Comparison of the number of recovered barcodes over the whole observation period with this theoretical value shows, that around 32% of expected ECCs engraft and are detectable. The number of recovered barcodes at a certain time point can then be set in relation to that value, to calculate the number of active, marked, ECCs. Extrapolating these values to the non-marked ECCs and adding up both populations allows assessing the number of active ECCs at a certain time point.

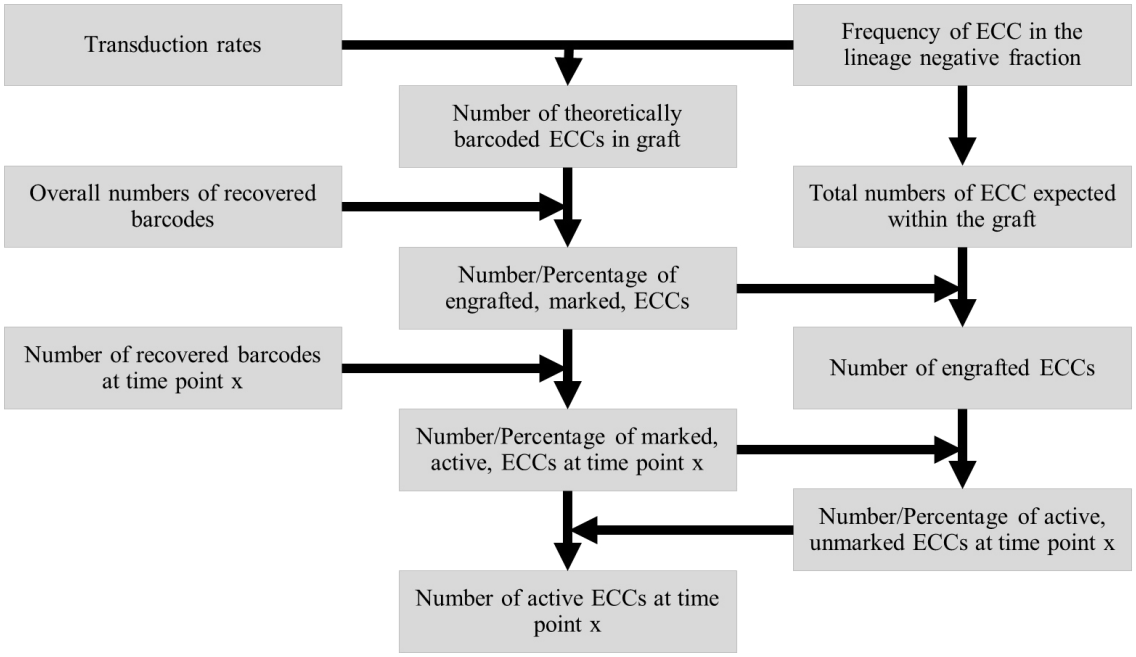


Figure 13 – Scheme to calculate the number of hematopoietically active cells. Details are given in the text, the calculation can be found in the appendix (Appendix figure 6, page 108)

7.14. Experimental Setup

The experimental setup of this work is shown in Figure 14. Up to three different, barcoded lentiviral vectors were used to independently transduce 8-week old male, lineage-negative, CD45.1 donor cells. These transduced cells were mixed to compose one graft, and transplanted into 8 week-old, lethally irradiated, female wild type recipients. Starting six weeks after transplantation, monthly blood samples were taken. After 8-12 months, animals were sacrificed and DNA was extracted from haematopoietic organs. T cell, B cell and granulocyte subpopulations were sorted by FACS. Lineage-negative cells of chosen animals were further used for secondary transplantation into 8-week old lethally irradiated female wild-type recipients. Those secondary transplants were observed for 6 – 8 months and the described final analysis was done. Chosen samples were then further analysed via NGS.

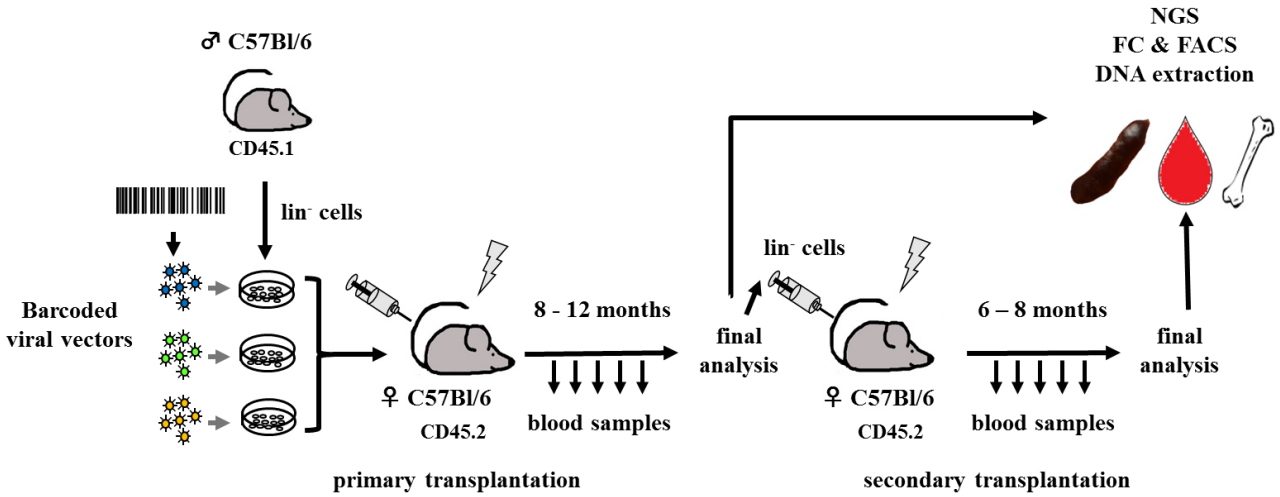


Figure 14 – Overview of the experimental workflow. Barcoded, alpha- or lentiviral, vector constructs were used to independently transduce lineage-negative cells from ~8-week old male, CD45.1 donor mice. Transduced cells were mixed to compose one graft, which was transplanted into lethally irradiated ~8 week-old female, wild-type, recipients. Starting around 6 weeks post transplantation, monthly blood samples were taken to enable monitoring of the reconstitution of the haematopoietic system in a time dependent context. After 8 to 12 months, the final analysis was done for blood, spleen and bone marrow. DNA was extracted from these organs, lineage-negative and -positive cells were purified and subpopulations from B cells, T cells as well as granulocytes sorted via FACS. Lineage-negative cells from chosen primary animals were used to transplant a lethally irradiated cohort of secondary recipients. Those animals were blood sampled for about 6–8 months, again followed by the described final analysis. Chosen samples were then further analysed via NGS.

8. Results

8.1. Production of barcoded plasmid libraries and viral particles

The aim of this thesis was to compare the influence of the internal promoter on a clonal level of haematopoietic reconstitution to determine the most "neutral" marking vector. To do so, I used various 3rd-generation alpha- and lentiviral self-inactivating (SIN) vector constructs in a competitive *in-vivo* transplantation model (Figure 14). Two alpharetroviral and three lentiviral vector constructs, whose fluorescent proteins (FP) were driven by various promoters with different potency, were generated (Figure 15). The first lentiviral construct was equipped with an enhanced blue fluorescent protein 2 (BFP) and a strong spleen focus-forming virus (SFFV) promoter, driving strong and broad transgene expression in the haematopoietic compartment (Baum et al., 1995; Schambach et al., 2006; Weber et al., 2008). A weaker promoter is represented by the intron-deleted version of the human elongation factor-1 alpha (EF-1 α short, EFS) promoter, which, is sufficient for the expression of clinically relevant genes (Schambach et al., 2006; Zychlinski et al., 2008). This promoter was used in both vector systems to drive the expression of an enhanced green fluorescent protein (eGFP, short GPF). Finally, we created an alpharetroviral as well as a lentiviral promoter-deprived (pd-) vector construct. It has already been shown that marking of the haematopoietic system with a pd-gammaretroviral vector did not lead to induced clonal imbalance in a serial bone marrow transplantation model (Cornils et al.,

2009). Although expression was not to be expected, pd-alpha- and pd-lentiviral constructs were equipped with TSapphire (TSapp) and Venus fluorescent proteins to roughly match the size of their promoter-carrying counterparts. All fluorescent proteins used in this study were chosen to be GFP derivatives and differ only in ~4.1% of their ~720 bases. It has been demonstrated that high GFP levels can cause toxicity due to inhibited polyubiquitination, (Baens et al., 2006). Still, GFP is better suited for tracing of haematopoietic cell compared to DsRed (Tao et al., 2007). Using GFP derivates in all vector constructs at least should equalize these possible toxic effects. Finally, all constructs were equipped with the already described barcode (BC) near the 3'LTR. We used the four different barcode backbones (introduced on page 30) to link each fluorescent protein to one specific barcode backbone.

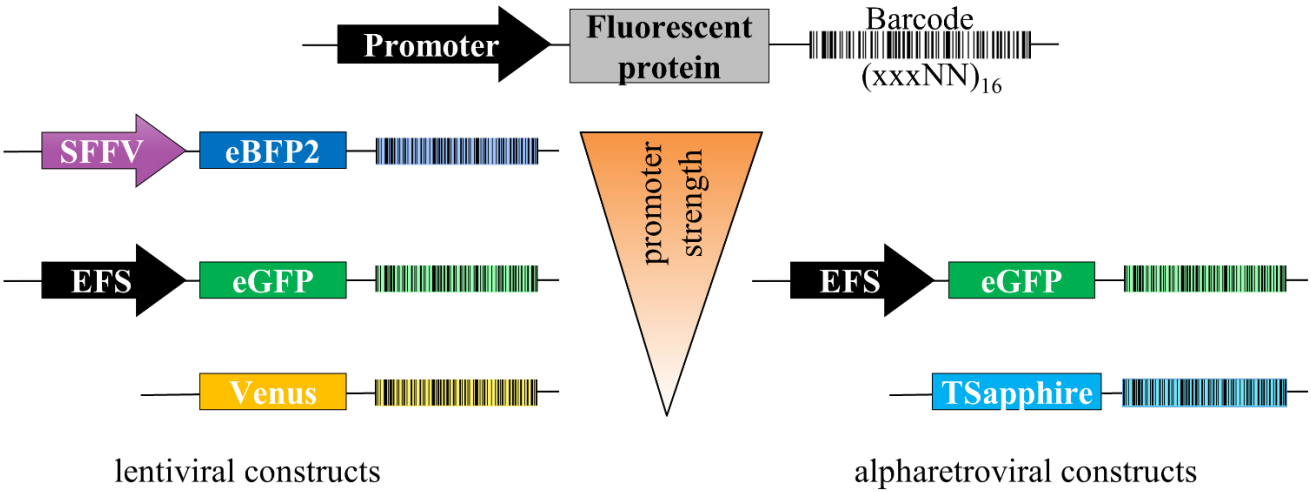


Figure 15 – Schematic representation of the five different vector constructs used in this work. The general structure, shown above in dark colours, consists of an internal promoter driving a fluorescent protein (FP). The genetic barcode is located near the 3'LTR. Three lentiviral constructs (left side) and two alpharetroviral vectors (right side) with different promoters, FPs and barcode backbones were cloned.

It is important to keep the number of barcodes (so-called complexity) during the whole experiment as large as possible and technically feasible to ensure proper error correction and barcode calling at the end. Yet, there are several limiting aspects during the production process (see section 7.6, page 40). Figure 16 summarizes the approximate number of possible barcodes after each step. Although, the BC32 system offers a theoretical complexity of ~10^{19} different combinations, it is impossible to start with every single one. A 144 bp ssDNA barcode oligo has a molecular weight of approx. 47 kDa and 1 ng contains roughly 1.2x10^{10} molecules. Hence, to cover all possible strands at once one would at least need 0.83 g of DNA. Accordingly, the 700 ng ssDNA starting material contained only a random selection of around 10^{12} barcodes. During the next steps, the complexity decreases due to purification procedures, where material is lost. The complexity after electrotransformation can be estimated by diluting colonies as mentioned in the method section. Counting those colonies typically indicates an order of 10^{6} to 10^{8} barcodes in the subsequent plasmid preparation. Analysis of the 16 wobble precursor barcode showed no declining complexity during adjacent virus production (Selich et al., 2016, supplementary figure 1). Still, it seems possible that a small fraction of individual barcodes creates secondary structures, sterically hampering packaging into viral particles. Later on, the order of

>10^9 viral particles are harvested, which should be sufficient for each barcode to be packaged. Transplantation of 400,000 lineage-negative cells with a transduction rate of ~20% should result in ~80,000 initial barcodes per animal. It is important to note that only a small amount of lineage-negative cells are true long-term HSCs and able to engraft. For NGS analysis, we can only use 200 ng of a sample (~33,000 genomes) for barcode recovery. Typically, the order of 10^1-10^2 biologically relevant barcodes per sample is recovered. Those ~100 barcodes are only a small part of the whole picture, given that ~80% of the initial cells were not transduced. Nevertheless, they are sufficient to draw conclusions and apply them to the whole system if we can ensure proper assignment and/or deletion of false positive barcodes. These false-positive barcodes are primarily generated by PCR errors during amplification or sequencing. To correct those errors, the previously introduced Hamming distance is used (see page 28). A computational model (Thielecke et al., 2017, Figure S1) suggests, randomly picking 1000 out of all possible BC32s should yield an average Hamming distance over 15, allowing the correction of seven PCR-induced errors.

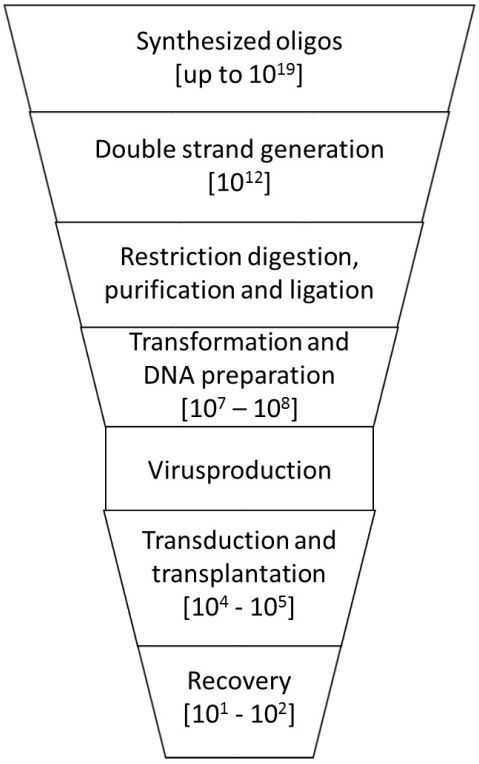


Figure 16 - Decline of barcode numbers during the process. An estimation of the number of barcodes at each step is given inside the brackets if possible. ~830 mg of synthesized oligos could, given a non-random distribution, contain all 10^{19} potential barcodes considering all possible combinations. For double-strand generation, 700 ng of oligos were used, representing barcodes in the magnitude of 10^{12}. During subsequent restriction, purification and ligation barcodes are lost. The next possibility, to estimate complexity is the transformation. At that step, $10^6 – 10^8$ colonies (= barcodes) can be detected by counting colonies on different plates. Virus production should not influence this number, except there are secondary structures interfering viral packaging. Transplanting 400,000 cells per animal with a transduction rate of ~20% would result in around 80,000 initial barcodes per animal. Not all cells are able to engraft long-term, as lineage-negative cells also contain many precursors. For NGS analysis, we can only amplify barcodes form 200 ng DNA at the end (~33.000 genomes). Typically, the order of 10^1-10^2 biologically relevant barcodes were recovered per sample.

Library complexity as well as the titers determined for the different constructs are shown in Table 12. Promoter-deprived constructs, whose titre could not be analysed via FP expression were produced in parallel with other constructs and the titre was equalized.

Table 12 – Complexity of barcoded plasmid libraries and titre of virus supernatants produced

Construct	Barcode backbone (see page 33)	Estimated complexity of the plasmid library [number of barcodes]	Titre [infectious particles/mL]
LeGO-pd-V-BC	Venus	3.6×10^6	1.6×10^8
LeGO-EFS-GFP-BC	GFP	2.1×10^6*	1.6×10^8
LeGO-SFFV-eBFP2-BC	Cherry	1.5×10^6	2.5×10^8
Alpha-pd-S-BC	Cerulean	7.3×10^5	4.2×10^8
Alpha-EFS-GFP-BC	GFP	4.3×10^6	4.2×10^8

*pooled from 2 different library preps with lower individual complexities

Other barcoded libraries at that time point (2013) contained either some thousand arrayed (Schepers et al., 2008), or a few hundred (Gerrits et al., 2010; Verovskaya et al., 2013) to around 10^5 (Lu et al., 2011, Supplement Table 2) different barcodes. With our protocol for the BC16, we were able to generate plasmid libraries with around 5×10^5 barcodes (Cornils et al., 2014). Using the new BC32 system, we were able to more than double the complexity of our plasmid libraries in most cases, at the same time significantly increases the Hamming distance between our barcodes.

8.1.1. Further optimisation of the BC32 constructs

During the course of this project, I further optimised our barcode constructs. As shown in the introduction (page 29), the first generation of BC32, used in this work, was inserted into the vector via *Xba*I/*Xho*I restriction sites. Unfortunately, some of the defined nucleotide triplets chosen were able to generate palindromic *Xba*I or *Xho*I recognition sequences with specific wobble base combinations. As consequence, a certain percentage of barcodes was digested within the variable sequence, generating truncated barcodes. These short barcodes were partly excluded, during agarose gel purification as only a small piece of gel around the desired size was further processed. However, some amount of the short sequences were incorporated into our plasmid preps. To prevent this, the second version of our BC32 system is inserted via *MreI/MauBI* into an altered multiple-cloning site. These two enzymes are unable to cut within the barcode sequence and in addition less commonly used in other vector sequences, e.g. within a multiple-cloning site, making adaption of the BC32 system, e.g. for Sleeping-Beauty transposons, easier. In addition, I did some optimisation and tweaking of the barcode production and cloning protocols. With all these changes, our *MreI/MauBI* BC32 plasmid libraries nowadays reach complexities of up to 10^8 barcodes – around 2 log scales higher than the initial libraries used in this work.

8.2. Competitive *in vivo* setup and engraftment of primary recipients

To compare the five different vector constructs, I set up four competitive *in-vivo* groups, 12 mice each. Each group received transduced cells from two or three different constructs (Figure 17). Mouse Group 1 (MG1) received cells either harbouring the lentiviral EFS-GFP or promoter-deprived (pd)-Venus construct. The transplant of Mouse Group 2 (MG2) contained cells either transduced with one of the two already mentioned vectors or the third, lentiviral SFFV-BFP construct. Mouse Group 3 (MG3) received cells transduced with the two alpharetroviral vectors, EFS-GFP or pd-TSapphire. The graft of the last mouse group 4 (MG4), was mixed from three independent transductions with both alpharetroviral vectors mentioned and the lentiviral SFFV-BFP control construct. Of note: group 3 and 4 were done in parallel with the same transduced grafts, whereas group 2 started around three months after group 1 due to logistical reasons. Group 4 was repeated with 9 animals, labelled Mouse Group 4.2, due to high initial mortality (see page 56).

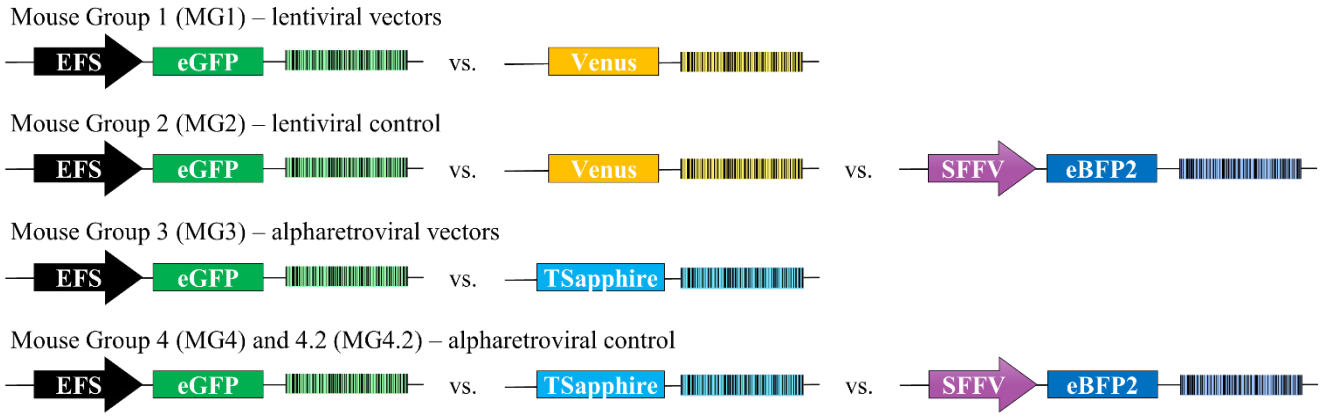


Figure 17 – Overview of the four competitive *in-vivo* mouse groups (MG). Group 1 received lentiviral EFS-GFP and pd-Venus barcoded cells. Group 2 received cells transduced with one of the mentioned lentiviral vectors or an SFFV-BFP construct. Group 3 got cells barcoded with alpharetroviral EFS-GFP or pd-TSapphire constructs. The transplant of control group 4 contained cells transduced with either one of the alpharetroviral vectors or the lentiviral SFFV-BFP construct. Mouse group 4.2 is a repetition of mouse group 4 with the same constructs used.

Transduction rates of the promoter carrying, lentiviral and alpharetroviral EFS-GFP and lentiviral SFFV-BFP, constructs were determined via FC 72h post transduction. For barcoding experiments, a single integration per cell is desired to avoid cells harbouring multiple different barcodes. At the same time, maximum number of cells should be transduced to ensure sufficient marking of the graft. Consequently, the optimal transduction rate would be around 20% transduced cells (Fehse et al., 2004). As Figure 18 shows, transduction rates for the GFP vectors were around the targeted 15-20%. In contrast, transduction rates for the BFP vectors were around 2-3 times that value, despite the same multiplicity of infection (MOI), the number of vector particles per cell ratio. This might be explained by a weaker GFP expression of the EFS promoter, not sufficient to lift cells into the GFP-positive gate. However, in our experience, the discrimination of FP transduced cells with the EFS promoter normally looks fine, although the separation in the FC plot shows a lower intensity and is not as clear-cut as for

the SFFV. The second explanation would be that the determined titre and, in consequence, calculated MOI were inaccurate. As titration was done the same way for all three vector preparations, and GFP values of two constructs were consistent, this explanation seems unlikely. Interestingly, when we lowered the MOI for MG4.2 from 50 to 30 to avoid too high transduction, transduction rates decreased around the same percentage for both constructs (22% for GFP and 28% for BFP), and SFFV-BFP still seemed to transduce the twofold amount of cells. There is one report of the same phenomenon with alpharetroviral SFFV- and EFS-carrying vectors transducing induced pluripotent stem cells with the same MOI, where the SFFV construct was found with a 35% higher vector copy number afterwards (Lin et al., 2015). Unfortunately, we were not able to determine vector copy numbers in our transduced cells, as we could not keep the lineage-negative cells in culture long-term. PCR analysis of DNA extracted three days after transplantation was skewed by remaining plasmid and vector fragments. Given these transduction rates, some cells will most likely harbour more than one barcode.

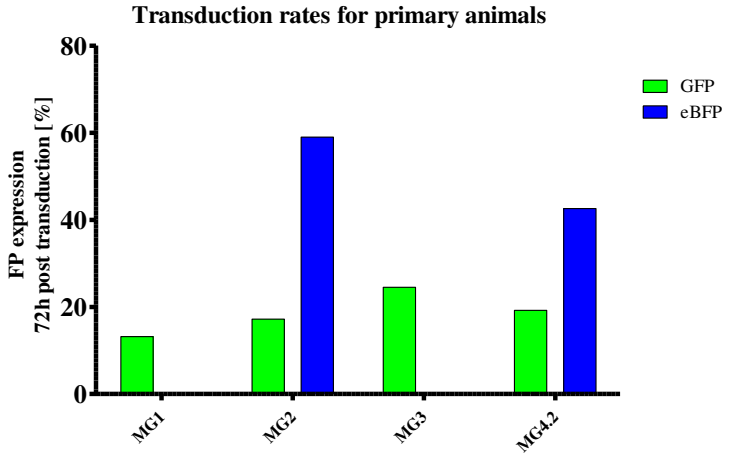


Figure 18 – Transduction rates for mouse groups used in this work. FP expression of cells transduced with the lentiviral SFFV-BFP and lenti- or alpharetroviral EFS-GFP carrying constructs was measured by FC 72h post transduction. GFP transduction was around the targeted 20%, but BFP transduction was much higher, even though the same multiplicities of infection (MOIs) were used. Groups 1-3 were transduced with MOI 50; the latter was decreased to 30 in MG4.2.

The survival curves for the different groups (Figure 19) show some differences. All animals from MG1 were alive up to day 280 post transplantation (tx). In contrast, I had to sacrifice 8 out of 12 animals from MG4 before day 81 due to worsening health conditions. It is important to note, that there were no animals with enlarged spleens or thymi, which could have indicate leukaemia development. The bone marrow of *Femur*, *Tibia* and *Ilium* in those eight animals often had a white to rose colour, compared to a more cherry-red colour in healthy animals. This, in combination with thin blood, points towards graft failure. Considering that short term HSC and progenitors are able to provide blood production for the first 8 – 12 weeks and afterwards long-term HSC are needed (Christensen and Weissman, 2001; Spangrude et al., 1988) it is likely, that LT-HSCs were either missing or unable to supply the necessary progeny in the respective animals. Interestingly, the recipients of MG3 and 4 were siblings, randomly assigned to the different groups. The groups were done in parallel and all received the same donor cells. The main difference were the additional, SFFV-BFP vector transduced cells only administered to MG4, thus using a lower number of cells transduced with the other

constructs to keep the overall number of cells constant. Therefore, survival curves of MG2 and 4 could indicate some negative impact of the lentiviral SFFV-BFP construct. Due to the high mortality rates in MG4 within the first 100 days post tx, I replicated this setup with 9 additional animals (termed MG4.2). Interestingly, animals in MG4.2 did not show any graft failures in the first months. The final analysis of MG1 was done at day 345 post transplantation, while all other primary groups were analysed around day 250 post tx to ensure sufficient animals alive for secondary transplantation. One animal in MG4.2 developed leukaemia indicated by an enlarged spleen and big lymph nodes. Analysis of the leukaemic cells revealed the CD45.2 (wild-type) surface marker. Since our transplanted cells had the CD45.1 point mutation, the observed leukaemia was recipient-derived, possibly induced by irradiation.

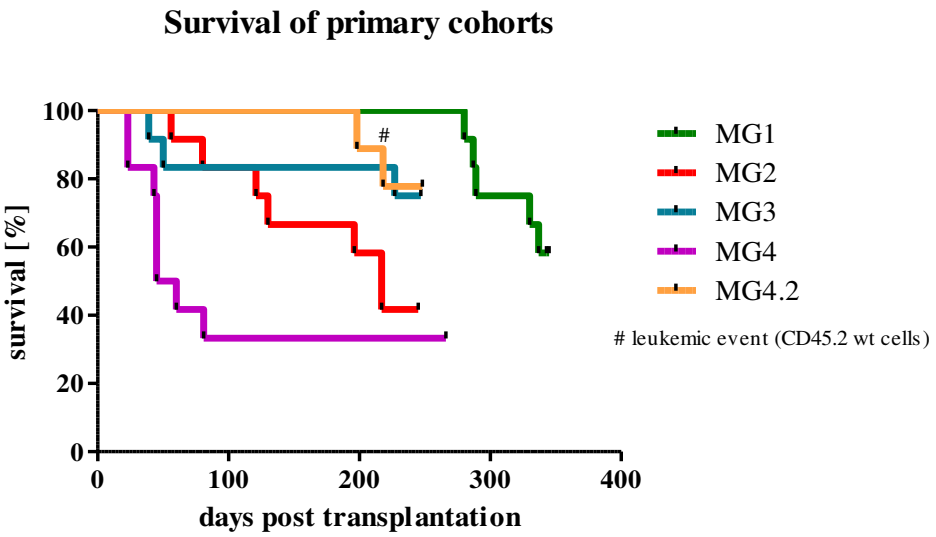


Figure 19 – Kaplan–Meier plot of the five primary cohorts over the observation period. Contrary to MG1, where all animals were still alive at day 280 post transplantation, only 7 animals of MG2 (same vectors as MG1 plus an additional SFFV-BFP) made it to day 200 post transplantation. In MG3, two early graft failures occurred, while all other animals were stable afterwards. In contrast, I had to sacrifice eight of the 12 animals from MG4, siblings of MG3, transduced at the same date with the same transduced cells and only receiving one additional construct, before day 100 due to worse conditions. For compensation, we set up an additional group (MG4.2) with the same constructs as MG4, which showed no signs of graft failures.

Around 6 weeks after tx (day 37 to 47) the first blood samples were taken and chimaerism analysed via FC. Mean chimaerism of all groups at that time point was 78.7% (range 72.2 – 85%). The next chimaerism analysis was done around day 200 after tx, where a mean chimaerism of 92.7% in all surviving animals (range 91.8-93.6%) had been established. In most animals, that level of chimaerism was maintained until the final analysis as exemplified in Figure 20. This data shows that we obtained stable, long-term engraftments in our animals.

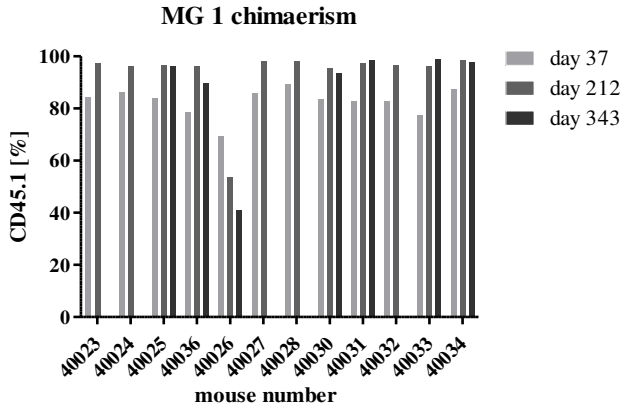


Figure 20 – Chimaerism of mouse group 1 (MG1) determined by FC analysis of blood samples at selected time points. Absence of bars indicates animals no longer in the experiment. The data for the other primary groups look similar, with at least 90% chimaerism for most animals at the later time points.

In addition to the chimaerism, FP expression for the promoter-carrying constructs was determined in the peripheral blood samples. As shown in Figure 21, most animals had decreasing levels of fluorescence protein expression over time.

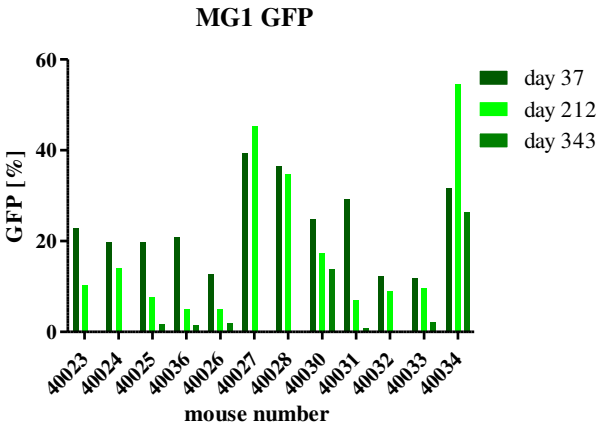
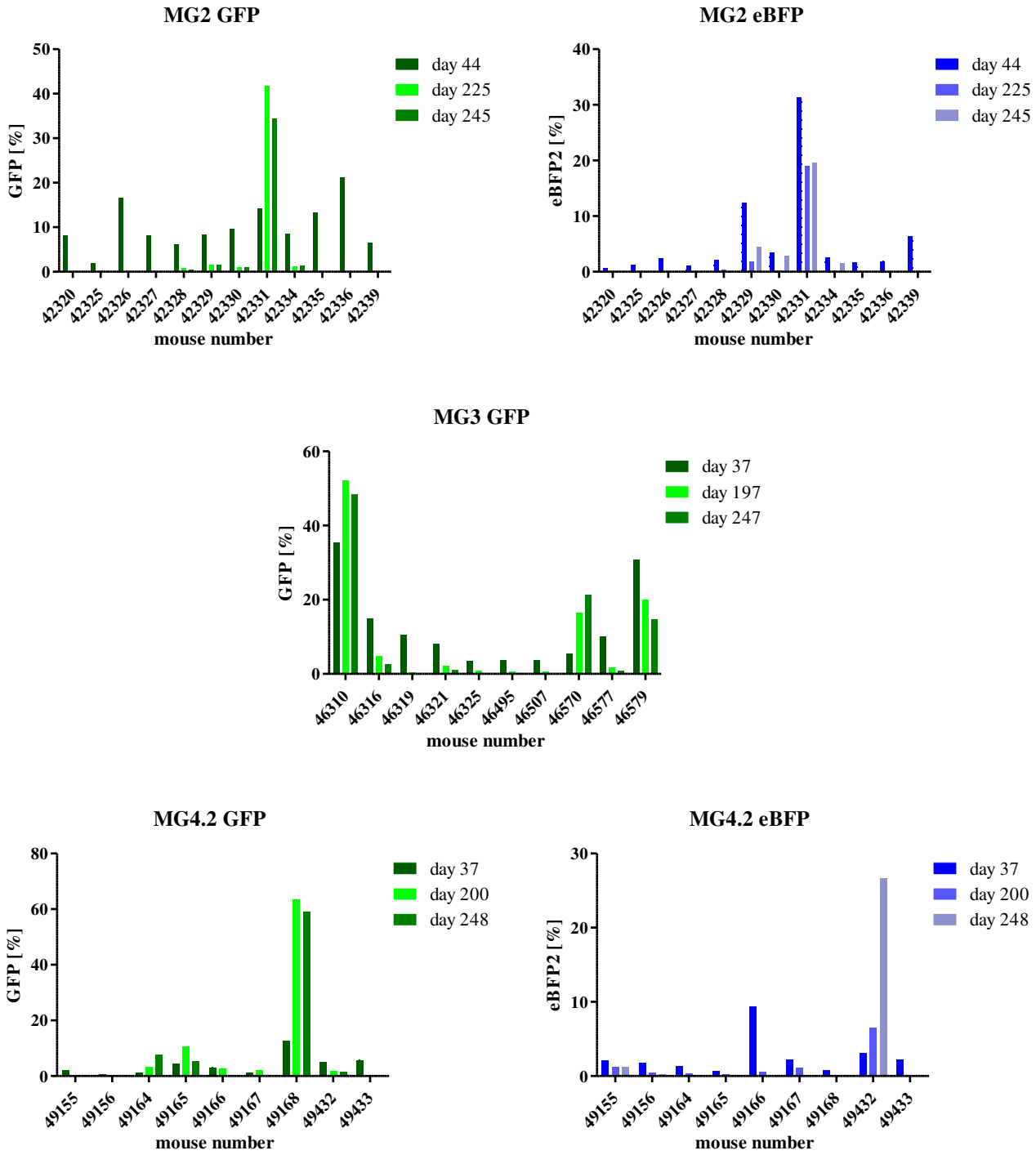


Figure 21 –FP expression over time in blood samples from MG1 (above) and all other mouse groups (see below). The percentage of cells expressing FP are shown for up to three blood samples per animal. Absence of bars indicates animals no longer in the experiment. Around 6 weeks after transplantation, the initial expression levels for GFP were determined. In most animals these levels decreased over time until the percentage of FP-positive cells was below 5%. Some animals showed the contrary, with over 30% FP expressing cells at the final analysis. Despite of higher transduction rates in MG2, BFP expression initially is below the GFP level. Possible explanations are given in the text. The difference disappeared until final analysis. It is unclear, why MG4.2s initial percentage of GFP expressing cells were much lower than the previous groups despite similar transduction rates.



In summary, FACS analysis revealed stable, long-term engraftment of marked cells with high chimaerism levels. The percentages of FP expressing cells varied between animals but expression was detectable over the whole observation period. Furthermore, we could detect both expected FP in animals of the respective group, showing that (at least) two marked populations were able to engraft in our competitive model situation in parallel.

8.3. Engraftment of secondary recipients

Four of the primary animals per group were chosen for secondary transplantation into three recipients each. If possible, one animal showing high fluorescence expression, one animal with intermediate expression and two animals with low expression according to FACS analysis were selected. 200,000 – 250,000 lineage-negative cells of one selected donor animal were transplanted into three, lethally irradiated secondary recipients each. As with the primary cohorts, chimaerism was determined within the first blood sample around 6 weeks post transplantation. With a mean of ~46% chimaerism over all animals (exemplified in Figure 23), chimaerism, unsurprisingly, was below the primary groups at that time point. Even at the final analysis time point (day 147 to 220 post transplantation) chimaerism only reached ~63% overall. As Figure 22 shows, there were several animals that had to be taken out of the experiment within 100 days post transplantation, indicating insufficient numbers of engrafting intermediate- and/or long-term HSCs. In addition, eight secondary animals, distributed over all groups, showed signs of leukaemia (enlarged lymph nodes, giant spleens and/or thymi). However, FC analysis did show a CD45.2 (wild-type) genotype of the leukaemic cells in all animals. Therefore, the development of these leukaemias was unrelated to our vector constructs, but initiated by cells damaged by irradiation. It is possible, that the number of transplanted lineage-negative cells was too low to support proper blood production, in line with the observed graft failures. This may put some kind of compensation pressure upon the remaining, radiation damaged, recipient cells. Another explanation could be that these cells had already developed a preleukaemic state in the primary animals, which developed into full-blown leukaemia during the secondary engraftment. The origin of the leukaemic clones remains unknown, as primary and secondary recipient animals were female, wild-type animals and thus are indistinguishable. As secondary transplantations have been shown to promote progression of (transduced) preleukaemic clones (Li et al., 2002) one could argue for a primary cohort origin.

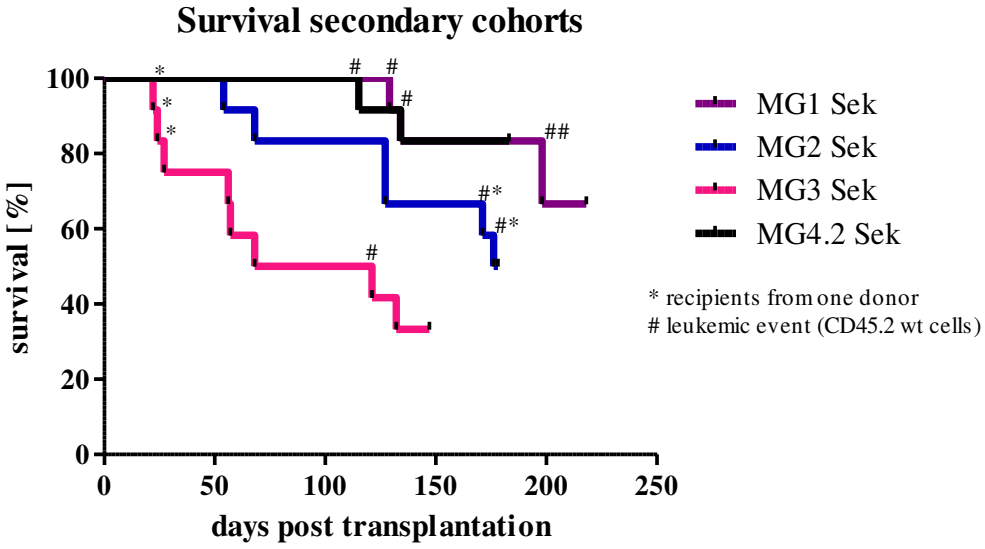


Figure 22 - Kaplan–Meier plot of the four secondary cohorts over the observation period. Several animals showed signs of graft failures, thin blood, white bones without marrow, and had to be taken out of the experiments within the first 100 days post transplantation. Several other animals (indicated with #) developed leukaemia's 100 – 200 days post transplantations. All leukaemias were CD45.2 positive, indicating wild-type origin.

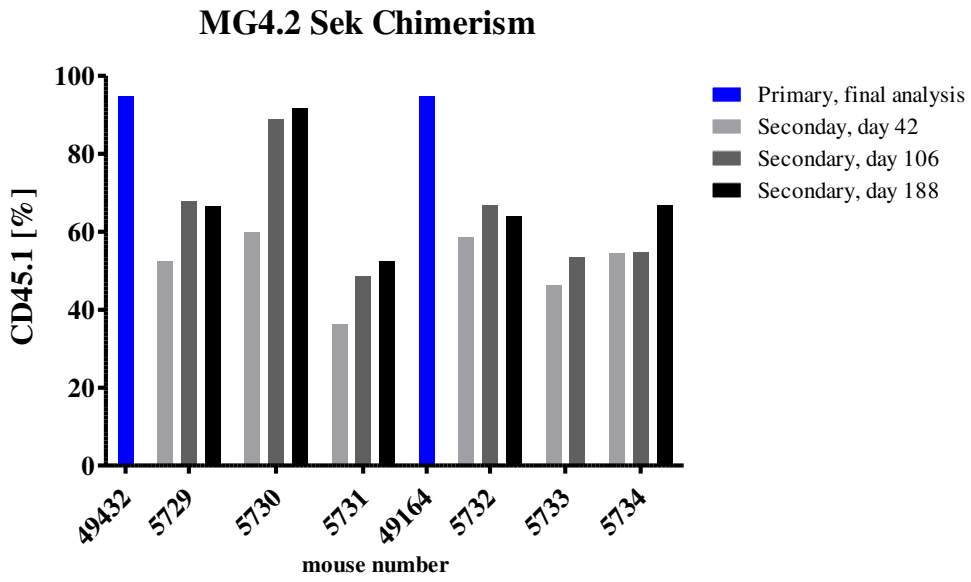


Figure 23 –Chimaerism of MG4.2 Sek, determined by FC analysis of selected blood samples (dark bars) compared to the primary donor (blue bars). Absence of bars indicates animals no longer in the experiment at that time point. Each primary animal (left, blue) was used to transplant the three secondary animals on its right side. Chimaerism analysis for the other secondary groups looks comparable.

FP positivity, within the secondary transplants, showed strong variances when compared to the primary donor for the two promoter-carrying constructs. As exemplified in Figure 24, there were secondary transplanted animals showing much higher numbers of FP-positive cells compared to their donor, while in others only a fraction of the initial FP content was found after engraftment. This effect, however, seems to be independent of the underlying vector construct, as there was no trend for any of the FPs to always be above/below the primary level. In general, recipients from the same donors often (but not always) showed similar trends of declining or rising levels of FP-positive cells, but absolute frequencies between the animals varied up to threefold.

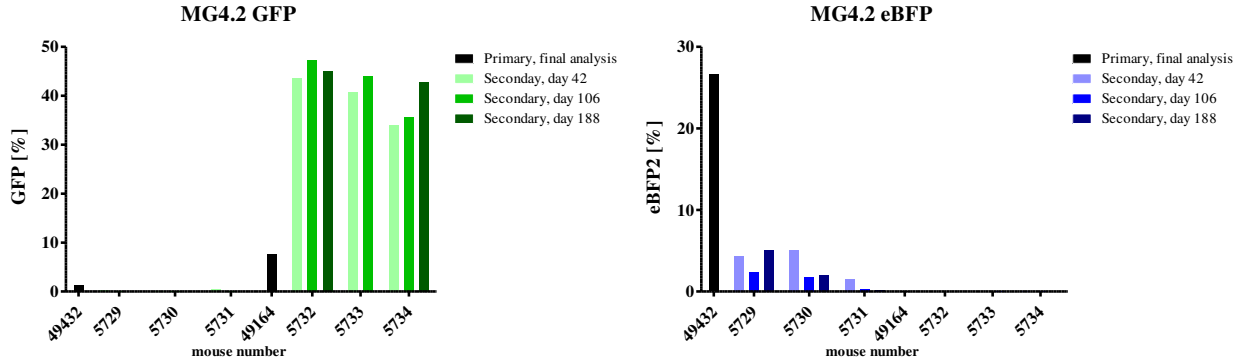


Figure 24 – Percentage of FP-positive cells of some secondary transplantations from MG4.2, determined by FC analysis of selected blood samples (coloured bars) compared to the primary donor (black bars). Absence of bars indicates animals no longer in the experiment at that time point or values below 0.4%. Each primary animal (black bars) was used to transplant the three secondary animals on its right side. While the 3 recipients of #49164 (left diagram) show higher GFP content after engraftment compared to donor, the opposite is true for the BFP content for the secondary recipients of #49432 (right diagram).

Due to the relatively poor survival rates, occurrence of leukaemic events and only moderate chimaerism, in most of the secondary transplants, we decided against sequencing those samples and instead focused on samples from the primary donors.

8.4. Barcode analysis via NGS in primary animals

I selected 176 samples for NGS, focussing on the 16 primary animals (4 per group) used for secondary transplantation. Samples selected contained genomic DNA from spleen, bone marrow, lineage-negative cells, an time course of blood samples and, if available, sorted subpopulations of T cells ($CD3e^+$), B cells ($B220^+$) and Granulocytes ($Ly.6G^+$ or $Gr\text{-}1^+$).

To prevent confusion and to clarify the upcoming part of this thesis, I want to reintroduce two important terms, briefly mentioned in the introduction, that sound similar but define different things:

When mentioning (barcode) backbone(s), I refer to the entirety of all barcodes associated to one specific vector construct. The structure shown below is used in the Alpha-pd-TSapphire construct, hence called TSapphire backbone.

NNNCAGNNATCNNCTTNNCGANNGGANNCTANNCTTNNCAGNNATCNNCTTNNCGANNGGANNCTANNCTTNNCAGNNATCNN

In contrast, the term barcode (BC) describes one discrete, defined sequence. For example, the barcode depicted below represents one defined TSapphire BC.

TCGCAGGCATCGCCTTGACGATGGGATCCTAGGCTTTTCAGGCATCGACTTTGCGATCGGATGCTAGTCTTCCCAGCTATCAG

8.4.1. Bioinformatic processing and quality control

Overall, we obtained more than 70 million reads. Sequences were quality-controlled and error corrected to trace back up to eight PCR and sequencing errors to the original barcodes. Literature suggests that a true stem cell has to contribute above 0.5% cell content in blood to be biologically relevant (Bystrykh and Belderbos, 2016; Dykstra et al., 2011; Verovskaya et al., 2013). However, there is an important difference stimulating us to define a different threshold. In this work, we dealt with a competitive setup of up to three different vector constructs (barcode backbones), which we wanted to evaluate against each other. Due to varying clonal situations between the individual animals, we needed a relative threshold, which could be applied over all mouse groups and vector constructs. Therefore, our threshold was not based on total blood/bone marrow contribution, but was set relative to the total reads of one specific barcode backbone. For every sample, the total number of reads representing one specific backbone, e.g. GFP, were summarized and frequencies of the individual GFP barcodes relative to this absolute number were calculated. We then omitted all barcodes present in the sample below a threshold frequency of 0.5%. In the end, around 50 million reads remain for all our samples. The average number of reads in a sample (= read count) was ~2.8×10^5. However, one bone marrow sample achieved the maximum read count of 1.56×10^6 reads, while sorted Ly.6G cells of

another sample were only sequenced with 1433 reads (Figure 25). To avoid skewing of frequencies due to low read counts we applied another filter, excluding barcode backbones with less than 800 backbone reads and below 5% of total sample reads from our analysis.

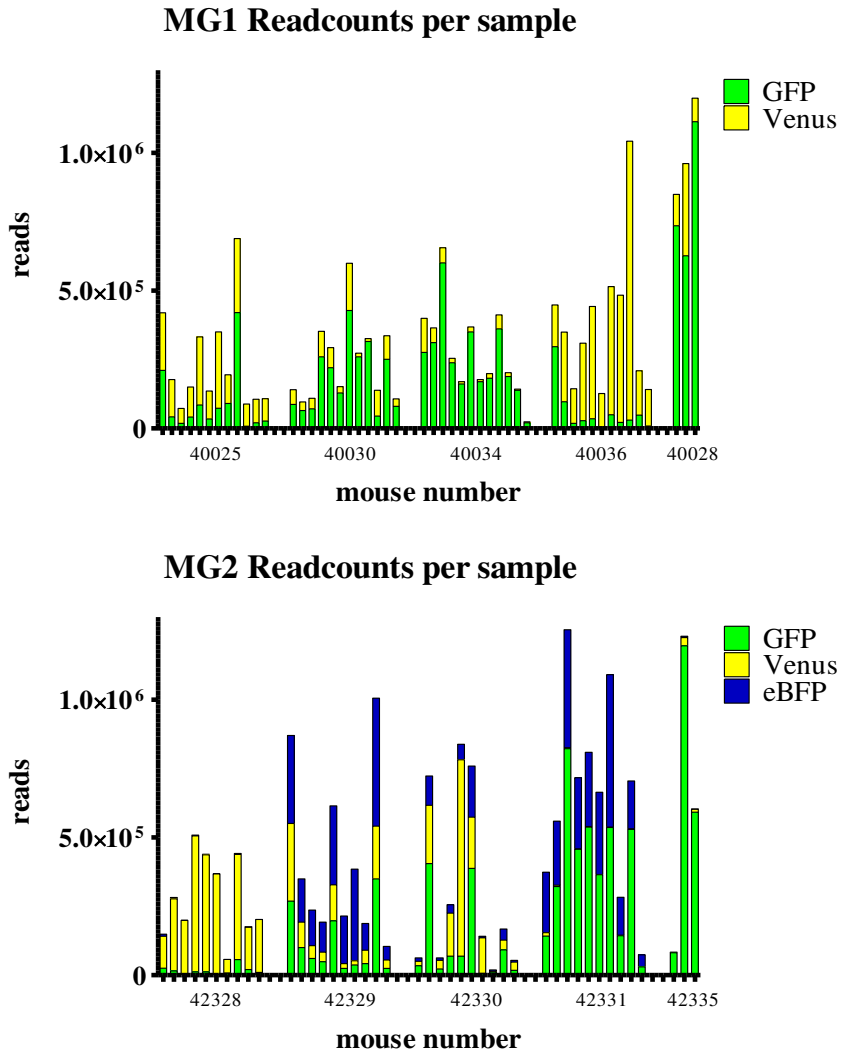


Figure 25 – Barcode-backbone specific read counts for individual samples for MG1 and MG2. 10 - 12 different samples from haematopoietic organs were sequenced for most animals. Each sample is represented by one bar. For clarity, only the mouse number is indicated. Total read counts show strong variance between samples and or animals (total bar height) and distribution of the individual barcode backbones (colours).

To assess possible background-contamination, I analysed the false positive reads of every sample. False-positive reads are barcode backbones, which should not be in that sample, e.g. BFP-barcodes in MG1. The average percentage of such false-positive reads was 0.98% (± 4.6%) of total sample reads. The high deviations were mainly caused by samples from one specific animal showing four samples with low read counts (three of them with < 4.000 reads) with 20-40% of them from false-positive backbones. Excluding these samples lowers the overall percentage of false-positive background to 0.19% (± 1.14%) of total reads. Based thereon, we can conclude that our efforts to prevent (cross)-contamination in the majority of cases worked and the dataset has only a very limited background.

In the next step, I had a look at the distribution of barcodes across different animals. Of note, this analysis was done with a frequency threshold of 0.1%, thus including biologically irrelevant barcodes to ensure higher sensitivity. As Figure 26 shows, over 91% of all barcodes found for an individual backbone, were unique to one animal. Since the same transduced donor cells were used for up to four recipients, it should be possible for individual barcodes to appear in more than one mouse.

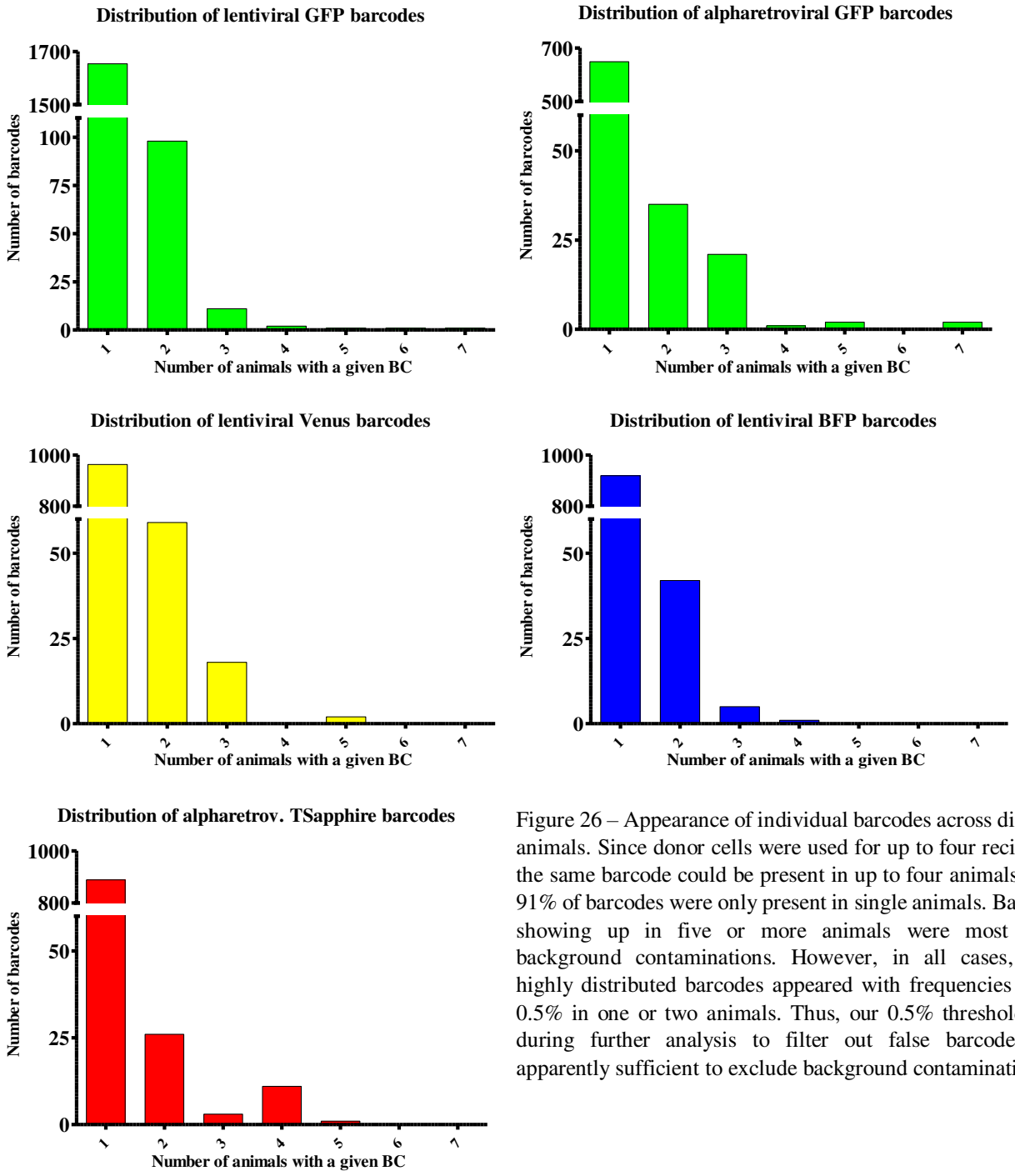


Figure 26 – Appearance of individual barcodes across different animals. Since donor cells were used for up to four recipients, the same barcode could be present in up to four animals. Over 91% of barcodes were only present in single animals. Barcodes showing up in five or more animals were most likely background contaminations. However, in all cases, those highly distributed barcodes appeared with frequencies below 0.5% in one or two animals. Thus, our 0.5% threshold used during further analysis to filter out false barcodes was apparently sufficient to exclude background contaminations.

Barcodes found in more than four animals most likely represented background contamination. Analysing the frequencies of these highly distributed barcodes revealed that all of them were present with < 0.5% in at least one of the animals. Thus, our 0.5% threshold used during further analysis to

filter out false barcodes was apparently sufficient to exclude background contaminations. This analysis strengthened the conclusion, that we had only a very low background contamination levels in our dataset. Additionally, we analysed the average Hamming Distance (HD) representing the difference between all individual barcodes for each backbone (exemplified in section 5.4.1). The samples showed a median Hamming Distance of 24 (± 2) after error correction. This implicates that two barcodes on average only share 8 out of 32 variable positions, clearly discriminating them from each other. This high HD justified our error-correction threshold of HD = 8. The latter allowed allocation of barcodes with up to eight different nucleotides, e.g. from PCR- or sequencing errors, back to one common ancestor.

Overall, the quality analysis of the dataset showed that we recovered barcodes with high Hamming Distances, allowing reliable discrimination of individual barcodes. After exclusion of some outliers, the background contamination level was low enough to be ignored. Nevertheless, I had to omit some barcode backbones of individual samples due to low read counts and therefore too low coverage to calculate reliable frequencies.

8.4.2. In depth analysis of the NGS dataset

8.4.2.1. Relative number of barcodes over time

Within the first analysis, we looked at absolute barcode numbers found in each sample and their development over time. There were strong variances in barcode numbers found within individual animals. For example, one animal had 54 GFP barcodes contributing to the first blood sample, while another animal from the same group only showed 26 GFP barcodes. Thus, a relative analysis was chosen where the number of barcodes found in the first blood sample, taken 6 weeks after transplantation, was used as a reference value and barcode numbers of all consecutive samples were set relative to that value. The general pattern of the resulting diagrams (Figure 27) shows a decreasing number of barcodes during the observation period. This is in line with our expectations, as short- and intermediate-term progenitors, present in the transplanted lineage-negative fraction, exhaust over time. However, some animals showed stable or even increasing numbers of barcodes over time. The number of initial barcodes found is indicated in the figure above the PB6w reference value. It is comparable between all four groups with some higher numbers for the lentiviral GFP construct (MG1 + MG2 GFP). Over time, numbers of barcodes found decreased by about 20 – 50%, which was apparently independent of the vector construct used. There might be an exception for the alpharetroviral pd-TSapphire construct (MG3 + MG4.2). Unfortunately, high variations in the latter one, where clone numbers for two animals rose while for two others they decreased by >70%, did not allow a clear-cut proposition. In general, the individual differences between animals seemed to be more prominent than the influence of different constructs. In addition, I did not observe any, positive or negative, correlation between FP expression as measured by flow cytometry and recovered clone numbers. For example, 34 barcodes were found in one animal with 31.5% GFP expression, but 40 clones from another animal with only 18.9% GFP.

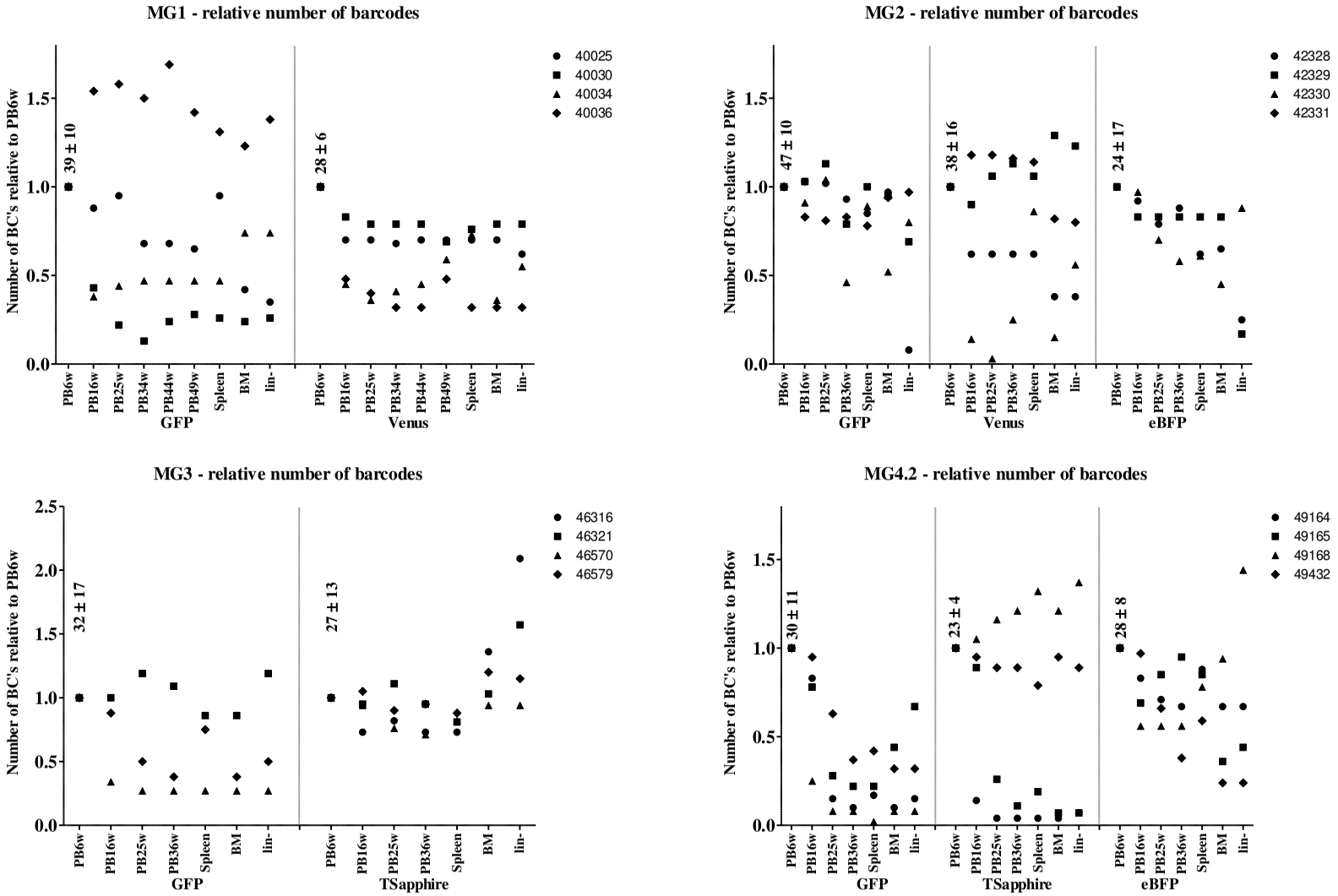


Figure 27 – Number of barcodes relative to the first blood sample (PB6w) for three to four animals per group. The actual number of barcodes found within PB6w is indicated above the first data point (mean + SD). Barcode numbers found in subsequent samples are set in relation to PB6w. In most animals, a decline of barcode numbers in the successive blood samples is seen. This indicates the expected loss of short/intermediate progenitors not able to support long-term blood production. The GFP backbone for mouse #46316 (MG3) is excluded due to a very dominant clone in PB6w skewing the reference value. PBxw: peripheral blood taken after x weeks; BM: bone marrow; lin-: lineage-negative fraction.

8.4.2.2. Temporal and spatial dynamics of haematopoietic reconstitution

In the next step, I analysed the temporal and spatial distribution of our sequenced barcodes. These can be best represented via stacked bar plot, where individual barcodes from one backbone are colour coded and stacked upon each other for every sample. Samples of the same animal are then displayed in one plot. The general colour scheme defines the backbone analysed (green = GFP, yellow = Venus, red = TSapphire, blue = BFP), so up to three plots are required per animal. The y-axis represents the frequency of an individual barcode within all backbone reads. Every horizontal colour represents the same barcode found within multiple samples. Blood samples from different time points on the left plot side represent the temporal dynamics. Samples from the final analysis, oriented to the right side of the plot, show the spatial distribution in blood, bone marrow, spleen and selected myeloid and lymphoid subsets at the final time point.

The first plots below from mouse #40025 (MG1, Lenti-EFS-GFP vs. Lenti-pd-Venus, Figure 28, upper panel) show a stable, polyclonal, situation for both marked populations. Nevertheless, there is one Venus clone, indicated by the light yellow bar at the very bottom appearing at PB16w, slowly rising over time and contributing 49% of the marked lineage-negative sample in the end. The plots for animal #40030 (MG1, Lenti-EFS-GFP vs. Lenti-pd-Venus, Figure 28, lower panel) show a less polyclonal situation, where seven big barcodes dominate the fraction of EFS marked haematopoiesis after 16 weeks. It is imaginable, that these seven barcodes represent only one or two clones with multiple integrations, as their pattern and dynamics look very even. The pd backbone shows two clones becoming prominent within the later blood samples, but seem to be outcompeted by another clone in the lineage-negative/bone marrow fraction.

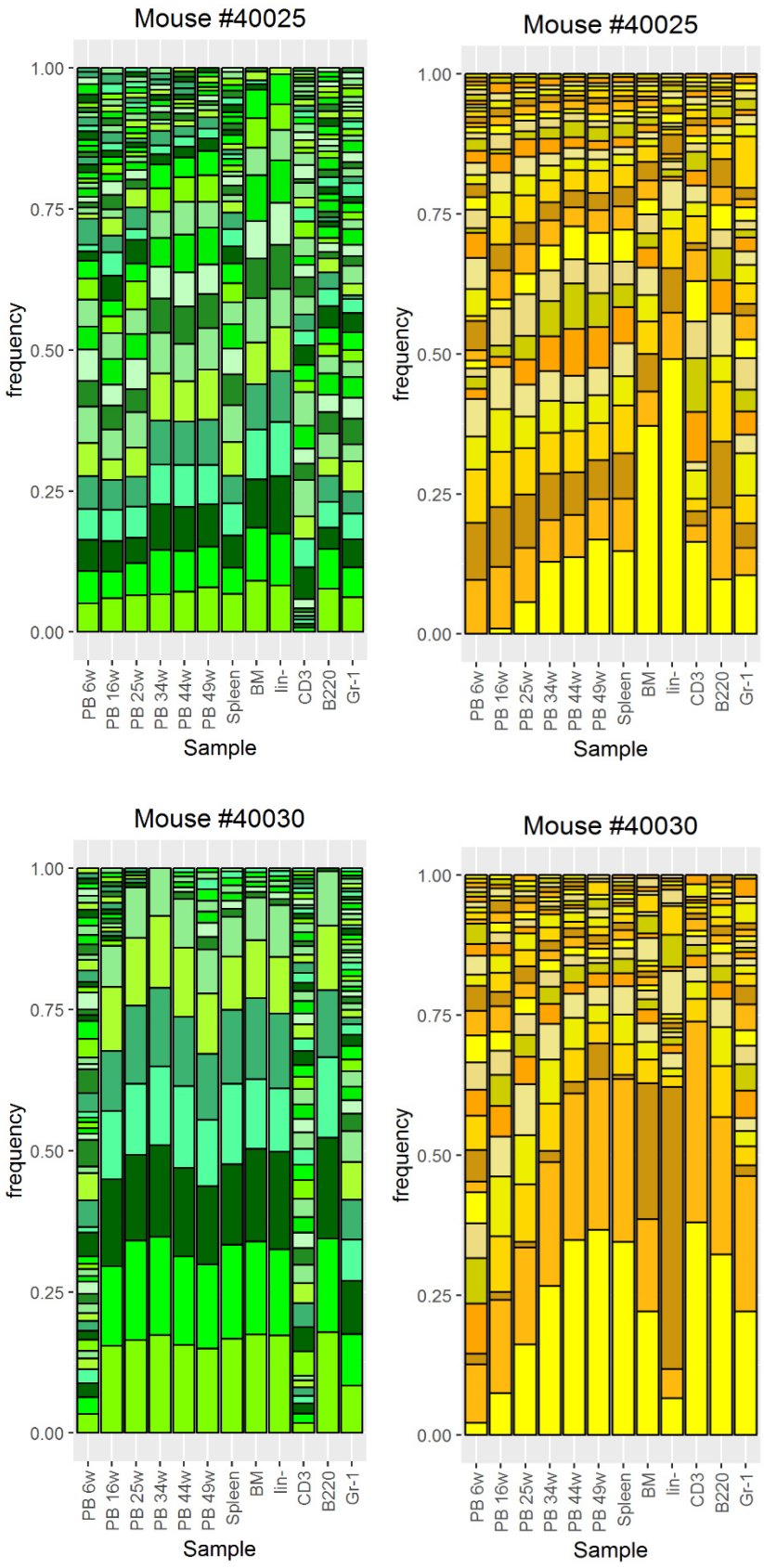
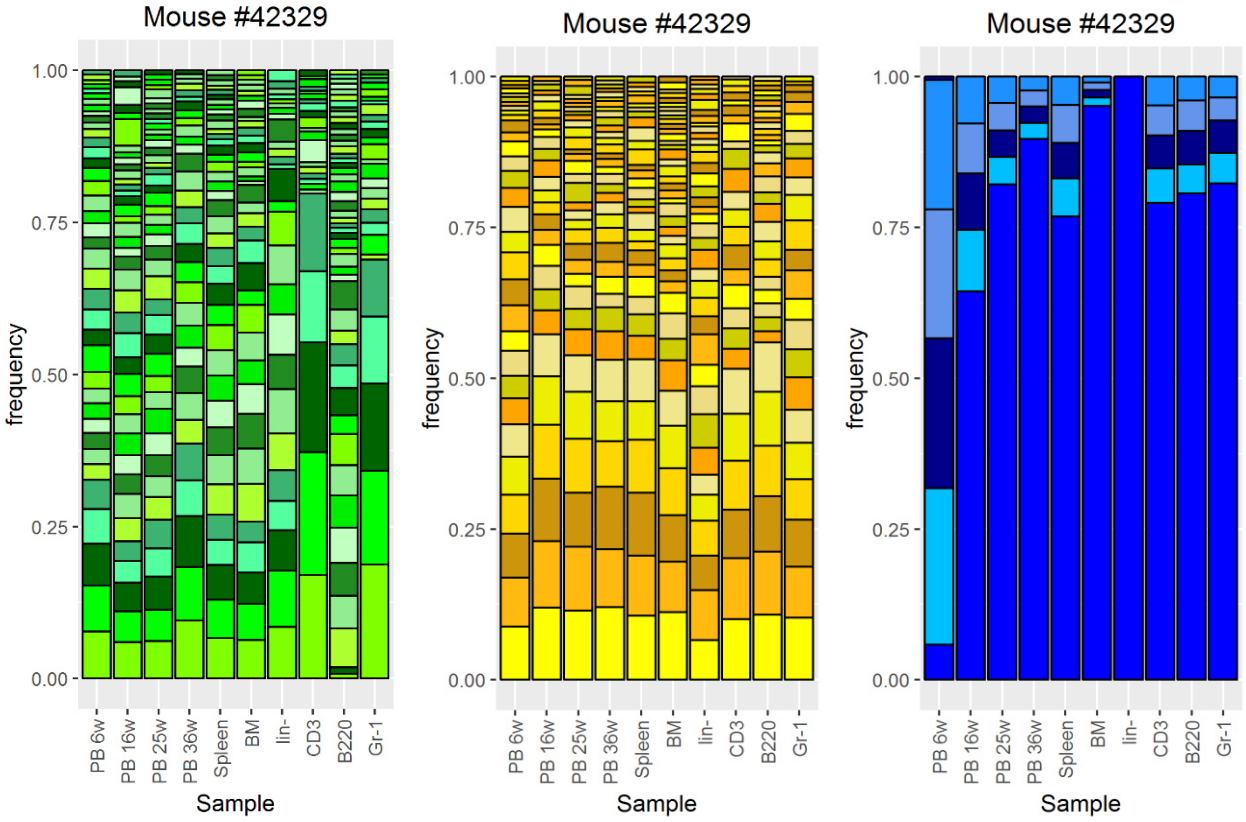


Figure 28 – Bar plots from mouse #40025 and #40030, both MG1. All barcodes from one backbone are colour coded and stacked upon each other based on the percentage of reads (frequency) an individual barcode is found. The general colour scheme defines the vector construct analysed (green = Lenti-EFS-GFP, yellow = Lenti-pd-Venus). Every constant horizontal colour represents the same barcode found within multiple samples. Blood samples from different time points

(bars on the left) represent the temporal dynamics, while samples from the final analysis (oriented to the right side of the plot) show the spatial distribution at the final time point. In mouse #40025, both constructs show a diverse, polyclonal pattern although one pd-Venus clone starts becoming prominent within the blood samples at later time points and represents approx. 50% of the marked lineage-negative fraction. Contrary, both constructs show development of few dominant clones in the other animal (#40030). PB xw: peripheral blood taken after x weeks; BM: bone marrow; lin-: lineage-negative fraction; CD3: T cell subset; B220: B cell subset; Gr-1: granulocyte subset.

Animals from MG2 (Lenti-EFS-GFP vs. Lenti-pd-Venus vs. Lenti-SFFV-BFP) show EFS and pd backbone patterns similar to the ones already shown, e.g. mouse #42330 with two dominant Venus barcodes. Again, the even pattern and proportions of these barcodes may point towards one clone bearing two vector integrations. In the SFFV-marked cells one clone gets dominant in each animal, which is not the same clone. A third animal sequenced from this group (#42328, plot not shown), shows too low SFFV read counts in the bone marrow and lineage-negative sample for analysis, but the final blood sample and spleen show one prominent clone with 31 and 41% contribution, respectively.



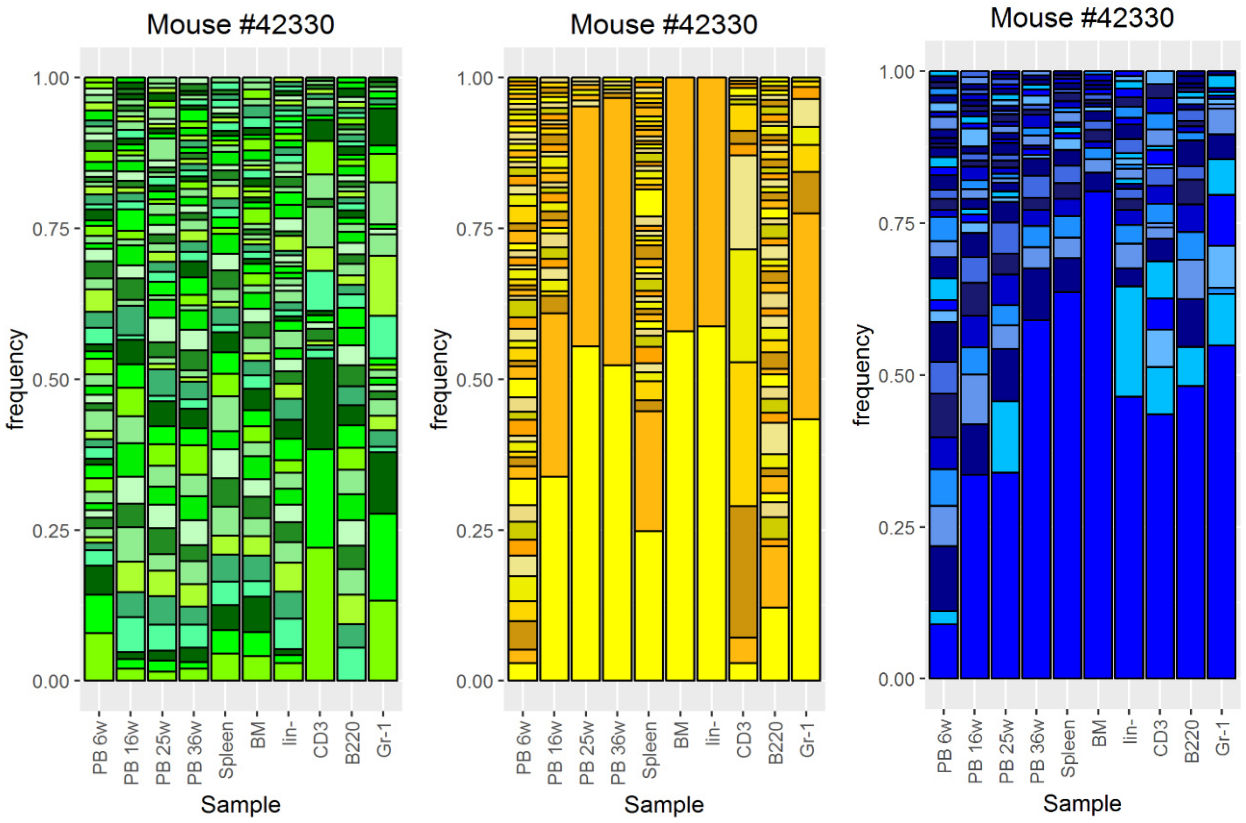


Figure 29 – Bar plots from mouse#42329 and 42330, both MG2. The general colour scheme defines the vector construct analysed (green = EFS-GFP, yellow = pd-Venus, blue = Lenti-SFFV-BFP). Both animals display a polyclonal situation for the EFS-GFP construct. Additionally, dominant SFFV-BFP clones emerge in both animals. Dynamics of pd-Venus constructs differ, as one animal keeps a constant polyclonal pattern, while dominant clones develop in #42330. PBxw: peripheral blood taken after x weeks; BM: bone marrow; lin-: lineage-negative fraction; CD3: T cell subset; B220: B cell subset; Gr-1: granulocyte subset.

Mouse #46316 from MG3 (Alpha-pd-TSapp vs. Alpha-EFS-GFP) was the only animal in our analysis showing clonal dominance within in the first blood sample (one GFP clone with 98% contribution) which disappears and is detectable only on a low level (1% in PB36w, 4% in BM) in later samples. However, two other barcodes, becoming prominent 16 weeks after transplantation, compensate disappearance of that clone. As the proportion between both of these emerging barcodes in all samples is quite even, it is likely one clone with a double integration. The TSapphire-marked haematopoietic fraction of that animal is dominated from the beginning by one clone showing at least 59% contribution in all samples. In contrast, mouse #46321 shows a fluctuating, polyclonal, dynamic situation for both vector constructs. Haematopoietic reconstitution from the marked clones for other two sequenced animals (plots not shown) shows a stable, polyclonal, situation in one of them. Analysis of the fourth animal reveals three prominent EFS and one bigger pd clone, contributing around 25% in blood samples.

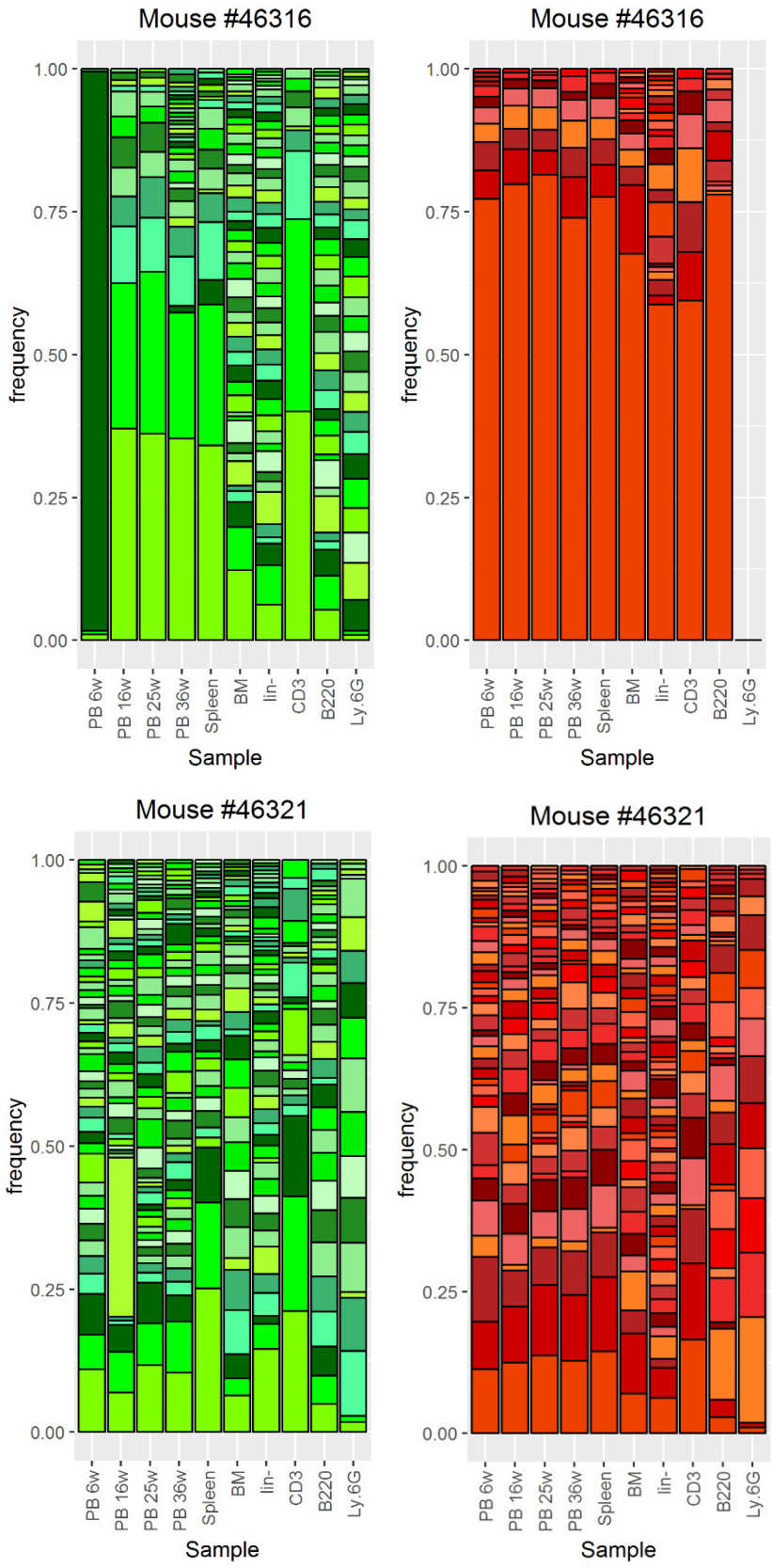
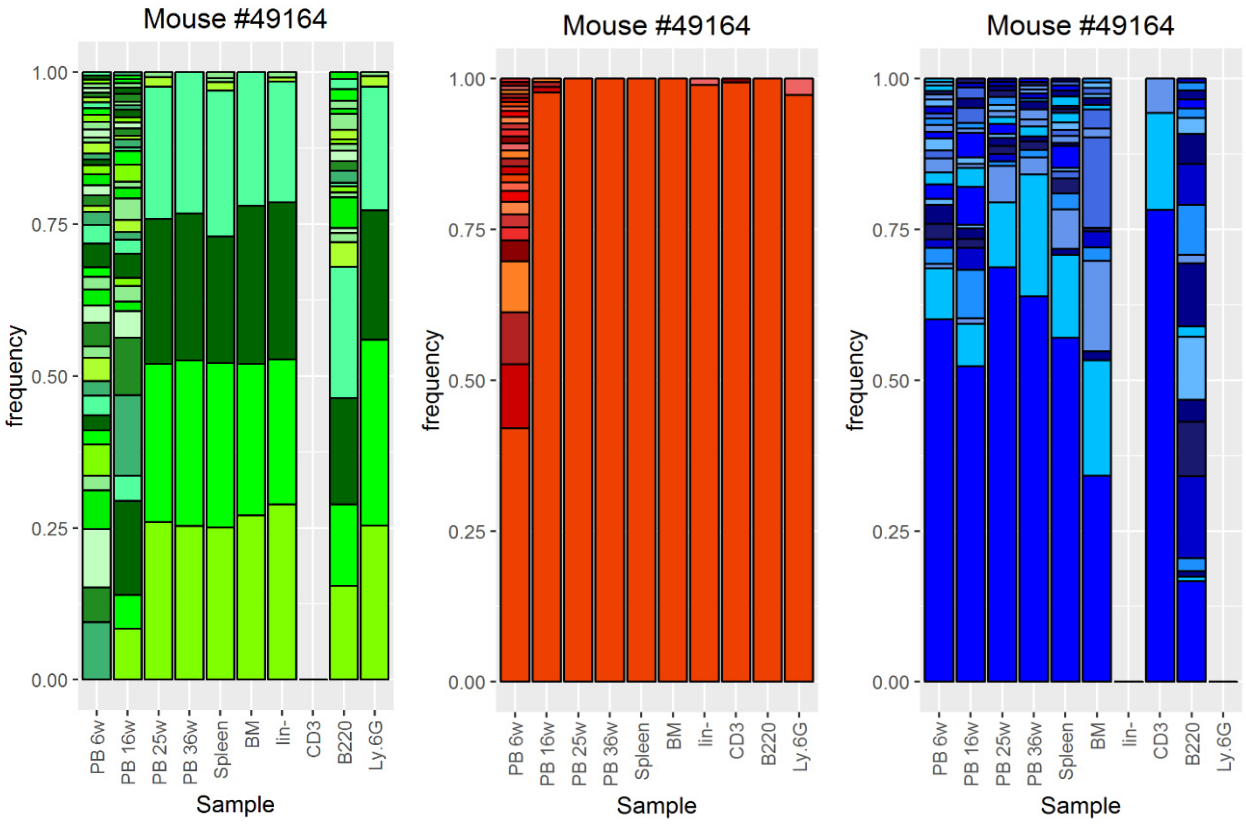


Figure 30 – Bar plots from mouse#46316 and 46321, both MG3. The general colour scheme defines the vector construct analysed (green = Alpha-EFS-GFP, red = -Alpha-pd-TSapp). Reconstitution dynamics of #46316 shows one EFS-GFP clone dominating 6 weeks after transplantation, nearly completely disappearing until 16 weeks after transplantation. One clone dominates haematopoiesis of the fraction transduced with the alpharetroviral pd-TSapp construct over the whole

observation period. Clonal dynamics of mouse #46321 show a stable, polyclonal situation over time. PBxw: peripheral blood taken after x weeks; BM: bone marrow; lin-: lineage-negative fraction; CD3: T cell subset; B220: B cell subset; Ly.6G: granulocyte subset.

In MG4.2 (Alpha-pd-TSapp vs. Alpha-EFS-GFP vs. Lenti-SFFV-BFP) all 4 sequenced animals show development of 1 to 4 dominant EFS clones, sometimes most likely with multiple integrations. Two animals have an almost monoclonal pd backbone situation in addition. Pd backbone analysis in the other two animals shows some bigger clones (contributions around 15-20%) but none of them becomes dominant. SFFV analysis can only be done in three out of the four animals, as the low read count filter omits 5 samples from the last animal. In all three animals, 1 or 2 dominant clones are present over the whole observation period, representing at least 60% of SFFV reads in one sample.



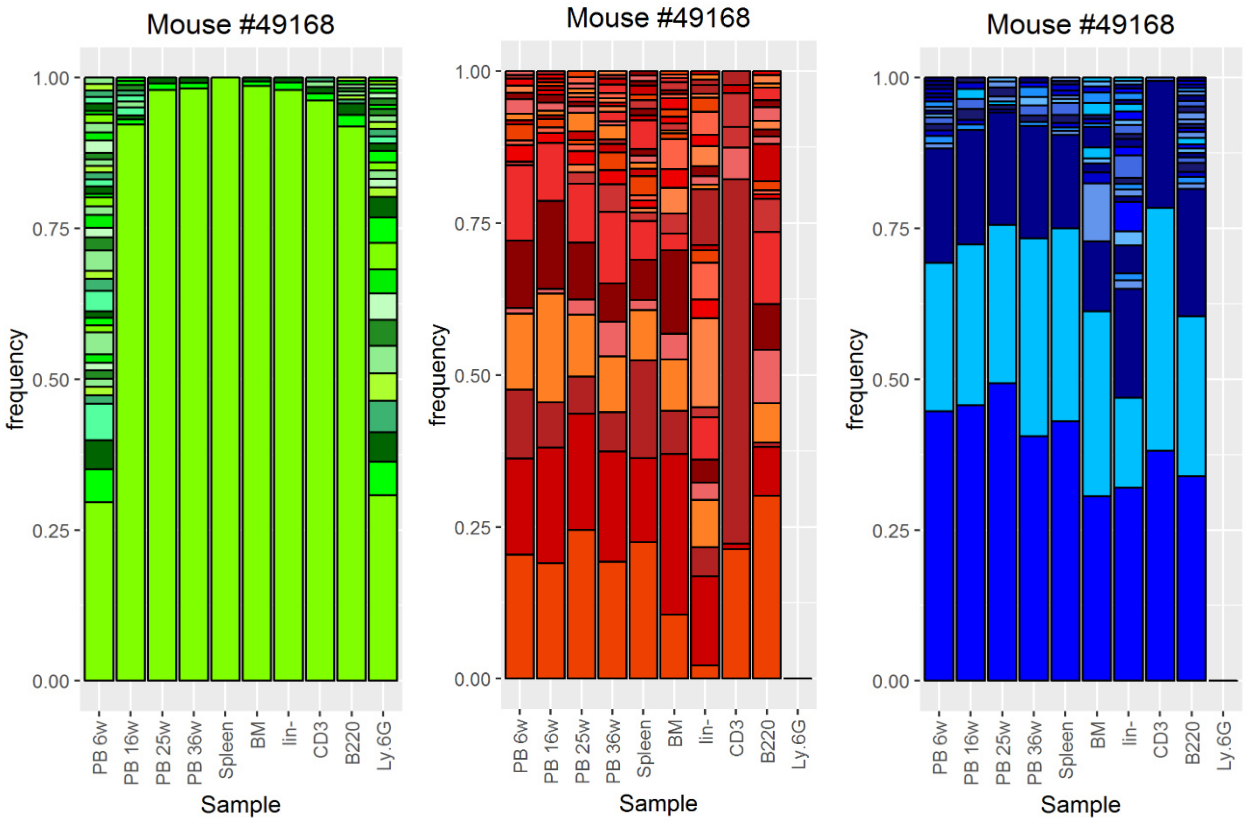


Figure 31 – Bar plots from mouse#49164 and 49168, both MG4.2. The general colour scheme defines the vector construct analysed (green = Alpha-EFS-GFP, red = -Alpha-pd-TSapp, blue = Lenti-SFFV-BFP). Both animals display emergence of prominent or dominant clones in all three constructs analysed. Absent bars indicate samples omitted due to low read counts. PBxw: peripheral blood taken after x weeks; BM: bone marrow; lin-: lineage-negative fraction; CD3: T cell subset; B220: B cell subset; Ly.6G: granulocyte subset.

In general, the number of contributing clones decreased over time, as the plots display less segmentation of the bars from PB6w to the later blood or organ samples. This is in line with the observations from the barcode number analysis. Most clones, prominent at later time points, were already detected in the first blood sample, 6 weeks after transplantation, although often less prominent. There were strong variations between individual animals. As already mentioned, the transduction rates (Figure 18), especially for the BFP constructs, were higher than intended. Thus, some cells should harbour more than one integration. Unfortunately, there is no definite way to identify those clones based on NGS data. Therefore, I can only speculate that barcodes showing similar tendencies and somewhat stable proportions between each other in the upcoming analyses may originate from the same cell.

Overall, bar plot frequency analysis reveals a diverse picture of varying monoclonal to polyclonal situations in our samples, depending on the individual animals. I was able to detect prominent/dominant clones in temporal and/or spatial dimensions for all five different vector constructs used. In general, two major patterns can be observed: In some animals, haematopoietic dynamics were already stable 6 weeks after transplantation and show only slight fluctuations in frequencies and distribution of clones observed over the following time points. I observed different levels of clonality within this stable haematopoiesis, ranging from dominant clones (#46316 TSapp,

page 70) to polyclonal situations (#40025 GFP, page 67). The second pattern displays emerging contribution from (one or several) clones between 6 and 16 weeks. The frequencies of these clones then either continue to rise towards the next time points (#42330 BFP, page 69) or transition into a plateau (#46316 GFP, page 70). I did not observe pattern specific prevalence for any of the tested constructs or within groups. Thus, general patters for the vector constructs look very similar, as differences seem to be more on the individual animal's level.

8.4.2.3. Validation of NGS frequencies via digital droplet PCR

Digital droplet PCR (ddPCR) is a technique that allows quantification of nucleic acids with high sensitivity and without the need of calibration curves (Hindson et al., 2011). To quantify the NGS frequencies found, we selected individual barcodes from our dataset and analysed them via ddPCR. As described in the Methods section (page 47) the first probe/primer set was promoter specific, while the second set was specific for the selected barcode, allowing quantification of the frequency from one specific barcode within one backbone. We focused on the more prominent/dominant clones, as they should be easy detectable, even if only 1% of DNA in our sample were transduced with the specific backbone. As the representative Figure 32 shows, we did not obtain clear-cut results. NGS and ddPCR values correlate really well in some cases, as demonstrated by the first two pairs of bars for the big TSapphire clone in bone marrow of mouse 49164 (page 71) as well as the big BFP clone in bone marrow of mouse 42329 (page 68). There are other barcodes showing some level of discrepancy between NGS and ddPCR values, like the BFP clone in the Spleen of mouse 49164 (page 71). Finally, there are some clones, like the dominant GFP clone at PB6w in mouse 46316 (page 70), where we only detect fractions of the NGS values via ddPCR.

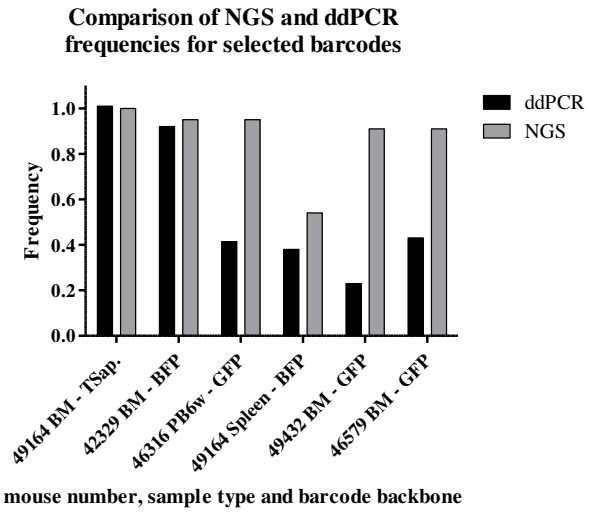


Figure 32 – Representative selection of individual barcodes analysed via digital droplet PCR (ddPCR, black) and correlated NGS frequencies (gray). The results are ambivalent, as data for some barcodes correlates very well, represented by the first two sets of bars. Others, like the fourth pair of bars, show some level of variation. In other cases ddPCR data indicates much lower contribution of individual barcodes than the corresponding NGS data. Possible explanations are discussed in the text.

There are several possible explanations. First, ddPCR in general shows around 10% variability (Stahl et al., (2016) and own observations, unpublished). Secondly, the barcode-specific primers were designed after a general procedure, starting at the second wobble pair and extending until a calculated annealing temperature of 59°C was reached. As consequence of the wobble positions between different BC specific primers, there are varying amounts of GC content, hairpin and dimerization possibilities that may influence the PCR outcome. Thirdly, bioinformatic processing of NGS results allows one mismatch within the barcode backbone sequence and summarizes up to 8 mismatches between individual wobbles to a common ancestor barcode, whose sequence was used for primer design. However, omitting all samples with more than one backbone mismatch and applying quality thresholds may lead to differences between NGS and ddPCR, as they would be detected in the latter one. Fourthly, it could just be a sample distribution or sample-size problem. We only processed 20 to 50 ng genomic DNA within the ddPCR reaction, both to minimise the use of material and due to the maximum volume usable. 50 ng represent around 8300 cells (50 ng/6 pg, the weight of diploid DNA per cell). Only a fraction of those cells is transduced with the specific barcode backbone and an even smaller part of that fraction contains our target barcode. Thus, barcode distribution might show some level of variation between several ddPCR and/or NGS reactions (Thielecke et al., 2017).

Overall, we were able to correlate NGS and ddPCR data for individual barcodes in some samples, whereas others showed different levels of discrepancies. We can only offer possible explanations, as the available murine material was too limited for detailed further investigation. However, we addressed the quantifiability of genetic barcodes in a different setup (Thielecke et al., 2017), further evaluated later on.

8.4.2.4. Diversity analysis and evenness distribution

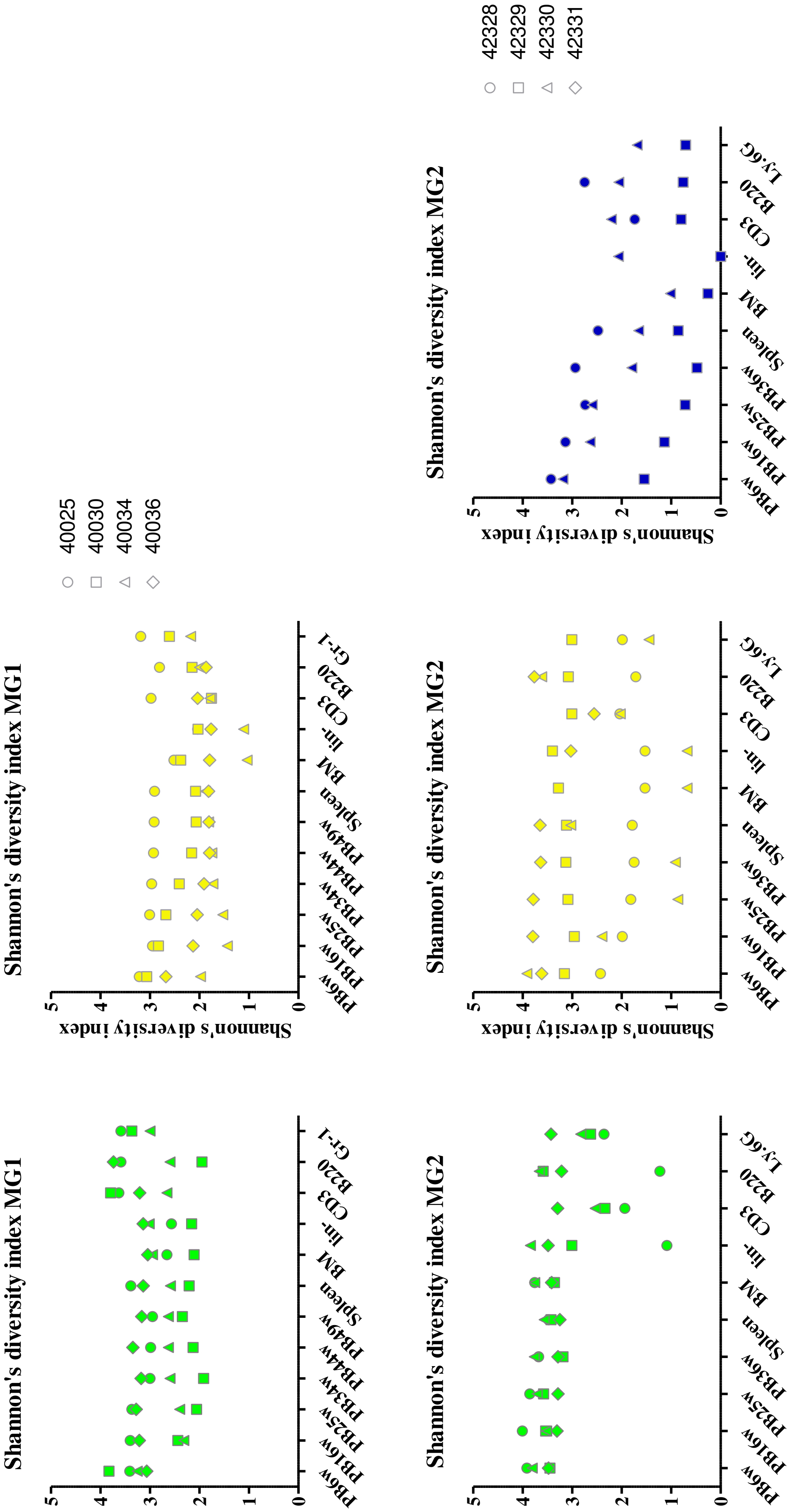
Bar plots, as shown above, are a good way of displaying clonal dynamics. However, comparing different groups or evaluating trends with them is suboptimal, as they do not generate tangible numbers or indices. I tried to address this issue utilizing the Shannon index. This index was originally developed for communication theory (Shannon, 1948), similar to the introduced Hamming distance. It became established in ecology, where it is used to describe species diversity (Hill, 1973; Spellerberg and Fedor, 2003). Furthermore, it can be used in the context of genetic barcoding (Bystrykh and Belderbos, 2016; Porter et al., 2014; Selich et al., 2016). The Shannon index (H) is defined as $H = -\sum_{i=1}^{s} p_i * \log_2 p_i$, where p_i is the proportional abundance of barcode i and s the total number of barcodes. The sum over all sequenced barcodes then results in an index starting at zero, with no upper limit. For barcode context, a high Shannon index represents a diverse, polyclonal situation. A rising index over time either indicates the appearance of new clones or changes in the existing clones to a more equal distribution. A decrease in the Shannon index represents either the loss of clones or development of prominent/dominant clones making the distribution more uneven.

Analysing the Shannon indices in our dataset (Figure 33) reveals some interesting patterns. First, most animals start at least with an index around 3 in the first sample six weeks after transplantation (PB6w).

As it turns out, Shannon indices of 1.8 to 2 correlate pretty well with the existence of prominent clones mentioned in the bar plot analysis. For example, in the Venus bar plot for mouse #40025 (page 67), the lineage-negative sample with the 45% clone has a Shannon index of 2.04. A Shannon index of 1.5 indicates the presence of (a) dominant clone(s), like the BFP bar plot for mouse #49168 Spleen (page 72), which displays a Shannon index of 1.46. Lower numbers imply the existence of only one or two clones dominating haematopoiesis marked with that backbone and (near-)absence of other clones, e.g. GFP bar plot for #49168 (page 72), where samples PB16w to B220 have a Shannon index between 0 and 0.45.

Comparing the Shannon-index plots (Figure 33) for the individual animals within one group shows a diverse picture. Except for MG4.2, which I will discuss later on, there are at least two animals with (the majority) of samples above a Shannon index of 2, representing an polyclonal situation. On the contrary, there is at least one animal below that threshold for all constructs, except Lenti-EFS-GFP (MG1 and MG2).

Comparison of backbones/constructs used in different groups, e.g. Lenti-EFS-GFP in MG1 and MG2, can be done by calculating means over all four animals per group and comparing these means with the other group. The results indicate that Lenti-EFS-GFP and Lenti-pd-Venus backbones of MG1 and MG2 have pretty similar mean Shannon indices, with quotients of 0.75 to 1.32. The same calculation done for the Alpha-EFS-GFP and Alpha-pd-TSapp backbones of MG3 and MG4.2, shows that the quotient of the first sample (PB6w) is still comparable (0.8 for GFP and 1.08 for TSapp). However, at the later time points and samples, this quotient rises to 1.31 – 3.56 for GFP and 1.59 – 2.57 for TSapphire. This lower mean Shannon index indicates a decreased barcode diversity in MG4.2. This observation appears obvious by looking at the two respective plots, at least for Alpha-EFS-GFP. As this high divergence is true for both backbones, it is most likely related to a global effect in MG4.2. This could either be the change in multiplicity of infection (MOI), decreased from 50 to 30 to avoid high transduction rates or due to biological variations within the grafts. Transduction rates, even with the lowered MOI in MG4.2, are not that different in both groups (Figure 18, page 55). Furthermore, the mean number of initially recovered barcodes after 6 weeks is quite similar for all groups and most constructs (Figure 27, page 65). Consequently, it seems unlikely that the lowered MOI explains the observed barcode diversity. The only noticeable biological difference between MG3 and MG4.2 were slightly older donor animals (10 – 12 weeks compared to the usual 8 weeks) for the latter group due to logistical reasons. All other parameters and steps, e.g. purification of lineage-negative cells, cell numbers after 3 days in culture, workflow and protocols, were similar as before and thus comparable. I cannot exclude that an age difference of 2 – 4 weeks may facilitate such an effect, although I would not have a good explanation why. The comparison of the mean quotients for the BFP backbone of MG4.2 to MG2 shows values of 0.89 – 1.37, indicating comparability. However, as indicated earlier there might be a negative impact of the SFFV promoter and/or the high initial copy number resulting in the loss of BFP-transduced cells. Taken together, there is a global effect specific to MG4.2 decreasing the barcode diversity for at least two of the three backbones (EFS-GFP and pd-TSapphire).



Shannon's diversity index MG1
Shannon's diversity index MG2
Shannon's diversity index
PB6w
PB16w
PB25w
PB34w
PB44w
PB49w
PB36w
Spleen
BM
lin-
CD3
B220
Gr-1
Ly.6G
40025
40030
40034
40036
42328
42329
42330
42331

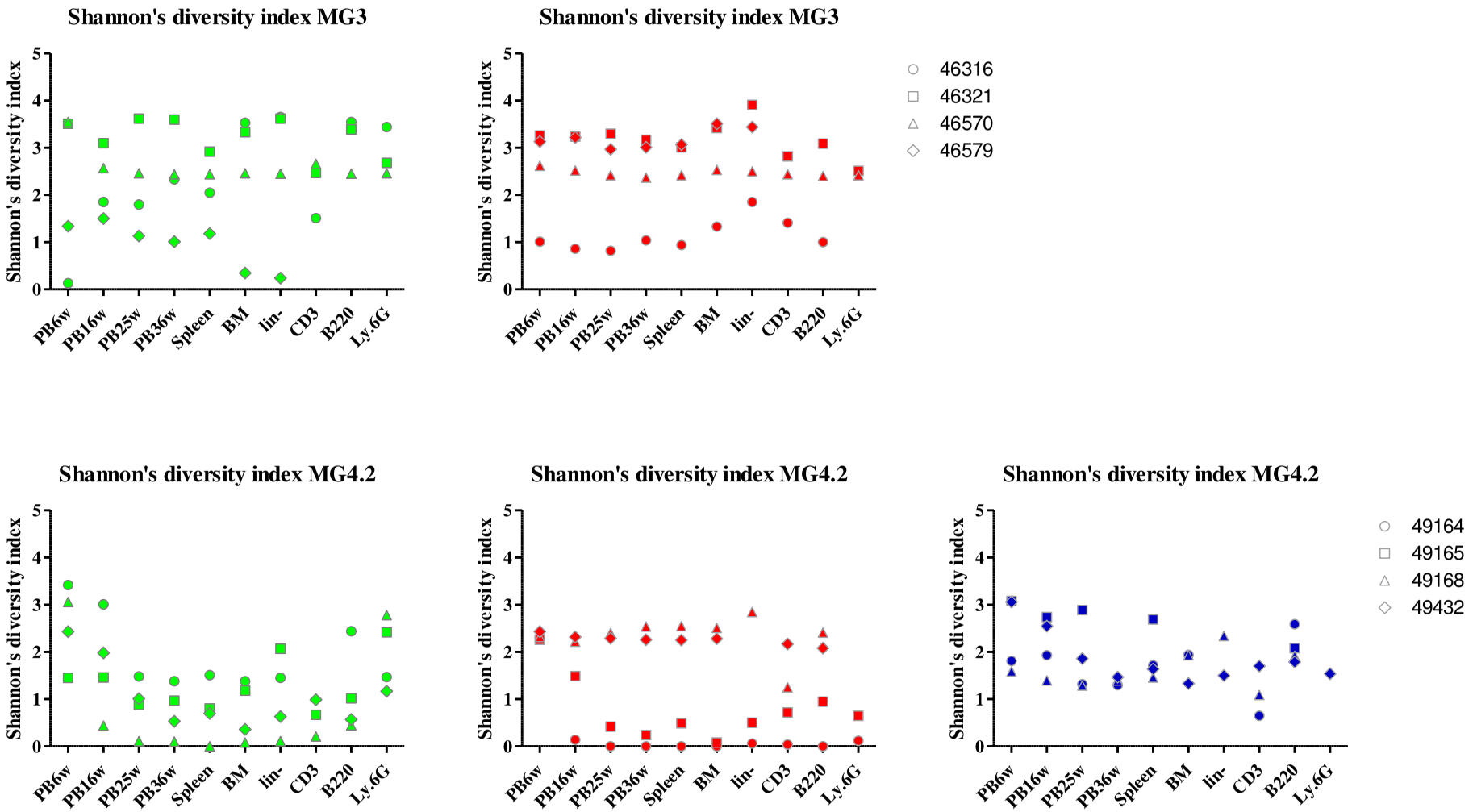


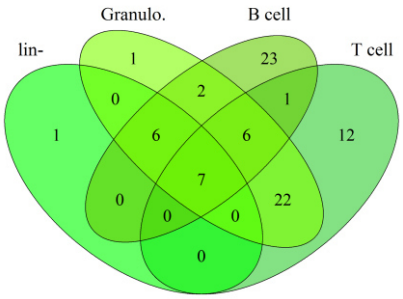
Figure 33 – Shannon index plots for all backbones and mouse groups. Adapting this biodiversity index to our barcode context allows to summarise barcode diversity and distribution in a single number per sample. Indices for every sample and barcode backbone (green = GFP, yellow = Venus, red = TSapphire, blue = BFP), from animals in one group are then summarised in one plot. As mentioned in the text, a Shannon index over 2 indicates a polyclonal situation, while an index below 2 points towards existence of prominent clones. Lower values represent situations with dominant clones. Missing data points indicate samples not available or taken out of the analysis due to low read counts. PBxw: peripheral blood taken after x weeks; BM: bone marrow; lin-: lineage-negative fraction; CD3: T cell subset; B200: B cell subset; Ly.6G or Gr-1: granulocyte subset.

In summary, the Shannon index values utilized here correlate very well with appearance of prominent or even dominant clones noted in the bar plots, providing numerical value for comparison of barcode diversity. On a grand scale, the Shannon plots show the same tendencies/patterns for all backbones and groups with the exception of MG4.2. The temporal a clonal dimensions, within most animal (horizontal tendencies within each plot) appear somewhat stable after an initial drop, which represents the loss in clone numbers between weeks 6 and 16, shown and discussed before. The differences become apparent when comparing individual animals within one plot, as there are several animals already starting with lower Shannon indices than others do. This supports the conclusion that intrinsic effects of individual animals on clonal reconstitution dynamics are more influential than the vector construct(s) or promoter(s) used.

8.4.2.5. Barcode distribution of sorted subsets

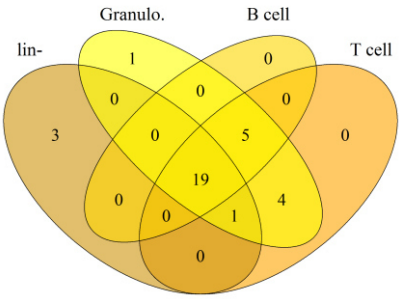
Fluorescence activated cell sorting of spleen samples at the final analysis time point allowed for sorting of three different subpopulations. Cells were sorted using the CD3e (T cell), B220 (B cell) and Gr-1/Ly.6G (granulocyte) surface markers. In most cases, only some ten-thousand granulocytes and around one million T and B cells could be obtained. Consequently, only low amounts (< 5ng/µL) of genomic DNA could be extracted. As only a maximum volume of 23 µL per sample was usable for the barcode retrieval PCR prior to sequencing, the 200 ng of genomic DNA used for all other samples were often not available with the subset samples. Therefore, the number of sampled cells in these subsets might be lower than in the other samples. In addition, barcode backbones of 16 samples (8x BFP, 4x TSapphire, 2x GFP and 2x Venus backbones), belonging to 9x granulocytes, 4x lineage-negatives and 2x T cells had to be omitted due to low read counts. Barcode distribution of the remaining samples was analysed and is represented here, using Venn diagrams (Figure 34). In this diagrams, numbers on the edges indicate barcodes found only in the specific subset, while numbers in overlapping areas represent barcodes found in multiple (or even all) analysed compartments. Analysing the distribution of barcodes shows no patterns or regularities between animals of the same group or barcode backbones. Some animals show a high number of clones unique to specific subsets, while most barcodes in other animals are shared between all four analysed compartments. This is in line with the observation(s) drawn from previous analyses, that there seems to be no construct-specific effect. Admittedly, with all the individual diversity depicted in the previous analyses and the additional challenges mentioned at the beginning of this section, it would be rather implausible to detect this effect in this subset analysis. Addition of a factor correcting the difference in sorted cells did not change the picture (data not shown).

Mouse #40025 GFP



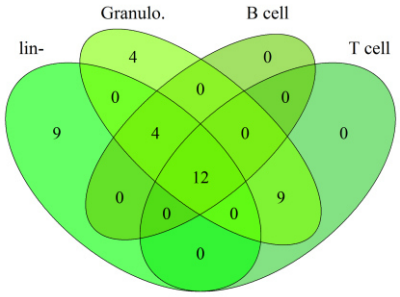
Total: 81 barcodes

Mouse #40025 Venus



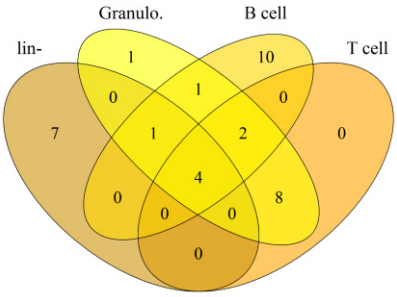
Total: 33 barcodes

Mouse #40034 GFP



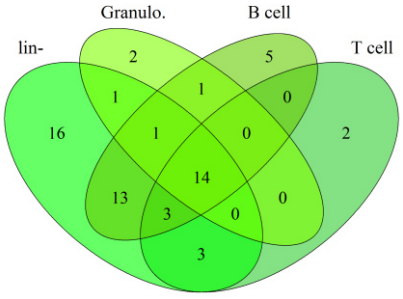
Total: 38 barcodes

Mouse #40034 Venus



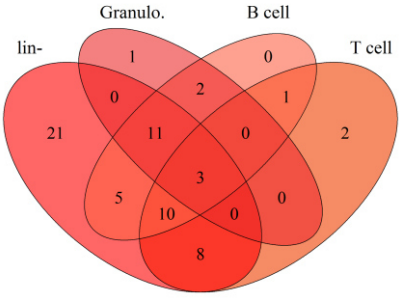
Total: 34 barcodes

Mouse #46321 GFP



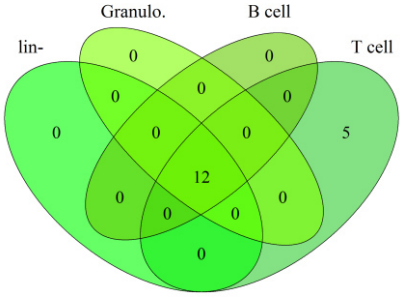
Total: 61 barcodes

Mouse #46321 TSapphire



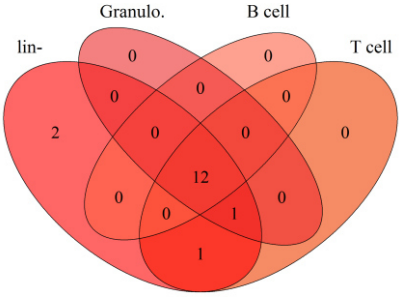
Total: 64 barcodes

Mouse #46570 GFP



Total: 17 barcodes

Mouse #46570 TSapphire



Total: 16 barcodes

Figure 34 – Barcode analysis within sorted splenic subsets. Genomic DNA from sorted T cells, B cells and granulocytes was extracted and barcodes sequenced. The numbers represent the same barcodes found in only one (edges) or multiple (overlap) subsets. Selected Venn diagrams for three out of the four mouse groups are shown. Subset analysis of MG4.2 was hampered by omitting nine backbones due to low read counts. T cell subset: CD3-positive; B cell subset: B220-positive; granulocyte-subset: Ly.6G or Gr-1 positive

9. Discussion

Marking of haematopoietic (stem) cells with genetic barcodes has become an established method to trace the fate and lineage-contributions of those cells as well as their progeny. Cutting-edge barcode systems combine high precision with high throughput, revealing clonal development in unmatched detail. Still, several open questions in the context of haematopoietic stem cell transplantation remain, as clonal development and dynamics during/after reconstitution are not fully investigated yet.

In the project presented here, I wanted to evaluate the influence of vector type and internal promoter on the haematopoietic reconstitution after bone marrow transplantation to determine the ideal vector for neutral labelling of cells. Hence, we analysed three different promoters (SFFV, EFS and promoter-deprived) in alpha- and lentiviral vector constructs equipped with our genetic barcoding system within four competitive *in-vivo* groups. I demonstrate that the BC32 system provides the tool, which is necessary to mark and trace cells or populations, transduced with different vector constructs or vector classes within a single animal in parallel. Barcode readout with NGS can be used to reveal clonal dynamics for each construct/population. This is possible by variation of the barcode specific consensus sequence (barcode backbone), adding another layer of complexity by "barcoding the barcodes". This backbone coding had previously only been used in a proof-of-principle experiment for lentiviral vectors, carrying different fluorescent proteins in a liver regeneration model utilizing the predecessor 16-wobble barcode system and Sanger sequencing (Cornils et al., 2014). My work shows *in vivo* application of the novel BC32 barcode system. I could prove that our system works as expected and facilitates stable long-term marking and tracking of populations and/or individual clones within the haematopoietic system. The high number of 32 variable positions allows over 10^{19} unique barcodes. Technical reasons reduce that number, but the estimated complexity (number of barcodes) was in the range of 10^6 barcodes for the plasmid libraries used in this work and meanwhile exceeds 10^8 barcodes with optimised protocols. This is more than sufficient for most, if not all, imaginable, technically feasible applications. Due to the high variance and consequently increased Hamming distances between recovered barcodes, we were able to correct up to eight PCR or sequencing-induced errors and achieve very detailed pictures of clonal composition and contribution within our samples. Vector-integrated NGS adapters allow for amplification and multiplexing of all our barcodes with only one PCR reaction, crucial for minimising PCR-induced bias and ensure the best level of quantifiability, as shown in Thielecke et al. (2017). In this thesis, I used up to three different vector constructs in parallel within a competitive *in-vivo* model. This number can be further increased and is only limited by the amount of genomic DNA usable for sequencing, NGS sample coverage and detection limit. Overall, this system allows analysis of several different parameters, like oncogenes, different mutations of the same gene or cell types/populations in parallel. The cumulative performance of one parameter can be

assessed through barcode-backbone analysis, whereas individual clonal dynamics are revealed by analysing the actual barcodes within the subgroups. All barcode backbones can be amplified with the same PCR reaction, minimising the likelihood of skewing or biasing sample composition and ensuring highest possible comparability. Utilising this system for a variety of possible applications and questions may greatly reduce the number of animals required.

Transduction rates are a critical parameter in barcoding experiments, as multiple vector integrations should be prevented to ensure unique labelling of cells. On the other hand, as many cells as possible should be labelled to exclude observations and conclusions based on limiting dilution or sample-size effects. Therefore, the optimal transduction rate is around 20% (Fehse et al., 2004; Kustikova et al., 2003). Even though I carefully titrated the viral supernatants before and transduced the murine cells in this work in a similar fashion, transduction rates between vector constructs varied (Figure 18, page 55). While the transduction rates for alpha- and lentiviral EFS-GFP constructs were around the targeted 20%, the lentiviral SFFV-BFP construct transduced nearly 60% of cells at the same MOI. Surprisingly, 6 weeks after transduction less cells expressed BFP in comparison to GFP, despite the higher transduction rate of the SFFV-BFP constructs (Figure 21). This might indicate a toxic effect caused by the high transduction rate, maybe due to an engraftment disadvantage associated with high numbers of inserted vector copies, or a silencing effect on the SFFV promoter. SFFV promoter silencing was already described in the context of haematopoietic and pluripotent stem cells (Herbst et al., 2012; Pfaff et al., 2013; Zhang et al., 2007). However, as the number of recovered BFP barcodes at certain time points in our dataset was comparable to that for other constructs (Figure 27, page 65), an effect promoted by promoter silencing seems unlikely. Brenner et al. (2003) reported higher engraftment potential for non-transduced human $CD34^+$ cells into immunodeficient mice compared to cells highly transduced with a gammaretroviral vector. They hypothesised that many HSC do not enter cell cycle within the transduction period and therefore can not be efficiently transduced with their gammaretroviral vectors. In contrast, more mature cells, efficiently stimulated into cell cycle progression “collected” many viral integrations, but were unable to engraft after transplantation. Observations from Schoedel et al., (2016) confirmed that long-term HSCs persist mainly in the G0-Phase, whereas short-term HSCs and more mature MPPs tend to show cell cycle progression. The lentiviral and alpharetroviral vectors used in this thesis do not require cell cycle progression to transduce their target cells (Katz et al., 2002; Yamashita and Emerman, 2006). Still, the argument with the more mature cells, displaying less engraftment potential could hold true in our model system too, as the cells were expanded *in vitro* for three days. Maetzig et al. (2011) hypothesised that the cytokine cocktail used for *ex-vivo* stimulation and expansion of HSCs may influence the experimental outcome by reducing the stem cell repertoire if chosen suboptimal. The cytokine stimulation cocktail used for expansion of lineage-negative cells in this thesis (stem cell factor (SCF), thrombopoietin (TPO), insulin-like growth factor 2 (IGF-2), and fibroblast growth factor-1 (FGF-1)) has previously been shown to expand the number of long-term HSC while not altering their function (Zhang and Lodish, 2005). Nevertheless, those authors reported differences in surface molecule expression between cultured and freshly isolated HSCs. As one can imagine, these changes in surface markers may have intrinsic reasons (or consequences) potentially associated with altered engraftment potential.

Additional negative effects of high viral load on engraftment potential are imaginable, potentially explaining the observed effect for SFFV-BFP discussed above. Unfortunately, I was unable to compare pre- and post-transplantation samples to further investigate this aspect in our setup as remaining barcodes, probably from non-integrated, reverse-transcribed vectors or vector plasmids transported within viral particles skew PCR data from the pre-transplant grafts.

High chimaerism as well as long-term FP expression data obtained via FC showed that the haematopoietic system in most animals was robustly reconstituted by donor cells. NGS of 176 samples selected from 16 animals created a dataset, which was used to investigate the clonal dynamics of each vector construct in a temporal and spatial context. Analysis revealed a decreasing number of barcodes (= clones) contributing to haematopoiesis over time. This was somewhat expected, as the transduced lineage-negative fraction contains short- and intermediate-term HSCs exhausting after several weeks or months (Benveniste et al., 2010). Interestingly, most clones with substantial haematopoietic contribution were detected over the whole observation period (6 weeks to 8 – 12 months), although with varying frequencies over time. This indicates a long-term HSC phenotype of these cells, as intermediate-term HSCs were reported to only sustain haematopoiesis for 6 – 8 months (Benveniste et al., 2010). Furthermore, my data are in line with the report of Sun et al. (2014) postulating that the composition of clones in non-transplanted, steady-state haematopoiesis changes over time, whereas clones in a transplantation setting are (to a certain degree) stable over time. As discussed earlier (page 24ff.), recombination-based, *in-situ*, barcoding systems, are currently insufficient to analyse steady-state haematopoiesis in high resolution. Technical progress should make research of this matter possible in the near future and give new insights.

Within our dataset, two major patterns of clonal reconstitution dynamics could be identified (section 8.4.2.2). In the first pattern, the haematopoietic system was already stably reconstituted 6 weeks after transplantation and clones showed only slight fluctuations in frequencies and distribution over the following time points. Thereby, varying levels of poly- to (almost) monoclonality were observed. The second major pattern of clonal haematopoietic dynamics showed emerging contribution from one or several clones between 6 and 16 weeks after transplantation. These contributions either continued to rise towards the subsequent time points or transitioned into a stable plateau. Similar to the first pattern described, varying levels of clonality were detectable. Overall, no pattern-specific prevalence for any of the tested constructs or within the competitive groups could be detected. The general patters of the vector constructs looked very similar, and differences seem to be more on the individual animals level.

NGS further revealed a relatively small number of barcodes reconstituting the marked fraction of peripheral blood at each time point. In most samples 20 – 40 clones per construct were detectable, resulting in 40 – 120 clones total. Often, 75% of the marked peripheral blood were contributed by only fifteen or less big(ger) clones per construct. The literature reports of recovered clones in transplantation settings vary, as there are reports of >60 (Naik et al., 2013) 30 – 50 (Lu et al., 2011) 10 – 50 (Verovskaya et al., 2013, 2014) or less clones (Gerrits et al., 2010). However, as different transduction and/or stimulation protocols, barcode systems, target populations, cell numbers, read-out methods and filtering parameters to answer their specific questions were used in the different studies, it is hard

compare these numbers. Furthermore, Brewer et al. (2016) could show that HSC differentiation is coupled to the transplantation dose, as high number of transplanted cells apparently increase the number of short-term HSC clones. I could not find literature reporting the clonal distribution in a non-transplantation setting via *in-situ* labelling approach. Sun et al. (2014) only reported the numbers, but not frequencies or distributions of tags recovered from certain subsets after *in-situ* labelling, and other groups investigating native haematopoiesis used inducible reporter genes rather than genetic barcodes (Busch et al., 2015). The total number of clones recovered in our system seems slightly higher than the ones reported by others. Some of this can be explained by our construct-specific thresholds needed to compare multiple barcode backbones. Other groups commonly used whole-blood-contribution thresholds for their single constructs. Additionally, similar dynamics and constant inter-barcode proportions observed within samples (section 8.4.2.2) indicate a high likelihood of clones with multiple integrations in several animals and vector constructs. Accordingly, our barcode numbers are potentially overestimating the number of existing clones. Yet, single cell sorting and PCRs would be required to unambiguously confirm the existence of multiple vector insertions and determine their barcode sequences. Bioinformatic correction of (potential) multi-barcode clones could be an option. However, we could demonstrate (Thielecke et al., 2017) that NGS read counts show some level of variation, even for samples with known barcodes and frequencies. Thus, there would be a risk to discard clones with only one integration, especially for low-contribution barcodes when filtering multi-barcode clones. Therefore, such a correction may add more uncertainty than the (probable) existence of some clones with multiple integrations.

Clonal dynamics, shown and analysed here in various ways, are very different between individual animals, even within groups receiving the same grafts. There is literature showing that lentiviral SIN vectors with an internal SFFV promoter, do possess low-level transforming potential in an *in-vitro* immortalisation assay (IVIM). In that assay (Arumugam et al., 2009; Modlich et al., 2006, 2009; Zychlinski et al., 2008), lineage-negative cells are transduced with retroviral vectors, expanded, and then distributed into 96-well plates in limiting dilutions. Some weeks later, the replating frequency is calculated as a quantitative readout of transforming events. The mentioned publications showed that the IVIM is rather specific to detect upregulation of *Evi1* and *Prdm16*, strong drivers of myeloid haematopoietic malignancies. IVIM results for lentiviral vectors with SFFV and EFS promoter were already reported (Modlich et al., 2009; Zychlinski et al., 2008). These publications demonstrated that lentiviral SFFV constructs show transformation rates of about $5x10^{-6}$, while no transforming potential was detectable for the EFS construct, even with over 20 copies per cell. Promoter-deprived constructs are unlikely to show higher transforming potential than the EFS construct. Additionally, cells transduced for this thesis harbour only limited numbers, ideally single integrations and barcodes. Using the data acquired during my work, I can estimate, that a total number of around 76,000 cells with engraftment potential (LSK $CD150^+$) have been transduced for this thesis (see the following section 9.1.1 and Appendix figure 6 for details), around 27,000 of them with SFFV-BFP. Given these numbers, it would be highly unlikely to observe malignant transformation. This is further supported by the observations of Montini et al. (2009) reporting that SIN-lentiviral SFFV vectors do not accelerate tumor onset within a tumor-prone mouse model *in vivo*. Promoter-deprived gammaretroviral vectors

were reported to not induce clonal imbalance in a similar transplantation setup to the one in this work, using LM-PCR as read-out (Cornils et al., 2009). This is somewhat different to the observation of dominant clones reported for the constructs tested here, but can potentially be explained with higher sensitivity of the BC32 system compared to LM-PCR, as well as variations in setups and vector constructs used.

There were several possible scenarios and outcomes with different effect sizes for the experimental setup used in this thesis. In the most clear-cut, but as mentioned unrealistic case, we would observe malignant transformations due to mutagenic effects of one vector. In an intermediate scenario, one vector construct, arguably SFFV, would produce dominant clones in the majority of cases, while the promoter-deprived and, maybe, EFS constructs would maintain stable, polyclonal situations. The last possibility would be that either there is no extrinsic effect or we are unable to detect it with our setup. In this scenario, all vector constructs used would behave in a similar way, with or without appearance of dominant clones. The data presented in this work indicates one of the latter two scenarios to be true. I observed the appearance of dominant clones, reconstituting the majority or even entirety of the marked fraction, for at least four out of the five constructs used. For the fifth construct (lentiviral EFS-GFP) we only found clones becoming more prominent, but they never reached a dominant status noted for the other constructs. Thus, I can conclude that we do not see a promoter and/or vector class (alpha- vs. lentiviral) induced effect with the setup we used. The exception could be the (strong) SFFV promoter. As discussed earlier, there seems to be a negative effect decreasing the variety of barcodes expected for the higher transduction rate. In addition, the second group receiving the SFFV-BFP construct exhibited a global effect of lowered barcode diversity due to unknown reasons (shown and discussed in section 8.4.2.4). As mentioned before, (albeit low) genotoxic potential of lentiviral SFFV vectors was reported (Modlich et al., 2009). This, in addition with the ambivalent data collected does not allow an unambiguous assessment of clonal influence for the SFFV promoter within our transplantation model. In general, the variability between individual animals, irrespective of constructs received, is too large, so a small effect for other constructs could be missed within the background variation.

It is important to emphasise that I did not observe any (obvious) malignant effects, like leukaemia development, related to the vector constructs tested. Furthermore, clonal dominance does not have to be malignant per se and may not be as bad as sometimes indicated. The literature reports at least one patient in a gene therapy trial where long-term (33 months at publication date) therapeutic benefit was mainly provided by one dominant clone displaying high transgene expression (Cavazzana-Calvo et al., 2010). However, a polyclonal system can most likely compensate some loss of clones/functionality to effects like promoter silencing or appearance of genomic instability, whereas a monoclonal situation may not (Ott et al., 2006; Stein et al., 2010). Therefore, a (stable) polyclonal situation should always be preferred compared to clonal dominance.

NGS analysis of FACS-sorted granulocyte as well as T- and B-cell subsets only returned a low number of recovered barcodes. This may be attributed to the low cell numbers sorted and consequently limited amounts of DNA available for sequencing or only few marked clones contributing to these lineages.

In addition, sequence reads from several subset samples had to be omitted due to low read counts. Therefore, I am unable to investigate the presence and/or dynamics of lineage-biased HSCs (introduced in section 5.4.3, page 31) within my dataset, which have been reported by several groups using different techniques (e.g. Adolfsson et al., 2005; Benz et al., 2012; Dykstra et al., 2007; Lu et al., 2011; Naik et al., 2013; Wu et al., 2014).

Quantification of individual barcodes based on NGS read counts proved challenging. The comparison of NGS and ddPCR data (section 8.4.2.3, page 73) gave ambivalent results, with very good correlation for some samples, whereas others showed high discrepancies. There are several possible explanations, already mentioned in that section, ranging from primer characteristics to sample distribution and sample sizes processed. Due to limited material, we were unable to further investigate the reasons on our primary murine samples. Instead, we used artificial low-complexity barcode preparations of our BC32 and the previous BC16 to systematically analyse quantifiability, accuracy and sensitivity using different variants of PCR cycle numbers and treatments (Thielecke et al., 2017). The created BC32 "minibulk" consisted of four single-cell clones, mixed equally, as shown by ddPCR. Each clone harboured one known BC32. Evaluation of frequencies after NGS, however, showed an uneven recovery, with one barcode constantly underrepresented. We tried to eliminate this bias by reducing PCR-cycle numbers, digesting genomic DNA prior to PCR and even generated 1-kb long fragments to exclude an influence of surrounding genomic features. The improved BC32 system showed less, but still present, systemic bias compared to the BC16. Despite all our efforts, a certain level of systemic bias remained, most probably related to intrinsic PCR and/or NGS features. Nevertheless, we were able to show that reduction of PCR cycles and rigorous quality controls are crucial when working with genetic barcodes. It can be concluded that absolute quantification of barcodes via NGS is much more complex than initially anticipated and sometimes presented in the literature.

In hindsight, there is one major point that, if done differently, might have lowered the observed variability. The lineage-negative fraction as used for transduction is enriched for haematopoietic stem and progenitor cells, but also contains cells already committed to specific lineages or populations (Doulatov et al., 2012). It might have been advantageous to transduce a more "stem-like" population, e.g. lineage-negative, $cKit^+$ $Sca1^+$ (LSK) cells. This subset is further enriched for HSC and progenitor cells, thought to contain around 10% long-term HCS (Challen et al., 2009; Okada et al., 1992) and represents approximately 20% of the lineage-negative fraction. In a subsequent project, we marked different stem and progenitor populations with FP coding lentiviral vectors. Only marked cells, initially Sca1 and cKit positive, were present three weeks after transplantation whereas all other chosen populations ($Sca1^-/cKit^+$ and $Sca1^{int}/cKit^{int}$) were gone or barely detectable (own observations, unpublished). As the first sample of the experiment in this thesis was taken six weeks after transplantation, most of our transduced, short-lived, lineage-negative cells were long gone. Therefore, it is tempting to speculate that transducing LSK cells would have enhanced the transduction rates of long-term and/or intermediate-term progenitors. This population could be even further enriched for LT-HSCs using markers like CD48 or CD150 (Challen et al., 2009; Doulatov et al., 2012; Oguro et al., 2013).Transducing more HSCs should increase the numbers of recovered barcodes and lower some

of the variability between animals due to more evenly marked grafts. Nevertheless, the competitive situations between vector constructs still would be the same and dominant clones in this work emerged unaffected from barcode numbers. This should be considered for further barcoding experiments, especially when creating competitive situations.

Overall, I can conclude that in the transplantation setup investigated here, at least four out of five vector constructs tested showed comparable clonal dynamics and/or trends in all analyses. A final assessment for the SFFV promoter cannot be done, as the data is not conclusive. Consequently, all of these constructs appear suitable for observing "neutral" reconstitution of the haematopoietic system after transplantation. Still, our data demonstrates that vector constructs without any promoter or transgene expression and neutral integration patterns may mark cells, which develop clonal dominance. However, this clonal dominance appears to be related to intrinsic cellular features rather than the vector construct or marking procedure.

9.1.1. Calculating the amount of cells driving haematopoiesis post transplantation

Using the data collected within the project presented, I should be able to estimate the number of, active, repopulating (stem) cells within our haematopoietic model. Due to the variations observed between individual animals as well as some necessary estimations, a certain level of blur is included in the following calculation. Consequently, the numbers calculated here represent more the magnitude, than the actual values. The calculation can be found in Appendix figure 6, page 108.

As I could not detect any major differences regarding the vector constructs or promoters in previous analyses, data from all animals and all groups was pooled for this calculation. Some data from MG4.2 was excluded due to the already discussed global effect in this group lowering barcode diversity (page 75). In the beginning, the number of cells with intermediate/long-term engraftment potential within our 400,000 lineage-negative graft needs to be determined. The literature is not conducive, as even frequencies for the lineage-negative fraction within bone marrow vary strongly, depending on protocols used, mouse strain, age and purification strategy. There are reports of 8% lineage-negative cells within total bone marrow (Mayle et al., 2013) while others report only 1.25% (Kumar et al., 2008). Due to this variety, I will use the percentages observed in our own experiments, sampled from ~8 week-old Bl6 males (unpublished data) to get comparable data to the donors used in this work. In bone marrow, I observed about 2.2% lineage-negative cells of which around 20% were $Sca1^+$ $cKit^+$ (LSK). These LSK cells are enriched for haematopoietic stem and progenitor cells and were described to contain ~10% true long-term HSCs (Challen et al., 2009; Okada et al., 1992). However, with our barcode data, we can assess cells contributing to haematopoiesis between 6 weeks after transplantation and the final analysis time point (8 – 12 months post transplantation), which includes contribution of short-, intermediate- as well as long-term HSCs. These cells can be characterised via expression of CD150 (Challen et al., 2009; Doulatov et al., 2012; Oguro et al., 2013). Our data (unpublished) indicate

that around 25% of LSK cells are CD150$^+$ (SLAM$^+$) which, for the upcoming calculations, will serve as an engraftment-capability marker. Based on these values, our 400,000 lineage-negative cell graft can be supposed to contain around 19,800 cells capable of engrafting. To avoid mixing up terms and for brevity, these cells are called potentially engraftment-capable cells (pECC) from now on (Figure 35).

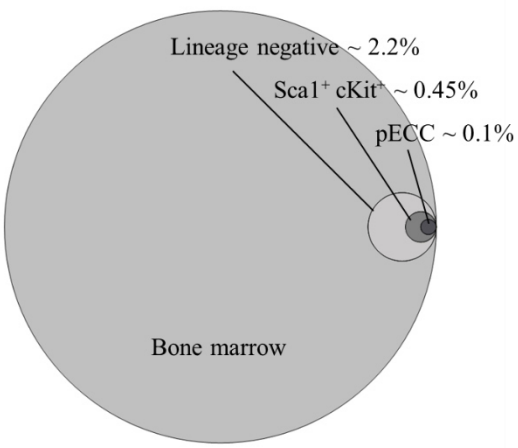


Figure 35 – Representation of the frequencies within the bone marrow for different subpopulations. The indicated percentages were obtained in an independent experiment (unpublished) and serve as scaffolding for the calculations presented here. pECC: potentially engraftment-capable cells, LSK CD150$^+$.

With the transduction rates presented previously (Figure 18, page 55, promoter-deprived constructs were equalized to GFP) we expect that we transduced ~1800 pECC within each graft with Lenti-EFS-GFP vectors in MG1 and MG2. The same amount of marked pECC would be assumed for Venus and TSapphire constructs, while SFFV-BFP, due to the higher transduction rate, should yield ~3300 marked cells per graft. Comparing these values to the total numbers of barcodes recovered (Figure 26, page 63) we only recover an average of 8% of these theoretically marked cells over all groups. Consequently, we lost 92% of our pECCs due to non-engraftment, dormancy, sample size or exhaustion/disappearance before the first sample time point. The literature reports homing frequencies of 10 to 20% (Camargo et al., 2006; van der Loo and Ploemacher, 1995) for transplantations done with purified HSCs with an colony-forming or limiting-dilution readout. Those values are slightly higher than the ones we observed, but this may be just due to the different experimental setups or different definition of HSC and pECC. The model presented here assumes that one barcode represents one pECC. As indicated before, there are probably some clones harbouring multiple barcode integrations. Consequently, this model may even overestimate the number of engrafted cells. As discussed before, there is no clear-cut way to correct our dataset for multiple-barcode clones using the NGS data or available gDNA. Therefore, I will continue building this model with the observed engraftment frequency of 8% and adjust the number of pECCs expected to engraft in our animals accordingly by factor 0.08. The fate of the cells disappearing can not be further assessed, thus it remains unclear if they did not engraft based on pure stochastics, lost their engraftment potential due to intrinsic properties, exhausted before the 6-week sample or just were dormant within the bone marrow. The successfully engrafted and active cells are called engraftment-capable cells (ECCs) from now on. Correlation of these expected ECCs with the mean number of barcodes observed 6 weeks after

transplantation (PB6w) showed a mean of 21.7% transduced ECCs actively contributing to the barcoded fraction found in peripheral blood at that time point. After several months, this number dropped to a mean of 16.2% (PB at final analysis). Based on the transduction rates, I was able to calculate absolute numbers of unmarked ECCs that should be active at these time points, which then can be summed up to the marked fraction. This reveals a contribution of roughly 350 active ECCs after six weeks and 260 ECCs eight to twelve months after transplantation (Figure 36). Admittedly, not all of them can be biologically relevant, as our calculations include a contribution threshold of 0.5% per vector backbone and thus only 0.16 – 0.25% of total blood, depending on the number of vector backbones. I could correct for a more biologically relevant threshold, but decided against adding further variables. This is primarily due to the observed high variabilities between individual animals. Using mean values from all animals for this calculation already adds some level of blur to my calculation. Furthermore, the purpose of this calculation was getting a dimension, rather than absolute values.

Busch et al. (2015) used a tamoxifen-inducible *in-situ* system of steady-state haematopoiesis and calculated that approximately 150 long-term HSCs per day are necessary to feed the short-term HSC pool, which in their work was the primary source of haematopoietic maintenance. Higher numbers of at least 400 active HSCs in steady-state haematopoiesis were reported by Zavidij et al. (2012), who used lentiviral vectors for *in-vivo* labelling. Our calculated 260 active ECCs are right in the middle. Naturally, steady-state haematopoiesis and a transplantation setting do have important differences (reviewed by Busch and Rodewald, 2016). Sun et al. (2014) postulate that the composition of clones in steady-state haematopoiesis changes over time, while clones in a transplantation setting are (to a certain degree) stable over time, the latter matching our observations presented in this work. Unfortunately, comparison of our calculated ECC values with reports from other transplantation setups is hampered by various differences in settings, transplanted cell types, numbers, observation period and lineage-tracking method. In addition, most publications only mention the number of recovered barcodes or insertion sites. Typically, the order of 30 – 100 (Naik et al., 2013; Verovskaya et al., 2013, 2014) reconstituting barcodes were retrieved after transplantation within individual samples, depending on filter parameters and thresholds. Those numbers are similar to the ones reported earlier in this work, if all clones for the different constructs within single animals were added up (Figure 27, page 65).

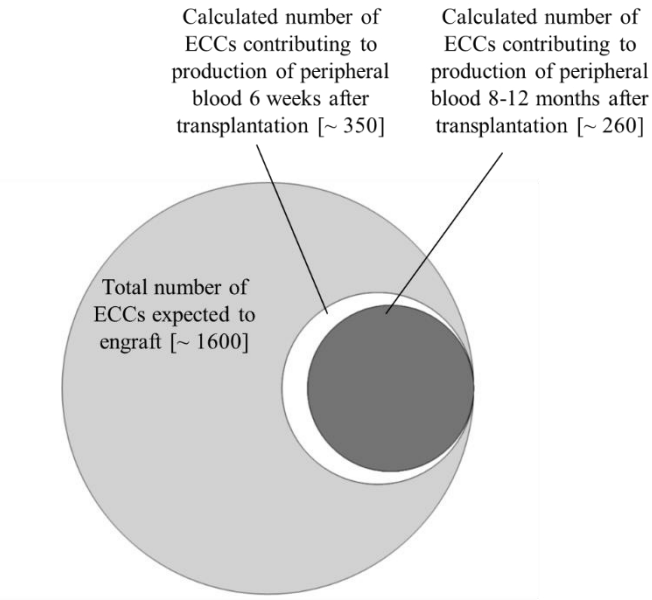


Figure 36 – Calculated numbers of active ECCs. Referencing transduction rates, the number of recovered barcodes over time and the frequencies of ECC within the bone marrow, I was able to estimate absolute numbers of cells contributing to production of peripheral blood at chosen time points. ECC: engraftment capable cells.

In summary, the data collected with our BC32 system allow to estimate the number of haematopoietically active cells at different time points after transplantation. According to my calculations, around 350 cells supply haematopoiesis 6 weeks after transplantation. At the final analysis time point, 8 – 12 months post transplantation, around 260 cells seem to be active.

In conclusion, the proposed barcoding stratefgy facilitates in-depth analysis of haematopoietic reconstitution after stem cell transplantation.

10. Publications

The following, chronologically ordered, publications are directly related to the barcode system presented in this thesis:

Roh, V., Abramowski, P., Hiou-Feige, A., Cornils, K., Rivals, J.-P., Truan, Z., Mermod, M., Monnier, Y., Prassolov, V., **Aranyossy, T.**, Fehse, B., Tolstonog, G. V. and Simon, C. (2017). High complexity cellular barcoding identifies clonal sweep as a hallmark of local recurrence in a surgical head and neck cancer mouse model. Manuscript submitted.

Cornils, K., Thielecke, L., Winkelmann, D., **Aranyossy, T.**, Lesche, M., Dahl, A., Roeder, I., Fehse, B., and Glauche, I. (2017). Clonal competition in BcrAbl-driven leukemia: how transplantations can accelerate clonal conversion. Mol. Cancer., accepted for publication on 24th May 2017

Thielecke, L., **Aranyossy, T.**, Dahl, A., Tiwari, R., Roeder, I., Geiger, H., Fehse, B., Glauche, I., and Cornils, K. (2017). Limitations and challenges of genetic barcode quantification. Sci. Rep. *7*, 43249.

Bigildeev, A.E., Cornils, K., **Aranyossy, T.**, Sats, N. V., Petinati, N.A., Shipounova, I.N., Surin, V.L., Pshenichnikova, O.S., Riecken, K., Fehse, B., et al. (2016). Investigation of the Mesenchymal Stem Cell Compartment by Means of a Lentiviral Barcode Library. Biochem. Biokhimiia *81*, 373–381.

I further contributed to the following publications, although their topic is not directly related to this thesis:

Mohme, M., Maire, C.L., Riecken, K., Zapf, S., **Aranyossy, T.**, Westphal, M., Lamszus, K., and Fehse, B. (2017). Optical Barcoding for Single-Clone Tracking to Study Tumor Heterogeneity. Mol. Ther. *0*, 306–313.

Badbaran, A., Fehse, B., Christopeit, M., **Aranyossy, T.**, Ayuk, F.A., Wolschke, C., and Kröger, N. (2016). Digital-PCR assay for screening and quantitative monitoring of calreticulin (CALR) type-2 positive patients with myelofibrosis following allogeneic stem cell transplantation. Bone Marrow Transplant. *51*, 872–873.

11. Declaration in lieu of an oath / Eidesstattliche Versicherung

I confirm that I wrote this dissertation on my own, without using any other than the declared sources, references and tools.

Hiermit erkläre ich an Eides statt, dass ich die vorliegende Dissertationsschrift selbst verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

Hamburg, den 09.06.2017

Tim Aranyossy

Signature/Unterschrift

Contact: Tim.Aranyossy@gmx.de

12. Bestätigung der Korrektheit der Sprache

UKE HAMBURG
Interdisziplinäre Klinik und Poliklinik für Stammzelltransplantation

Prof. Dr. med. Nicolaus Kröger
Direktor

Onkologisches Zentrum

Martinistraße 52, Gebäude O 24
20246 Hamburg

Forschungsabteilung
Zell- und Gentherapie
Prof. Dr. Boris Fehse
fehse@uke.de

Universitätsklinikum Hamburg-Eppendorf | Martinistraße 52 | 20246 Hamburg
Interdisziplinäre Klinik und Poliklinik für Stammzelltransplantation

Studienbüro Biologie
z.H. Frau Sült-Wüpping
MIN Fakultät
Universität Hamburg
Biozentrum Klein Flottbek
Ohnhorststr. 18
22609 Hamburg

Hamburg, 23.05.2017

Bestätigung der Korrektheit der Sprache

Sehr geehrte Damen und Herren,

hiermit bestätige ich, dass die von Herr Tim Dominic Aranyossy mit dem Titel "Using barcode vectors for neutral genetic marking to study clonal dynamics of hematopoietic reconstitution " vorgelegte Doktorarbeit in korrektem Englisch geschrieben ist.

Mit freundlichen Grüßen,

Carol Stocking

Dr. Carol Stocking
Klinik für Stammzelltransplantation
Universitätsklinikum Hamburg-Eppendorf
(Amerikanerin)
Email: c.stocking@uke.de

13. Bibliography

Adolfsson, J., Månsson, R., Buza-Vidas, N., Hultquist, A., Liuba, K., Jensen, C.T., Bryder, D., Yang, L., Borge, O.-J., Thoren, L.A.M., et al. (2005). Identification of Flt3+ lympho-myeloid stem cells lacking erythro-megakaryocytic potential a revised road map for adult blood lineage commitment. Cell *121*, 295–306.

Aiuti, A., Biasco, L., Scaramuzza, S., Ferrua, F., Cicalese, M.P., Baricordi, C., Dionisio, F., Calabria, A., Giannelli, S., Castiello, M.C., et al. (2013). Lentiviral hematopoietic stem cell gene therapy in patients with Wiskott-Aldrich syndrome. Science (80-.). *341*, 1233151.

Akashi, K., Traver, D., Miyamoto, T., and Weissman, I.L. (2000). A clonogenic common myeloid progenitor that gives rise to all myeloid lineages. Nature *404*, 193–197.

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

Arumugam, P.I., Higashimoto, T., Urbinati, F., Modlich, U., Nestheide, S., Xia, P., Fox, C., Corsinotti, A., Baum, C., and Malik, P. (2009). Genotoxic Potential of Lineage-specific Lentivirus Vectors Carrying the β-Globin Locus Control Region. Mol. Ther. *17*, 1929–1937.

Badbaran, A., Fehse, B., Christopeit, M., Aranyossy, T., Ayuk, F.A., Wolschke, C., and Kröger, N. (2016). Digital-PCR assay for screening and quantitative monitoring of calreticulin (CALR) type-2 positive patients with myelofibrosis following allogeneic stem cell transplantation. Bone Marrow Transplant. *51*, 872–873.

Baens, M., Noels, H., Broeckx, V., Hagens, S., Fevery, S., Billiau, A.D., Vankelecom, H., and Marynen, P. (2006). The Dark Side of EGFP: Defective Polyubiquitination. PLoS One *1*, e54.

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

Barre-Sinoussi, F., Chermann, J., Rey, F., Nugeyre, M., Chamaret, S., Gruest, J., Dauguet, C., Axler-Blin, C., Vezinet-Brun, F., Rouzioux, C., et al. (1983). Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS). Science (80-.). *220*, 868–871.

Barry, S.C., Harder, B., Brzezinski, M., Flint, L.Y., Seppen, J., and Osborne, W.R.A. (2001). Lentivirus vectors encoding both central polypurine tract and posttranscriptional regulatory element provide enhanced transduction and transgene expression. Hum. Gene Ther. *12*, 1103–1108.

Bartz, S.R., and Vodicka, M. a (1997). Production of high-titer human immunodeficiency virus type 1 pseudotyped with vesicular stomatitis virus glycoprotein. Methods *12*, 337–342.

Basu, V.P., Song, M., Gao, L., Rigby, S.T., Hanson, M.N., and Bambara, R.A. (2008). Strand transfer events during HIV-1 reverse transcription. Virus Res. *134*, 19–38.

Baum, C. (2011). Prävention der Insertionsmutagenese. Unvermeidbar oder beherrschbar? Pharm. Unserer Zeit *40*, 248–252.

Baum, C., Hegewisch-Becker, S., Eckert, H.G., Stocking, C., and Ostertag, W. (1995). Novel retroviral vectors for efficient expression of the multidrug resistance (mdr-1) gene in early hematopoietic cells. J. Virol. *69*, 7541–7547.

Baum, C., Düllmann, J., Li, Z., Fehse, B., Meyer, J., Williams, D.A., and von Kalle, C. (2003). Side effects of retroviral gene transfer into hematopoietic stem cells. Blood *101*, 2099–2114.

Baum, C., von Kalle, C., Staal, F.J.T., Li, Z., Fehse, B., Schmidt, M., Weerkamp, F., Karlsson, S., Wagemaker, G., and Williams, D.A. (2004). Chance or necessity? Insertional mutagenesis in gene therapy and its consequences. Mol. Ther. *9*, 5–13.

Becker, A.J., McCulloch, E.A., and Till, J.E. (1963). Cytological Demonstration of the Clonal Nature of Spleen Colonies Derived from Transplanted Mouse Marrow Cells. Nature *197*, 452–454.

Benveniste, P., Frelin, C., Janmohamed, S., Barbara, M., Herrington, R., Hyam, D., and Iscove, N.N. (2010). Intermediate-term hematopoietic stem cells with extended but time-limited reconstitution potential. Cell Stem Cell *6*, 48–58.

Benz, C., Copley, M.R., Kent, D.G., Wohrer, S., Cortes, A., Aghaeepour, N., Ma, E., Mader, H., Rowe, K., Day, C., et al. (2012). Hematopoietic stem cell subtypes expand differentially during development and display distinct lymphopoietic programs. Cell Stem Cell *10*, 273–283.

Bhang, H.C., Ruddy, D.A., Krishnamurthy Radhakrishna, V., Caushi, J.X., Zhao, R., Hims, M.M., Singh, A.P., Kao, I., Rakiec, D., Shaw, P., et al. (2015). Studying clonal dynamics in response to cancer therapy using high-complexity barcoding. Nat. Med. *21*, 440–448.

Biffi, A., Bartolomae, C.C., Cesana, D., Cartier, N., Aubourg, P., Ranzani, M., Cesani, M., Benedicenti, F., Plati, T., Rubagotti, E., et al. (2011). 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.

Bigildeev, A.E., Cornils, K., Aranyossy, T., Sats, N. V., Petinati, N.A., Shipounova, I.N., Surin, V.L., Pshenichnikova, O.S., Riecken, K., Fehse, B., et al. (2016). Investigation of the Mesenchymal Stem Cell Compartment by Means of a Lentiviral Barcode Library. Biochem. Biokhimiia *81*, 373–381.

Bittner, J.J. (1936). SOME POSSIBLE EFFECTS OF NURSING ON THE MAMMARY GLAND TUMOR INCIDENCE IN MICE. Science *84*, 162.

Blaese, R.M., Culver, K.W., Miller, A.D., Carter, C.S., Fleisher, T., Clerici, M., Shearer, G., Chang, L., Chiang, Y., Tolstoshev, P., et al. (1995). T lymphocyte-directed gene therapy for ADA- SCID: initial trial results after 4 years. Science *270*, 475–480.

Blundell, J.R., and Levy, S.F. (2014). Beyond genome sequencing: Lineage tracking with barcodes to study the dynamics of evolution, infection, and cancer. Genomics *104*, 417–430.

Braun, C.J., Boztug, K., Paruzynski, A., Witzel, M., Schwarzer, A., Rothe, M., Modlich, U., Beier, R., Göhring, G., Steinemann, D., et al. (2014). Gene therapy for Wiskott-Aldrich syndrome--long-term efficacy and genotoxicity. Sci. Transl. Med. *6*, 227ra33.

Breckpot, K., Aerts, J.L., and Thielemans, K. (2007). Lentiviral vectors for cancer immunotherapy: transforming infectious particles into therapeutics. Gene Ther. *14*, 847–862.

Brenner, S., Whiting-Theobald, N.L., Linton, G.F., Holmes, K.L., Anderson-Cohen, M., Kelly, P.F., Vanin, E.F., Pilon, A.M., Bodine, D.M., Horwitz, M.E., et al. (2003). 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.

Brewer, C., Chu, E., Chin, M., and Lu, R. (2016). Transplantation Dose Alters the Differentiation Program of Hematopoietic Stem Cells. Cell Rep. *15*, 1–10.

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

Busch, K., and Rodewald, H.-R. (2016). Unperturbed vs. post-transplantation hematopoiesis: both in vivo but different. Curr. Opin. Hematol. *23*, 295–303.

Busch, K., Klapproth, K., Barile, M., Flossdorf, M., Holland-Letz, T., Schlenner, S.M., Reth, M., Höfer, T., and Rodewald, H.-R. (2015). Fundamental properties of unperturbed haematopoiesis from stem cells in vivo. Nature *518*, 542–546.

Buschmann, T., and Bystrykh, L. V (2013). Levenshtein error-correcting barcodes for multiplexed DNA sequencing. BMC Bioinformatics *14*, 272.

Bush, D.L., and Vogt, V.M. (2014). In Vitro Assembly of Retroviruses. Annu. Rev. Virol. *1*, 561–580.

Bushman, F., Lewinski, M., Ciuffi, A., Barr, S., Leipzig, J., Hannenhalli, S., and Hoffmann, C. (2005). Genome-wide analysis of retroviral DNA integration. Nat. Rev. Microbiol. *3*, 848–858.

Bystrykh, L. V. (2012). Generalized DNA Barcode Design Based on Hamming Codes. PLoS One *7*, e36852.

Bystrykh, L. V., and Belderbos, M.E. (2016). Clonal Analysis of Cells with Cellular Barcoding: When Numbers and Sizes Matter. In Methods in Molecular Biology, pp. 257–284.

Bystrykh, L. V, Verovskaya, E., Zwart, E., Broekhuis, M., and de Haan, G. (2012). Counting stem cells: methodological constraints. Nat. Methods *9*, 567–574.

Camargo, F.D., Chambers, S.M., Drew, E., McNagny, K.M., and Goodell, M.A. (2006). Hematopoietic stem cells do not engraft with absolute efficiencies. Blood *107*, 501–507.

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

Carlson, C.M., and Largaespada, D. a (2005). Insertional mutagenesis in mice: new perspectives and tools. Nat. Rev. Genet. *6*, 568–580.

Cattoglio, C., Pellin, D., Rizzi, E., Maruggi, G., Corti, G., Miselli, F., Sartori, D., Guffanti, A., Di Serio, C., Ambrosi, A., et al. (2010). High-definition mapping of retroviral integration sites identifies active regulatory elements in human multipotent hematopoietic progenitors. Blood *116*, 5507–5517.

Cavazza, A., Moiani, A., and Mavilio, F. (2013). Mechanisms of Retroviral Integration and Mutagenesis. Hum. Gene Ther. *24*, 119–131.

Cavazzana-Calvo, M., Hacein-Bey, S., de Saint Basile, G., Gross, F., Yvon, E., Nusbaum, P., Selz, F., Hue, C., Certain, S., Casanova, J.L., et al. (2000). Gene therapy of human severe combined immunodeficiency (SCID)-X1 disease. Science *288*, 669–672.

Cavazzana-Calvo, M., Payen, E., Negre, O., Wang, G., Hehir, K., Fusil, F., Down, J., Denaro, M., Brady, T., Westerman, K., et al. (2010). Transfusion independence and HMGA2 activation after gene therapy of human β-thalassaemia. Nature *467*, 318–322.

Challen, G.A., Boles, N., Lin, K.-Y.K.-Y., and Goodell, M.A. (2009). Mouse hematopoietic stem cell identification and analysis. Cytom. Part A *75A*, 14–24.

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

Cheung, A.M.S., Nguyen, L. V., Carles, A., Beer, P., Miller, P.H., Knapp, D.J.H.F., Dhillon, K., Hirst, M., and Eaves, C.J. (2013). Analysis of the clonal growth and differentiation dynamics of primitive barcoded human cord blood cells in NSG mice. Blood *122*, 3129–3137.

Christensen, J.L., and Weissman, I.L. (2001). Flk-2 is a marker in hematopoietic stem cell differentiation: A simple method to isolate long-term stem cells. Proc. Natl. Acad. Sci. *98*, 14541–14546.

Cornils, K., Lange, C., Schambach, A., Brugman, M.H., Nowak, R., Lioznov, M., Baum, C., and Fehse, B. (2009). Stem Cell Marking With Promotor-deprived Self-inactivating Retroviral Vectors Does Not Lead to Induced Clonal Imbalance. Mol. Ther. *17*, 131–143.

Cornils, K., Thielecke, L., Huser, S., Forgber, M., Thomaschewski, M., Kleist, N., Hussein, K., Riecken, K., Volz, T., Gerdes, S., et al. (2014). Multiplexing clonality: combining RGB marking and genetic barcoding. Nucleic Acids Res. *42*, e56–e56.

Cornils, K., Thielecke, L., Winkelmann, D., Aranyossy, T., Lesche, M., Dahl, A., Roeder, I., Fehse, B., and Glauche, I. (2017). Clonal competition in BcrAbl-driven leukemia: how transplantations can accelerate clonal conversion. Mol. Cancer.

Craigie, R., and Bushman, F.D. (2014). Host Factors in Retroviral Integration and the Selection of Integration Target Sites. Microbiol. Spectr. *2*, 1367–1671.

Cronin, J., Zhang, X.-Y., and Reiser, J. (2005). Altering the tropism of lentiviral vectors through pseudotyping. Curr. Gene Ther. *5*, 387–398.

Deakin, C.T., Deakin, J.J., Ginn, S.L., Young, P., Humphreys, D., Suter, C.M., Alexander, I.E., and Hallwirth, C. V. (2014). Impact of next-generation sequencing error on analysis of barcoded plasmid libraries of known complexity and sequence. Nucleic Acids Res. *42*, e129–e129.

Derse, D., Crise, B., Li, Y., Princler, G., Lum, N., Stewart, C., McGrath, C.F., Hughes, S.H., Munroe, D.J., and Wu, X. (2007). Human T-Cell Leukemia Virus Type 1 Integration Target Sites in the Human Genome: Comparison with Those of Other Retroviruses. J. Virol. *81*, 6731–6741.

Dick, J.E., Magli, M.C., Huszar, D., Phillips, R.A., and Bernstein, A. (1985). 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.

Doulatov, S., Notta, F., Eppert, K., Nguyen, L.T., Ohashi, P.S., and Dick, J.E. (2010). Revised map of the human progenitor hierarchy shows the origin of macrophages and dendritic cells in early lymphoid development. Nat. Immunol.

11, 585–593.

Doulatov, S., Notta, F., Laurenti, E., and Dick, J.E. (2012). Hematopoiesis: A Human Perspective. Cell Stem Cell *10*, 120–136.

Dufait, I., Liechtenstein, T., Lanna, A., Laranga, R., Padella, A., Bricogne, C., Arce, F., Kochan, G., Breckpot, K., and Escors, D. (2013). Lentiviral Vectors in Immunotherapy. In Gene Therapy - Tools and Potential Applications, (InTech), p.

Dull, T., Zufferey, R., Kelly, M., Mandel, R.J., Nguyen, M., Trono, D., and Naldini, L. (1998). A third-generation lentivirus vector with a conditional packaging system. J. Virol. *72*, 8463–8471.

Dykstra, B., Kent, D., Bowie, M., McCaffrey, L., Hamilton, M., Lyons, K., Lee, S., Brinkman, R., and Eaves, C. (2007). Long-Term Propagation of Distinct Hematopoietic Differentiation Programs In Vivo. Cell Stem Cell *1*, 218–229.

Dykstra, B., Olthof, S., Schreuder, J., Ritsema, M., and de Haan, G. (2011). Clonal analysis reveals multiple functional defects of aged murine hematopoietic stem cells. J. Exp. Med. *208*, 2691–2703.

Ellermann, V., and Bang, O. (1908). Experimentelle Leukämie bei Hühnern. Zentralbl. Bakteriol. Parasitenkd. Infect. Hyg. Abt. I. *46*, 595–609.

Engelman, A., and Cherepanov, P. (2008). The Lentiviral Integrase Binding Protein LEDGF/p75 and HIV-1 Replication. PLoS Pathog. *4*, e1000046.

Escors, D., and Breckpot, K. (2010). Lentiviral Vectors in Gene Therapy: Their Current Status and Future Potential. Arch. Immunol. Ther. Exp. (Warsz). *58*, 107–119.

Fehse, B., Kustikova, O.S., Bubenheim, M., and Baum, C. (2004). Pois(s)on – It's a Question of Dose…. Gene Ther. *11*, 879–881.

Felice, B., Cattoglio, C., Cittaro, D., Testa, A., Miccio, A., Ferrari, G., Luzi, L., Recchia, A., and Mavilio, F. (2009). Transcription Factor Binding Sites Are Genetic Determinants of Retroviral Integration in the Human Genome. PLoS One *4*, e4571.

Freed, E.O. (2015). HIV-1 assembly, release and maturation. Nat. Rev. Microbiol. *13*, 484–496.

Friedmann, T., and Roblin, R. (1972). Gene Therapy for Human Genetic Disease? Science (80-.). *175*, 949–955.

Gabriel, R., Eckenberg, R., Paruzynski, A., Bartholomae, C.C., Nowrouzi, A., Arens, A., Howe, S.J., Recchia, A., Cattoglio, C., Wang, W., et al. (2009). Comprehensive genomic access to vector integration in clinical gene therapy. Nat. Med. *15*, 1431–1436.

Gerrits, A., Dykstra, B., Kalmykowa, O.J., Klauke, K., Verovskaya, E., Broekhuis, M.J.C., de Haan, G., and Bystrykh, L. V. (2010). Cellular barcoding tool for clonal analysis in the hematopoietic system. Blood *115*, 2610–2618.

Golden, J.A., Fields-Berry, S.C., and Cepko, C.L. (1995). Construction and characterization of a highly complex retroviral library for lineage analysis. Proc. Natl. Acad. Sci. *92*, 5704–5708.

Grosselin, J., Sii-Felice, K., Payen, E., Chretien, S., Tronik-Le Roux, D., and Leboulch, P. (2013). Arrayed lentiviral barcoding for quantification analysis of hematopoietic dynamics. Stem Cells *31*, 2162–2171.

Hacein-Bey-Abina, S., Von Kalle, C., Schmidt, M., McCormack, M.P., Wulffraat, N., Leboulch, P., Lim, A., Osborne, C.S., Pawliuk, R., Morillon, E., et al. (2003a). LMO2-associated clonal T cell proliferation in two patients after gene therapy for SCID-X1. Science (80-.). *302*, 415–419.

Hacein-Bey-Abina, S., von Kalle, C., Schmidt, M., Le Deist, F., Wulffraat, N., McIntyre, E., Radford, I., Villeval, J.-L., Fraser, C.C., Cavazzana-Calvo, M., et al. (2003b). A Serious Adverse Event after Successful Gene Therapy for X-Linked Severe Combined Immunodeficiency. N. Engl. J. Med. *348*, 255–256.

Hacein-Bey-Abina, S., Garrigue, A., Wang, G.P., Soulier, J., Lim, A., Morillon, E., Clappier, E., Caccavelli, L., Delabesse, E., Beldjord, K., et al. (2008). Insertional oncogenesis in 4 patients after retrovirus-mediated gene therapy of SCID-X1. J. Clin. Invest. *118*, 3132–3142.

Haeckel, E. (1868). Natürliche Schöpfungsgeschichte. Gemeinverständliche wissenschaftliche Vorträge über die Entwickelungslehre im Allgemeinen und diejenige von Darwin, Goethe und Lamarck und Besonderen. (Reimer).

Hahn, W.C., and Weinberg, R.A. (2002). Rules for making human tumor cells. N. Engl. J. Med. *347*, 1593–1603.

Hamming, R.W. (1950). Error Detecting and Error Correcting Codes. Bell Syst. Tech. J. *29*, 147–160.

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

Hayward, A., Grabherr, M., and Jern, P. (2013). Broad-scale phylogenomics provides insights into retrovirus-host evolution. Proc. Natl. Acad. Sci. U. S. A. *110*, 20146–20151.

Herbst, F., Ball, C.R., Tuorto, F., Nowrouzi, A., Wang, W., Zavidij, O., Dieter, S.M., Fessler, S., van der Hoeven, F., Kloz, U., et al. (2012). Extensive Methylation of Promoter Sequences Silences Lentiviral Transgene Expression During Stem Cell Differentiation In Vivo. Mol. Ther. *20*, 1014–1021.

Hill, M.O. (1973). Diversity and Evenness: A Unifying Notation and Its Consequences. Ecology *54*, 427–432.

Hindson, B.J., Ness, K.D., Masquelier, D.A., Belgrader, P., Heredia, N.J., Makarewicz, A.J., Bright, I.J., Lucero, M.Y., Hiddessen, A.L., Legler, T.C., et al. (2011). High-Throughput Droplet Digital PCR System for Absolute Quantitation of DNA Copy Number. Anal. Chem. *83*, 8604–8610.

Höfer, T., Barile, M., and Flossdorf, M. (2016). Stem-cell dynamics and lineage topology from in vivo fate mapping in the hematopoietic system. Curr. Opin. Biotechnol. *39*, 150–156.

Hu, W.-S., and Hughes, S.H. (2012). HIV-1 reverse transcription. Cold Spring Harb. Perspect. Med. *2*, a006882–a006882.

Hughes, S.H. (2004). The RCAS vector system. Folia Biol. (Praha). *50*, 107–119.

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

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

International Human Genome Sequencing Consortium, Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., et al. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921.

Jacobson, L.O., Simmons, E.L., Marks, E.K., and Eldredge, J.H. (1951). Recovery from Radiation Injury. Science (80-.). *113*, 510–511.

Jordan, C.T., and Lemischka, I.R. (1990). Clonal and systemic analysis of long-term hematopoiesis in the mouse. Genes Dev. *4*, 220–232.

Kafri, T., van Praag, H., Gage, F.H., and Verma, I.M. (2000). Lentiviral vectors: regulated gene expression. Mol. Ther. *1*, 516–521.

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

Klauke, K., Broekhuis, M.J.C., Weersing, E., Dethmers-Ausema, A., Ritsema, M., González, M.V., Zwart, E., Bystrykh, L. V., and de Haan, G. (2015). Tracing Dynamics and Clonal Heterogeneity of Cbx7-Induced Leukemic Stem Cells by Cellular Barcoding. Stem Cell Reports *4*, 74–89.

Kohn, D.B., Sadelain, M., Dunbar, C., Bodine, D., Kiem, H.-P., Candotti, F., Tisdale, J., Riviére, I., Blau, C.A., Richard, R.E., et al. (2003). American Society of Gene Therapy (ASGT) ad hoc subcommittee on retroviral-mediated gene transfer to hematopoietic stem cells. Mol. Ther. *8*, 180–187.

Kondo, M., Weissman, I.L., and Akashi, K. (1997). Identification of Clonogenic Common Lymphoid Progenitors in Mouse Bone Marrow. Cell *91*, 661–672.

Kondo, M., Wagers, A.J., Manz, M.G., Prohaska, S.S., Scherer, D.C., Beilhack, G.F., Shizuru, J.A., and Weissman, I.L. (2003). Biology of hematopoietic stem cells and progenitors: implications for clinical application. Annu. Rev. Immunol. *21*, 759–806.

Krause, D.S., Theise, N.D., Collector, M.I., Henegariu, O., Hwang, S., Gardner, R., Neutzel, S., and Sharkis, S.J. (2001). Multi-organ, multi-lineage engraftment by a single bone marrow-derived stem cell. Cell *105*, 369–377.

Krueger, A., Ziętara, N., and Łyszkiewicz, M. (2016). T Cell Development by the Numbers. Trends Immunol. *xx*, 1–12.

Krueger, F., Andrews, S.R., and Osborne, C.S. (2011). Large scale loss of data in low-diversity illumina sequencing libraries can be recovered by deferred cluster calling. PLoS One *6*, e16607.

Kumar, R., Fossati, V., Israel, M., and Snoeck, H.-W. (2008). Lin-Sca1+Kit- Bone Marrow Cells Contain Early Lymphoid-Committed Precursors That Are Distinct from Common Lymphoid Progenitors. J. Immunol. *181*, 7507–7513.

Kustikova, O., Fehse, B., Modlich, U., Yang, M., Düllmann, J., Kamino, K., von Neuhoff, N., Schlegelberger, B., Li, Z., and Baum, C. (2005). Clonal dominance of hematopoietic stem cells triggered by retroviral gene marking. Science (80-.). *308*, 1171–1174.

Kustikova, O.S., Wahlers, A., Kuhlcke, K., Stahle, B., Zander, A.R., Baum, C., and Fehse, B. (2003). Dose finding with retroviral vectors: correlation of retroviral vector copy numbers in single cells with gene transfer efficiency in a cell population. Blood *102*, 3934–3937.

Lever, A., Gottlinger, H., Haseltine, W., and Sodroski, J. (1989). Identification of a sequence required for efficient packaging of human immunodeficiency virus type 1 RNA into virions. J. Virol. *63*, 4085–4087.

Lewinski, M.K., Yamashita, M., Emerman, M., Ciuffi, A., Marshall, H., Crawford, G., Collins, F., Shinn, P., Leipzig, J., Hannenhalli, S., et al. (2006). Retroviral DNA Integration: Viral and Cellular Determinants of Target-Site Selection. PLoS Pathog. *2*, e60.

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

Lewis, P., Hensel, M., and Emerman, M. (1992). Human immunodeficiency virus infection of cells arrested in the cell cycle. EMBO J. *11*, 3053–3058.

Li, Z., Düllmann, J., Schiedlmeier, B., Schmidt, M., von Kalle, C., Meyer, J., Forster, M., Stocking, C., Wahlers, A., Frank, O., et al. (2002). Murine leukemia induced by retroviral gene marking. Science (80-.). *296*, 497.

Lin, H.-T., Masaki, H., Yamaguchi, T., Wada, T., Yachie, A., Nishimura, K., Ohtaka, M., Nakanishi, M., Nakauchi, H., and Otsu, M. (2015). An assessment of the effects of ectopic gp91phox expression in XCGD iPSC-derived neutrophils. Mol. Ther. - Methods Clin. Dev. *2*, 15046.

Llano, M., Vanegas, M., Fregoso, O., Saenz, D., Chung, S., Peretz, M., and Poeschla, E.M. (2004). LEDGF/p75 Determines Cellular Trafficking of Diverse Lentiviral but Not Murine Oncoretroviral Integrase Proteins and Is a Component of Functional Lentiviral Preintegration Complexes. J. Virol. *78*, 9524–9537.

Lombardo, A., Cesana, D., Genovese, P., Di Stefano, B., Provasi, E., Colombo, D.F., Neri, M., Magnani, Z., Cantore, A., Lo Riso, P., et al. (2011). Site-specific integration and tailoring of cassette design for sustainable gene transfer. Nat. Methods *8*, 861–869.

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

Lu, R., Neff, N.F., Quake, S.R., and Weissman, I.L. (2011). Tracking single hematopoietic stem cells in vivo using high-throughput sequencing in conjunction with viral genetic barcoding. Nat. Biotechnol. *29*, 928–933.

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

Maetzig, T., Brugman, M.H., Bartels, S., Heinz, N., Kustikova, O.S., Modlich, U., Li, Z., Galla, M., Schiedlmeier, B., Schambach, A., et al. (2011). Polyclonal fluctuation of lentiviral vector-transduced and expanded murine hematopoietic stem cells. Blood *117*, 3053–3064.

Mann, R., Mulligan, R.C., and Baltimore, D. (1983). Construction of a retrovirus packaging mutant and its use to produce helper-free defective retrovirus. Cell *33*, 153–159.

Masumi, A. (2013). Hematopoietic Stem Cells and Response to Interferon. In Stem Cell Biology in Normal Life and Diseases, (InTech), p.

Matreyek, K.A., and Engelman, A. (2011). The requirement for nucleoporin NUP153 during human immunodeficiency virus type 1 infection is determined by the viral capsid. J. Virol. *85*, 7818–7827.

Maximow, A. (1909). Der Lymphozyt als gemeinsame Stammzelle der verschiedenen Blutelemente in der embryonalen

Entwicklung und im postfetalen Leben der Säugetiere. Folia Haematol. *8*, 125–134.

Maximow, A.A. (1906). Über experimentelle Erzeugung von Knochenmarks-Gewebe. Anat. Anz. *28*, 24–38.

Mayle, A., Luo, M., Jeong, M., and Goodell, M.A. (2013). Flow cytometry analysis of murine hematopoietic stem cells. Cytom. Part A *83A*, 27–37.

Mazarakis, N.D., Azzouz, M., Rohll, J.B., Ellard, F.M., Wilkes, F.J., Olsen, A.L., Carter, E.E., Barber, R.D., Baban, D.F., Kingsman, S.M., et al. (2001). Rabies virus glycoprotein pseudotyping of lentiviral vectors enables retrograde axonal transport and access to the nervous system after peripheral delivery. Hum. Mol. Genet. *10*, 2109–2121.

Mitchell, R.S., Beitzel, B.F., Schroder, A.R.W., Shinn, P., Chen, H., Berry, C.C., Ecker, J.R., and Bushman, F.D. (2004). Retroviral DNA Integration: ASLV, HIV, and MLV Show Distinct Target Site Preferences. PLoS Biol. *2*, e234.

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

Mizutani, S., and Temin, H.M. (1970). An RNA-Dependent DNA Polymerase in Virions of Rous Sarcoma Virus. Cold Spring Harb. Symp. Quant. Biol. *35*, 847–849.

Modlich, U., Kustikova, O.S., Schmidt, M., Rudolph, C., Meyer, J., Li, Z., Kamino, K., von Neuhoff, N., Schlegelberger, B., Kuehlcke, K., et al. (2005). Leukemias following retroviral transfer of multidrug resistance 1 (MDR1) are driven by combinatorial insertional mutagenesis. Blood *105*, 4235–4246.

Modlich, U., Bohne, J., Schmidt, M., von Kalle, C., Knoss, S., Schambach, A., and Baum, C. (2006). Cell-culture assays reveal the importance of retroviral vector design for insertional genotoxicity. Blood *108*, 2545–2553.

Modlich, U., Navarro, S., Zychlinski, D., Maetzig, T., Knoess, S., Brugman, M.H., Schambach, A., Charrier, S., Galy, A., Thrasher, A.J., et al. (2009). Insertional Transformation of Hematopoietic Cells by Self-inactivating Lentiviral and Gammaretroviral Vectors. Mol. Ther. *17*, 1919–1928.

Modrow, S., Falke, D., and Truyen, U. (2003). Molekulare Virologie, 2. Auflage.

Mohme, M., Maire, C.L., Riecken, K., Zapf, S., Aranyossy, T., Westphal, M., Lamszus, K., and Fehse, B. (2017). Optical Barcoding for Single-Clone Tracking to Study Tumor Heterogeneity. Mol. Ther. *0*, 306–313.

Moiani, A., Suerth, J., Gandolfi, F., Rizzi, E., Severgnini, M., De Bellis, G., Schambach, A., and Mavilio, F. (2014). Genome-Wide Analysis of Alpharetroviral Integration in Human Hematopoietic Stem/Progenitor Cells. Genes (Basel). *5*, 415–429.

Montini, E., Cesana, D., Schmidt, M., Sanvito, F., Bartholomae, C.C., Ranzani, M., Benedicenti, F., Sergi, L.S., Ambrosi, A., Ponzoni, M., et al. (2009). The genotoxic potential of retroviral vectors is strongly modulated by vector design and integration site selection in a mouse model of HSC gene therapy. J. Clin. Invest. *119*, 964–975.

Mooslehner, K., Karls, U., and Harbers, K. (1990). Retroviral integration sites in transgenic Mov mice frequently map in the vicinity of transcribed DNA regions. J. Virol. *64*, 3056–3058.

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

Mueller, P.R., and Wold, B. (1990). In Vivo Footprinting of a Muscle Specific Enhancer by Ligation Mediated PCR. Science (80-.). *246*, 802–802.

Naik, S.H., Perié, L., Swart, E., Gerlach, C., van Rooij, N., de Boer, R.J., and Schumacher, T.N. (2013). Diverse and heritable lineage imprinting of early haematopoietic progenitors. Nature *496*, 229–232.

Naldini, L. (2015). Gene therapy returns to centre stage. Nature *526*, 351–360.

Naldini, L., Blomer, U., Gallay, P., Ory, D., Mulligan, R., Gage, F.H., Verma, I.M., and Trono, D. (1996). In Vivo Gene Delivery and Stable Transduction of Nondividing Cells by a Lentiviral Vector. Science (80-.). *272*, 263–267.

Narezkina, A., Taganov, K.D., Litwin, S., Stoyanova, R., Hayashi, J., Seeger, C., Skalka, A.M., and Katz, R. a (2004). Genome-Wide Analyses of Avian Sarcoma Virus Integration Sites. J. Virol. *78*, 11656–11663.

Nguyen, L. V., Makarem, M., Carles, A., Moksa, M., Kannan, N., Pandoh, P., Eirew, P., Osako, T., Kardel, M., Cheung, A.M.S., et al. (2014). Clonal Analysis via Barcoding Reveals Diverse Growth and Differentiation of Transplanted Mouse and Human Mammary Stem Cells. Cell Stem Cell *14*, 253–263.

Ocwieja, K.E., Brady, T.L., Ronen, K., Huegel, A., Roth, S.L., Schaller, T., James, L.C., Towers, G.J., Young, J.A.T., Chanda, S.K., et al. (2011). HIV integration targeting: a pathway involving Transportin-3 and the nuclear pore protein RanBP2. PLoS Pathog. *7*, e1001313.

Ogert, R.A., Lee, L.H., and Beemon, K.L. (1996). Avian retroviral RNA element promotes unspliced RNA accumulation in the cytoplasm. J. Virol. *70*, 3834–3843.

Oguro, H., Ding, L., and Morrison, S.J. (2013). SLAM Family Markers Resolve Functionally Distinct Subpopulations of Hematopoietic Stem Cells and Multipotent Progenitors. Cell Stem Cell *13*, 102–116.

Okada, S., Nakauchi, H., Nagayoshi, K., Nishikawa, S., Miura, Y., and Suda, T. (1992). In vivo and in vitro stem cell function of c-kit- and Sca-1-positive murine hematopoietic cells. Blood *80*, 3044–3050.

Ott, M.G., Schmidt, M., Schwarzwaelder, K., Stein, S., Siler, U., Koehl, U., Glimm, H., Kühlcke, K., Schilz, A., Kunkel, H., et al. (2006). 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.

Patwardhan, A., Samit, R., and Amit, R. (2014). Molecular Markers in Phylogenetic Studies-A Review. J. Phylogenetics Evol. Biol. *2*, 1–9.

Peikon, I.D., Gizatullina, D.I., and Zador, a. M. (2014). In vivo generation of DNA sequence diversity for cellular barcoding. Nucleic Acids Res. *42*, e127–e127.

Perié, L., Hodgkin, P.D., Naik, S.H., Schumacher, T.N., de Boer, R.J., and Duffy, K.R. (2014). Determining Lineage Pathways from Cellular Barcoding Experiments. Cell Rep. *6*, 617–624.

Perié, L., Duffy, K.R., Kok, L., de Boer, R.J., and Schumacher, T.N. (2015). The Branching Point in Erythro-Myeloid Differentiation. Cell *163*, 1655–1662.

Pfaff, N., Lachmann, N., Ackermann, M., Kohlscheen, S., Brendel, C., Maetzig, T., Niemann, H., Antoniou, M.N., Grez, M., Schambach, A., et al. (2013). A ubiquitous chromatin opening element prevents transgene silencing in pluripotent stem cells and their differentiated progeny. Stem Cells *31*, 488–499.

Piéroni, L., Bouillé, P., Auclair, C., Guillosson, J.J., and Nafziger, J. (1999). Early steps of replication of moloney murine leukemia virus in resting lymphocytes. Virology *262*, 408–415.

Poiesz, B.J., Ruscetti, F.W., Gazdar, A.F., Bunn, P.A., Minna, J.D., and Gallo, R.C. (1980). Detection and isolation of type C retrovirus particles from fresh and cultured lymphocytes of a patient with cutaneous T-cell lymphoma. Proc. Natl. Acad. Sci. U. S. A. *77*, 7415–7419.

Pollard, V.W., and Malim, M.H. (1998). The HIV-1 Rev protein. Annu. Rev. Microbiol. *52*, 491–532.

Popovic, M., Sarngadharan, M., Read, E., and Gallo, R. (1984). Detection, isolation, and continuous production of cytopathic retroviruses (HTLV-III) from patients with AIDS and pre-AIDS. Science (80-.). *224*, 497–500.

Porter, S.N., Baker, L.C., Mittelman, D., and Porteus, M.H. (2014). Lentiviral and targeted cellular barcoding reveals ongoing clonal dynamics of cell lines in vitro and in vivo. Genome Biol. *15*, R75.

Rama, A.R., Zafra, I., Burgos, M., and Prados, J. (2014). On Advances in Cancer Suicide-genes Therapy. SOJ Genet Sci *1*, 1–6.

Ramalho-Santos, M., and Willenbring, H. (2007). On the origin of the term “stem cell”. Cell Stem Cell *1*, 35–38.

Reed, I.S., and Solomon, G. (1960). Polynomial Codes Over Certain Finite Fields. J. Soc. Ind. Appl. Math. *8*, 300–304.

Reiser, J., Lai, Z., Zhang, X.Y., and Brady, R.O. (2000). Development of multigene and regulated lentivirus vectors. J. Virol. *74*, 10589–10599.

Rhoads, A., and Au, K.F. (2015). PacBio Sequencing and Its Applications. Genomics. Proteomics Bioinformatics *13*, 278–289.

Rio, P., Banos, R., Lombardo, A., Quintana-Bustamante, O., Alvarez, L., Garate, Z., Genovese, P., Almarza, E., Valeri, A., Diez, B., et al. (2014). Targeted gene therapy and cell reprogramming in Fanconi anemia. EMBO Mol. Med. *6*, 835–848.

Roe, T., Reynolds, T.C., Yu, G., and Brown, P.O. (1993). Integration of murine leukemia virus DNA depends on mitosis. EMBO J. *12*, 2099–2108.

Rohdewohld, H., Weiher, H., Reik, W., Jaenisch, R., and Breindl, M. (1987). Retrovirus integration and chromatin structure: Moloney murine leukemia proviral integration sites map near DNase I-hypersensitive sites. J. Virol. *61*, 336–343.

Rous, P. (1910). A TRANSMISSIBLE AVIAN NEOPLASM. (SARCOMA OF THE COMMON FOWL.). J. Exp. Med. *12*, 696–705.

Rous, P. (1911). A SARCOMA OF THE FOWL TRANSMISSIBLE BY AN AGENT SEPARABLE FROM THE TUMOR CELLS. J. Exp. Med. *13*, 397–411.

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

Sawai, C.M., Babovic, S., Upadhaya, S., Knapp, D.J.H.F., Lavin, Y., Lau, C.M., Goloborodko, A., Feng, J., Fujisaki, J., Ding, L., et al. (2016). Hematopoietic Stem Cells Are the Major Source of Multilineage Hematopoiesis in Adult Animals. Immunity *45*, 597–609.

Schambach, A., Bohne, J., Chandra, S., Will, E., Margison, G.P., Williams, D.A., and Baum, C. (2006). Equal potency of gammaretroviral and lentiviral SIN vectors for expression of O6-methylguanine-DNA methyltransferase in hematopoietic cells. Mol. Ther. *13*, 391–400.

Schepers, K., Swart, E., van Heijst, J.W.J., Gerlach, C., Castrucci, M., Sie, D., Heimerikx, M., Velds, A., Kerkhoven, R.M., Arens, R., et al. (2008). Dissecting T cell lineage relationships by cellular barcoding. J. Exp. Med. *205*, 2309–2318.

Schmidt, M., Zickler, P., Hoffmann, G., Haas, S., Wissler, M., Muessig, A., Tisdale, J.F., Kuramoto, K., Andrews, R.G., Wu, T., et al. (2002). Polyclonal long-term repopulating stem cell clones in a primate model. Blood *100*, 2737–2743.

Schmidt, M., Carbonaro, D. a, Speckmann, C., Wissler, M., Bohnsack, J., Elder, M., Aronow, B.J., Nolta, J. a, Kohn, D.B., and von Kalle, C. (2003). Clonality analysis after retroviral-mediated gene transfer to CD34+ cells from the cord blood of ADA-deficient SCID neonates. Nat. Med. *9*, 463–468.

Schmidt, M., Schwarzwaelder, K., Bartholomae, C., Zaoui, K., Ball, C., Pilz, I., Braun, S., Glimm, H., and von Kalle, C. (2007). High-resolution insertion-site analysis by linear amplification–mediated PCR (LAM-PCR). Nat. Methods *4*, 1051–1057.

Schoedel, K.B., Morcos, M.N.F., Zerjatke, T., Roeder, I., Grinenko, T., Voehringer, D., Gothert, J.R., Waskow, C., Roers, A., and Gerbaulet, A. (2016). The bulk of the hematopoietic stem cell population is dispensable for murine steady-state and stress hematopoiesis. Blood *128*, 2285–2296.

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

Seita, J., and Weissman, I.L. (2010). Hematopoietic stem cell: self-renewal versus differentiation. Wiley Interdiscip. Rev. Syst. Biol. Med. *2*, 640–653.

Selich, A., Daudert, J., Hass, R., Philipp, F., von Kaisenberg, C., Paul, G., Cornils, K., Fehse, B., Rittinghausen, S., Schambach, A., et al. (2016). Massive Clonal Selection and Transiently Contributing Clones During Expansion of Mesenchymal Stem Cell Cultures Revealed by Lentiviral RGB-Barcode Technology. Stem Cells Transl. Med. *5*, 591–601.

Shannon, C.E. (1948). A Mathematical Theory of Communication. Bell Syst. Tech. J. *27*, 379–423.

Sharma, A., Larue, R.C., Plumb, M.R., Malani, N., Male, F., Slaughter, A., Kessl, J.J., Shkriabai, N., Coward, E., Aiyer, S.S., et al. (2013). BET proteins promote efficient murine leukemia virus integration at transcription start sites. Proc. Natl. Acad. Sci. *110*, 12036–12041.

Shizuru, J.A., Negrin, R.S., and Weissman, I.L. (2005). Hematopoietic Stem and Progenitor Cells: Clinical and Preclinical Regeneration of the Hematolymphoid System. Annu. Rev. Med. *56*, 509–538.

Sorge, J., Ricci, W., and Hughes, S.H. (1983). cis-Acting RNA packaging locus in the 115-nucleotide direct repeat of Rous sarcoma virus. J. Virol. *48*, 667–675.

Sorge, J., Wright, D., Erdman, V.D., and Cutting, A.E. (1984). Amphotropic retrovirus vector system for human cell gene transfer. Mol. Cell. Biol. *4*, 1730–1737.

Spangrude, G.J., Heimfeld, S., and Weissman, I.L. (1988). Purification and characterization of mouse hematopoietic stem cells. Science (80-.). *241*, 58–62.

Spellerberg, I.F., and Fedor, P.J. (2003). A tribute to Claude Shannon (1916-2001) and a plea for more rigorous use of species richness, species diversity and the “Shannon-Wiener” Index. Glob. Ecol. Biogeogr. *12*, 177–179.

Stahl, T., Rothe, C., Böhme, M.U., Kohl, A., Kröger, N., and Fehse, B. (2016). Digital PCR Panel for Sensitive Hematopoietic Chimerism Quantification after Allogeneic Stem Cell Transplantation. Int. J. Mol. Sci. *17*, 1515.

Stehelin, D., Varmus, H.E., Bishop, J.M., and Vogt, P.K. (1976). DNA related to the transforming gene(s) of avian sarcoma viruses is present in normal avian DNA. Nature *260*, 170–173.

Stein, S., Ott, M.G., Schultze-Strasser, S., Jauch, A., Burwinkel, B., Kinner, A., Schmidt, M., Krämer, A., Schwäble, J., Glimm, H., et al. (2010). Genomic instability and myelodysplasia with monosomy 7 consequent to EVI1 activation after gene therapy for chronic granulomatous disease. Nat. Med. *16*, 198–204.

Stoye, J.P. (2012). Studies of endogenous retroviruses reveal a continuing evolutionary saga. Nat. Rev. Microbiol. *10*, 395–406.

Suerth, J., Labenski, V., and Schambach, A. (2014). Alpharetroviral Vectors: From a Cancer-Causing Agent to a Useful Tool for Human Gene Therapy. Viruses *6*, 4811–4838.

Suerth, J.D., Maetzig, T., Galla, M., Baum, C., and Schambach, A. (2010). Self-Inactivating Alpharetroviral Vectors with a Split-Packaging Design. J. Virol. *84*, 6626–6635.

Suerth, J.D., Maetzig, T., Brugman, M.H., Heinz, N., Appelt, J.-U., Kaufmann, K.B., Schmidt, M., Grez, M., Modlich, U., Baum, C., et al. (2012). Alpharetroviral Self-inactivating Vectors: Long-term Transgene Expression in Murine Hematopoietic Cells and Low Genotoxicity. Mol. Ther. *20*, 1022–1032.

Sun, J., Ramos, A., Chapman, B., Johnnidis, J.B., Le, L., Ho, Y.-J., Klein, A., Hofmann, O., and Camargo, F.D. (2014). Clonal dynamics of native haematopoiesis. Nature *514*, 322–327.

Svoboda, J., Chyle, P., Simkovic, D., and Hilgert, I. (1963). Demonstration of the absence of infectious Rous virus in rat tumour XC, whose structurally intact cells produce Rous sarcoma when transferred to chicks. Folia Biol. (Praha). *9*, 77–81.

Tao, W., Evans, B.-G., Yao, J., Cooper, S., Cornetta, K., Ballas, C.B., Hangoc, G., and Broxmeyer, H.E. (2007). Enhanced green fluorescent protein is a nearly ideal long-term expression tracer for hematopoietic stem cells, whereas DsRed-express fluorescent protein is not. Stem Cells *25*, 670–678.

Tatum, E.L. (2009). Molecular biology, nucleic acids, and the future of medicine. Repr. Cell Ther Transplant. Orig. Publ. Perspect. Biol. Med. *10*, 19–32.

Temin, H.M. (1964). The Participation of DNA in Rous Sarcoma Virus Production. Virology *23*, 486–494.

Thielecke, L., Aranyossy, T., Dahl, A., Tiwari, R., Roeder, I., Geiger, H., Fehse, B., Glauche, I., and Cornils, K. (2017). Limitations and challenges of genetic barcode quantification. Sci. Rep. *7*, 43249.

Till, J.E., and McCulloch, E.A. (1961). A direct measurement of the radiation sensitivity of normal mouse bone marrow cells. Radiat. Res. *14*, 213–222.

Till, J.E., McCulloch, E.A., and Siminovitch, L. (1964). A stochastic model of stem cell proliferation, based on the growth of spleen colony-forming cells. Proc. Natl. Acad. Sci. U. S. A. *51*, 29–36.

VandenDriessche, T., Thorrez, L., Naldini, L., Follenzi, A., Moons, L., Berneman, Z., Collen, D., and Chuah, M.K.L. (2002). Lentiviral vectors containing the human immunodeficiency virus type-1 central polypurine tract can efficiently transduce nondividing hepatocytes and antigen-presenting cells in vivo. Blood *100*, 813–822.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., et al. (2001). The sequence of the human genome. Science (80-.). *291*, 1304–1351.

Verovskaya, E., Broekhuis, M.J.C., Zwart, E., Ritsema, M., van Os, R., de Haan, G., and Bystrykh, L. V. (2013). Heterogeneity of young and aged murine hematopoietic stem cells revealed by quantitative clonal analysis using cellular barcoding. Blood *122*, 523–532.

Verovskaya, E., Broekhuis, M.J.C., Zwart, E., Weersing, E., Ritsema, M., Bosman, L.J., van Poele, T., de Haan, G., and

Bystrykh, L. V (2014). Asymmetry in skeletal distribution of mouse hematopoietic stem cell clones and their equilibration by mobilizing cytokines. J. Exp. Med. *211*, 487–497.

Wahlestedt, M., Erlandsson, E., Kristiansen, T., Lu, R., Brakebusch, C., Weissman, I.L., Yuan, J., Martin-Gonzalez, J., and Bryder, D. (2017). Clonal reversal of ageing-associated stem cell lineage bias via a pluripotent intermediate. Nat. Commun. *8*, 14533.

Wang, G.P., Levine, B.L., Binder, G.K., Berry, C.C., Malani, N., McGarrity, G., Tebas, P., June, C.H., and Bushman, F.D. (2009). Analysis of Lentiviral Vector Integration in HIV+ Study Subjects Receiving Autologous Infusions of Gene Modified CD4+ T Cells. Mol. Ther. *17*, 844–850.

Weber, K., Bartsch, U., Stocking, C., and Fehse, B. (2008). A Multicolor Panel of Novel Lentiviral "Gene Ontology" (LeGO) Vectors for Functional Gene Analysis. Mol. Ther. *16*, 698–706.

Weber, T.S., Dukes, M., Miles, D.C., Glaser, S.P., Naik, S.H., and Duffy, K.R. (2016). Site-specific recombinatorics: in situ cellular barcoding with the Cre Lox system. BMC Syst. Biol. *10*, 43.

Weiss, R.A. (1996). Retrovirus classification and cell interactions. J. Antimicrob. Chemother. *37 Suppl B*, 1–11.

Weiss, R.A. (2006). The discovery of endogenous retroviruses. Retrovirology *3*, 67.

Weissman, I.L. (2000). Stem cells: units of development, units of regeneration, and units in evolution. Cell *100*, 157–168.

Weissman, I.L., and Shizuru, J.A. (2008). The origins of the identification and isolation of hematopoietic stem cells, and their capability to induce donor-specific transplantation tolerance and treat autoimmune diseases. Blood *112*, 3543–3553.

Williams, K.M., and Gress, R.E. (2008). Immune reconstitution and implications for immunotherapy following haematopoietic stem cell transplantation. Best Pract. Res. Clin. Haematol. *21*, 579–596.

Wu, C., Li, B., Lu, R., Koelle, S.J., Yang, Y., Jares, A., Krouse, A.E., Metzger, M., Liang, F., Loré, K., et al. (2014). Clonal Tracking of Rhesus Macaque Hematopoiesis Highlights a Distinct Lineage Origin for Natural Killer Cells. Cell Stem Cell *14*, 486–499.

Wu, X., Li, Y., Crise, B., and Burgess, S.M. (2003). Transcription start regions in the human genome are favored targets for MLV integration. Science (80-.). *300*, 1749–1751.

Yamashita, M., and Emerman, M. (2006). Retroviral infection of non-dividing cells: old and new perspectives. Virology *344*, 88–93.

Zavidij, O., Ball, C.R., Herbst, F., Oppel, F., Fessler, S., Schmidt, M., von Kalle, C., and Glimm, H. (2012). Stable long-term blood formation by stem cells in murine steady-state hematopoiesis. Stem Cells *30*, 1961–1970.

Zhang, C.C., and Lodish, H.F. (2005). Murine hematopoietic stem cells change their surface phenotype during ex vivo expansion. Blood *105*, 4314–4320.

Zhang, F., Thornhill, S.I., Howe, S.J., Ulaganathan, M., Schambach, A., Sinclair, J., Kinnon, C., Gaspar, H.B., Antoniou, M., and Thrasher, A.J. (2007). Lentiviral vectors containing an enhancer-less ubiquitously acting chromatin opening element (UCOE) provide highly reproducible and stable transgene expression in hematopoietic cells. Blood *110*, 1448–1457.

Ziętara, N., Łyszkiewicz, M., Puchałka, J., Witzlau, K., Reinhardt, A., Förster, R., Pabst, O., Prinz, I., and Krueger, A. (2015). Multicongenic fate mapping quantification of dynamics of thymus colonization. J. Exp. Med. *212*, 1589–1601.

Zufferey, R., Nagy, D., Mandel, R.J., Naldini, L., and Trono, D. (1997). Multiply attenuated lentiviral vector achieves efficient gene delivery in vivo. Nat. Biotechnol. *15*, 871–875.

Zufferey, R., Donello, J.E., Trono, D., and Hope, T.J. (1999). Woodchuck hepatitis virus posttranscriptional regulatory element enhances expression of transgenes delivered by retroviral vectors. J. Virol. *73*, 2886–2892.

Zychlinski, D., Schambach, A., Modlich, U., Maetzig, T., Meyer, J., Grassman, E., Mishra, A., and Baum, C. (2008). Physiological Promoters Reduce the Genotoxic Risk of Integrating Gene Vectors. Mol. Ther. *16*, 718–725.

13.1. Web sources bibliography

https://archive.org/details/natrlichesch1868haec, accessed 20170608 (german)

http://blogs.nature.com/news/2012/11/gene-therapy-hits-european-market.html, accessed 20170608

http://dictionary.cambridge.org/dictionary/english/bar-code, accessed 20170608

http://flexikon.doccheck.com/de/Knochenmark, accessed 20170608 (german)

http://flexikon.doccheck.com/de/Erythrozyt, accessed 20170608 (german)

http://medical-dictionary.thefreedictionary.com/viral+tropism, accessed 20170608

http://www.biology-online.org/dictionary/Hematopoiesis, accessed 20170608

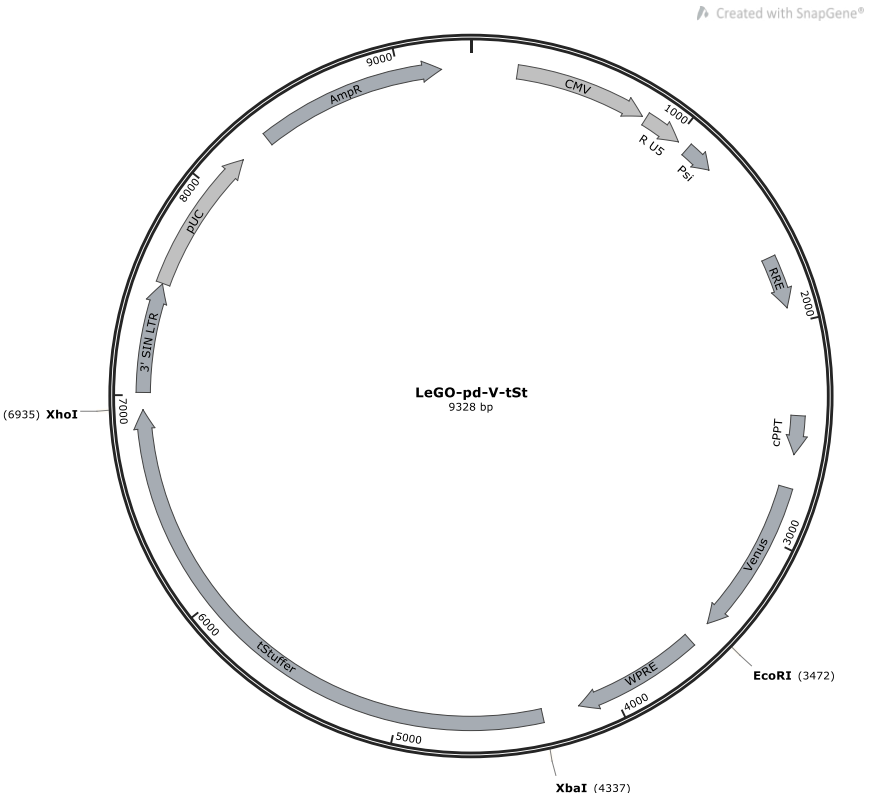
http://www.ema.europa.eu/ema/index.jsp?curl=pages/medicines/human/medicines/003854/human_med_001985.jsp&mid=WC0b01ac058001d124, accessed 20170608

https://www.firstwordpharma.com/node/1386939?tsid=28&region_id=3, accessed 20170608

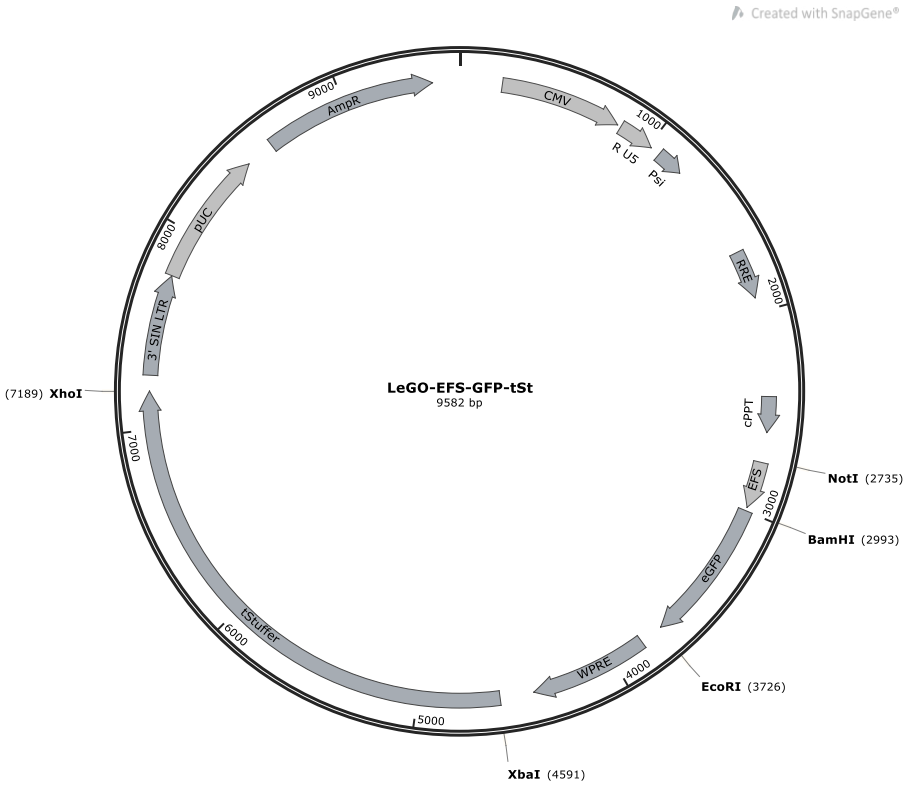
http://www.illumina.com/systems/sequencing.html, accessed 20170608

http://www.lentigo-vectors.de/, accessed 20170608

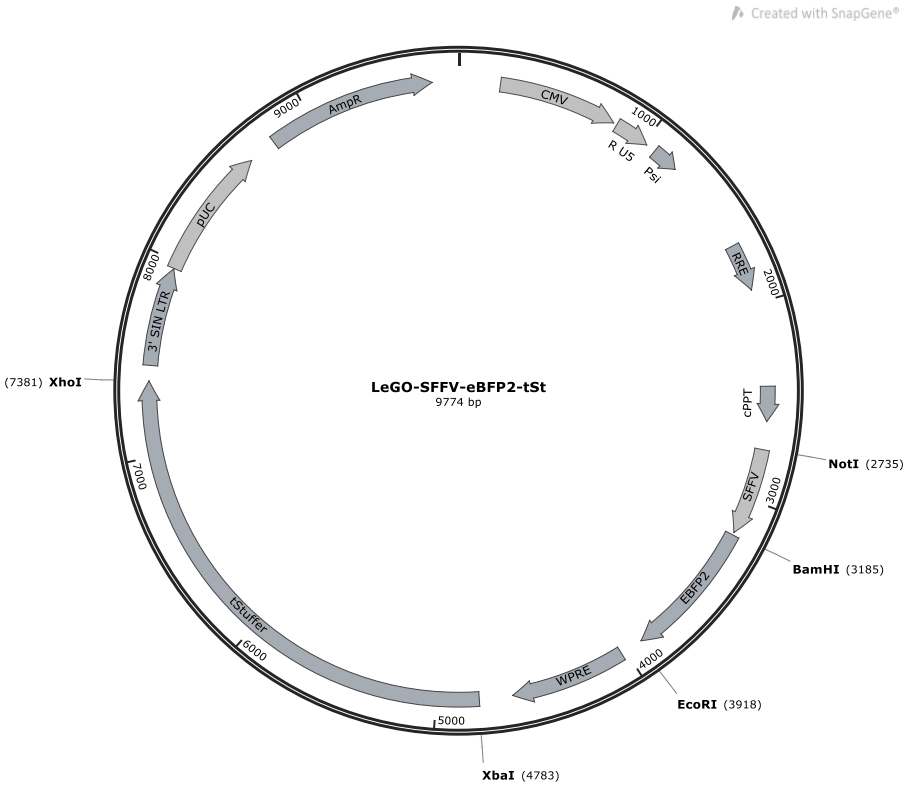
14. Appendix



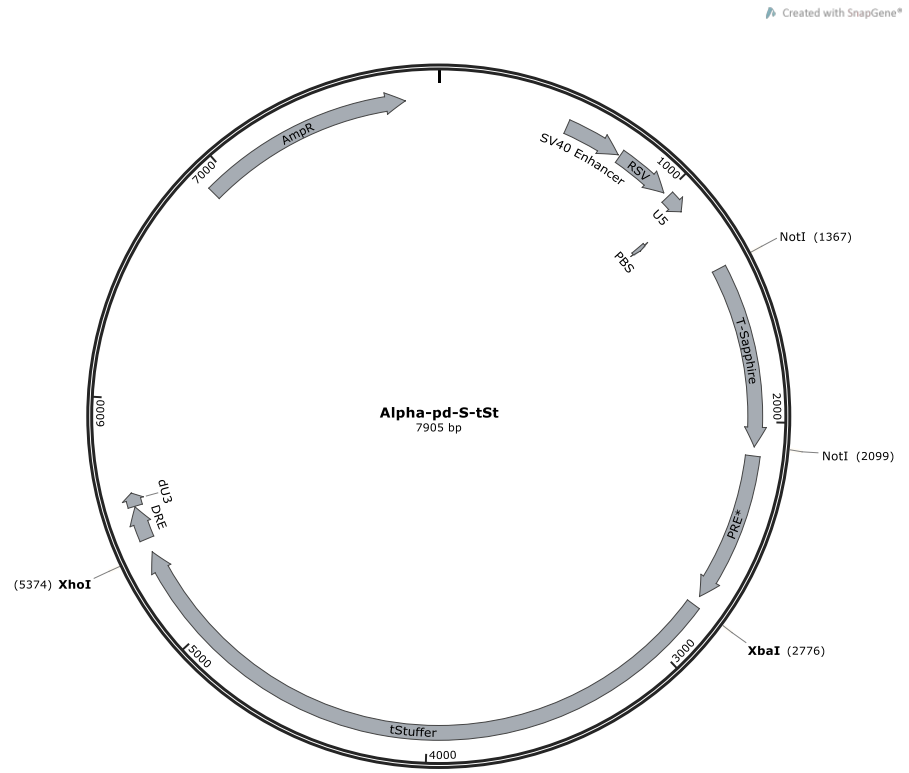
Appendix figure 1 – Promoter-deprived lentiviral construct with Venus FP and toxic stuffer



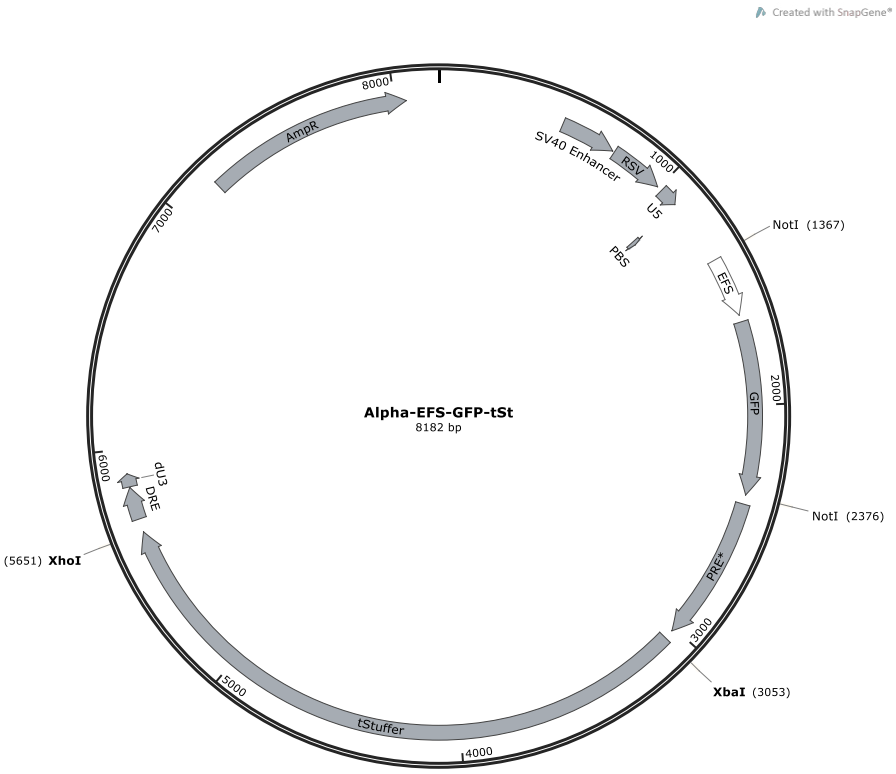
Appendix figure 2 – Lentiviral construct with GFP FP, an intermediate EFS promoter and toxic stuffer



Appendix figure 3 – Lentiviral construct with BFP FP, a strong SFFV promoter and toxic stuffer



Appendix figure 4 – Promoter-deprived alpharetroviral construct with TSapphire FP and toxic stuffer



Appendix figure 5 – Alpharetroviral construct with GFP FP, EFS promoter and toxic stuffer

A

Tranduction rates [% FP expression]

	GFP	eBFP
MG1	13.20	
MG2	17.20	59.00
MG3	24.50	
MG4.2	19.20	42.60
mean	**18.53**	**50.80**

B

ECCs in BM (data from independen experiment)

Sample	total BM x10^8	lin- x10^7	%lin-	% LSK in lin-	LSK CD150+ = pECC **% pECC in LSK**
1	4.00	0.86	2.15	18.80	25.90
2	3.00	1.03	3.43	24.20	26.20
3	5.00	0.82	1.64	19.60	23.70
4	2.00	0.87	4.35	20.50	21.20
5	5.00	0.99	1.98	15.00	22.00
6	4.00	1.20	3.00	20.50	21.20
7	5.00	0.38	0.76	18.70	26.10
8	5.00	0.32	0.64	13.40	21.70
9				18.60	28.10
10				32.10	29.70
		mean	2.24	20.14	**24.58**
		SD	1.21	4.88	2.90

C

Expected pECC in graft:

per barcode backbone **2 backbone graft**		per barcode backbone **3 backbone graft**		within total graft
200000	lin-	133000	lin-	400000 lin-
40280	cKit/Sca1	26786	cKit/Sca1	80560 LSK
9901	**ECC**	**6584**	**ECC**	19802 ECC

D

Theoretically transduced pECCs (expected pECC*transduction rates) per graft

Transduciton rates pd constructs = GFP

2 backbone graft		**3 backbone graft**			
GFP	Venus/TSapp	Venus/TSapp	GFP	BFP	
1834	**1834**	**1220**	**1220**	**3345**	**transduced pECCs per animal**

E

Total number expected of marked pECCs in one graft (sum of backbones)

2 backbone graft	3 backbone graft	
3668	**5784**	**transduced pECCs per graft**

Expected number of unmarked pECCs in one graft

(exp. pECC in graft - number of marked pECCs)

2 backbone graft	3 backbone graft	
16133	14018	unmarked pECCs per graft

F

Overall number of recovered barcodes:

4 animals with 2 backbone, 4 animals with 3 backbones

Multiply animal number * theoretically transd. ECCs per graft

Example: Lenti-GFP: 4*1834 + 4*1220 = 12215

12215 total Lenti-GFP transduced pECCs in MG1+2
1768 barcodes recovered overall for Lenti-GFP
14 % Recovery

12215 total Lenti-Venus transduced pECCs in MG1+2
1042 barcodes recovered overall for Lenti-Venus
9 % Recovery

12215 total Alpha-GFP transduced pECCs in MG4+4.2
710 barcodes recovered overall for Alpha-GFP
6 % Recovery

12215 total Alpha-TSapp transduced pECCs in MG4+4.2
930 barcodes recovered overall for Alpha-TSapp
8 % Recovery

26758 total Lenti-BFP transduced pECCs in MG2+4.2
967 barcodes recovered overall for Lenti-BFP
4 % Recovery

8 % mean recovery

G

Correction of expected pECCs by 8% recovery factor

(Expected number of cells * 0.08)

2 backbone graft	3 backbone graft	
294	**463**	**marked ECCs per graft**
1292	**1123**	**non-marked ECCs per graft**

H

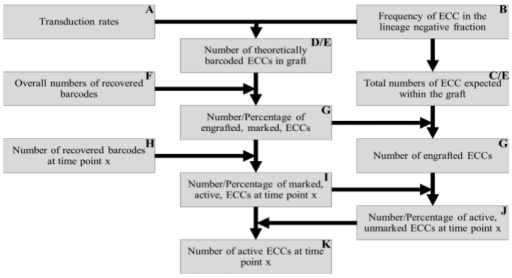
Observed clone numbers (mean over all backbones, 0.5% threshold)

2 backbone graft		3 backbone graft	
59.0	**PB6w**	**108.5**	**PB6w**
43.9	**Pbfinal**	**81.8**	**Pbfinal**
41.8	Spleen	93.3	Spleen
45.8	KM	76.8	KM
51.4	lin-	68.3	lin-

pECCs: potentially engraftment-capable cells

ECCS: engraftment-capable cells

difference: only 8% of the pECCs are found to engraft during the experiment althougth all of them should be able to. ECCs are the cells that did engraft



I

% of marked ECCs active at time point x (Observed/(Expected/100))

2 backbone graft		3 backbone graft		mean
20.08	**PB6w**	**23.42**	**PB6w**	21.75
14.93	**Pbfinal**	**17.65**	**Pbfinal**	16.29
14.21	Spleen	20.13	Spleen	17.17
15.57	KM	16.57	KM	16.07
17.49	lin-	14.73	lin-	16.11

J

Number of active non-marked ECCs at time point x

(% observed of marked ECCs*Expected unmarked ECCs/100)

2 backbone graft		3 backbone graft	
259	PB6w	263	PB6w
193	Pbfinal	198	Pbfinal

K

ECCs active at time point x (Sum of observed clone numbers + active non-marked ECCs)

2 backbone graft		3 backbone graft		**mean**	
318	PB6w	371	PB6w	**345**	**PB6w**
237	Pbfinal	280	Pbfinal	**258**	**Pbfinal**

Appendix figure 6 - Calculation of active ECCs