Dissertation

submitted to the
Combined Faculties for the Natural Sciences and for Mathematics
of the Ruperto-Carola University of Heidelberg, Germany

for the degree of
Doctor of Natural Science

presented by
Peer Wünsche, M.Sc.
born in Lübeck, Germany

Oral examination: September 28th, 2018

Patient derived γ-retroviral integration sites reveal active gene-regulatory regions in human repopulating long-term hematopoietic stem cells

Referees:

Prof. Dr. Michael Boutros

Prof. Dr. Hanno Glimm

“Nein”, erwiderte der Alchemist. „Was du noch wissen musst, ist folgendes: Immer, bevor ein Traum in Erfüllung geht, prüft die Weltenseele all das, was auf dem Weg gelernt wurde. Sie macht das nicht etwa aus Bosheit, sondern damit wir gemeinsam mit unserem Traum auch die Lektionen in Besitz nehmen, die wir auf dem Pfad dorthin gelernt haben. […] Eine Suche beginnt immer mit dem Anfänger Glück. Und sie endet immer mit der Prüfung des Eroberers.“

- Paulo Coelho, Der Alchemist –

Für meinen Vater

SUMMARY

Hematopoietic stem cell (HSC) research largely relies on cell culture models or mouse transplantation studies. Moreover, HSCs are rare and immunophenotypic definitions are incomplete, rendering the characterization of HSCs difficult. In this study, we circumvented these restrictions using >180,000 γ-retroviral (γRV) integration sites (ISs) from a gene therapy trial on 10 Wiskott-Aldrich-Syndrome patients. γRV ISs leave a unique tag to hematopoietic stem and progenitor cells (HSPCs) that engraft in patients, which are passed on to all progeny, making them suitable to track clonal reconstitution dynamics. Moreover, γRV ISs can be used to map active promoters and enhancers, due to their predilection to integrate at such sites. ISs recovered during stable long-term hematopoiesis would therefore point towards active promoters and genes that originate from true repopulating long-term HSCs. However, due to the genotoxic potential of γRVs, ISs are often regarded as molecular tags that point towards proto-oncogenes.

To examine this in more detail, we first cloned 20 protein-coding genes that showed a large number of ISs in their vicinity and established a pooled lentiviral overexpression library to study their influence on proliferation, self-renewal and differentiation of HSPCs. Although the characterization of individual candidate genes was limited by transduction efficiencies and library representation, we observed that not a single candidate gene led to clonal expansion or measurable increase in self-renewal during both *in vitro* and *in vivo* experiments, suggesting that γRV genotoxicity is less universal than expected.

Based on this, we assessed the cumulative number of ISs per gene over time and statistically compared γRV IS pattern before and after transplantation, demonstrating that the clonal skewing of IS pattern is indeed restricted to only few known leukemogenic loci. We next modeled the hematopoietic reconstitution after transplantation in humans and used these insights to define long-term HSC specific ISs, which confirmatively showed the highest ATAC-seq signal intensity at HSC specific peaks, efficiently enriched for HSC specific gene sets and strongly correlated with hematopoietic risk variants. Finally, through integration of publicly available ATAC-seq,

ChIP-seq, capture Hi-C as well as GWAS SNP data, we were able to create the first genome wide map for active gene-regulatory regions in functionally defined human repopulating long-term HSCs.

ZUSAMMENFASSUNG

Forschung an hämatopoetischen Stammzellen (HSZ) basiert weitgehend auf Zellkulturmodellen oder Maus Transplantationsstudien. Darüber hinaus sind HSZ sehr selten und deren immunphänotypische Aufreinigung oft unzureichend, was die Charakterisierung von HSZ zusätzlich erschwert. Durch die Verwendung von mehr als 180,000 γ-retrovirale (γRV) Integrationsstellen (IS) aus einer Gentherapie-Studie an 10 Wiskott-Aldrich-Syndrom Patienten konnten wir jedoch die vorher genannten Einschränkungen umgehen. Dies ist möglich, da die γRV IS eine unverwechselbare Markierung in hämatopoetische Stamm- und Vorläuferzellen (HSVZ) hinterlassen, die an alle Nachkommen weitergegeben wird. Da γRV präferentiell in aktiven regulatorischen Einheiten integrieren, können mithilfe der genomischen Positionen von IS nicht nur klonale Rekonstitutionsdynamiken analysiert werden, sondern auch aktive Promotoren und Enhancer kartografiert werden. IS die während der stabilen Langzeit-Hämatopoese detektiert wurden, weisen daher auch auf aktive Promotoren und Gene hin, die von repopulierenden Langzeit-HSZ abstammen. Allerdings werden γRV IS aufgrund ihres genotoxischen Potentials auch oft als molekulare Markierungen für Proto-Onkogene angesehen.

Um dies weiter zu untersuchen, klonierten wir zunächst 20 Protein-kodierende Gene, die eine große Anzahl von IS in ihrer Nähe aufwiesen, in eine lentivirale Überexpression Bibliothek um deren Einfluss auf Proliferation, Selbsterneuerung und Differenzierung in HSVZ zu untersuchen. Obwohl die Charakterisierung einzelner Kandidatengene durch Transduktionseffizienzen limitiert war, konnten wir weder klonale Expansion oder messbare Zunahme der Selbsterneuerung während der *in vitro* noch der *in vivo* Experimenten feststellen. Dies lies vermuten, dass γRV Genotoxizität weniger universal ist als bisher angnommen.

Basierend darauf haben wir sowohl die zeitliche Zunahme an IS pro Gen gemessen, als auch das IS Muster vor und nach der Transplantation statistisch verglichen. Diese Analysen zeigten ebenfalls, dass klonale Verzerrungen des IS Musters tatsächlich auf einige bekannte leukämogene Loci beschränkt sind. Als nächstes modellierten wir die hämatopoetische Rekonstitution nach Transplantation bei Menschen und nutzten diese Erkenntnisse, um HSZ-spezifisches IS zu definieren. Die HZS-Spezifität konnte weiterhin sowohl durch eine hohe Korrelation mit HSZ-spezifischen „ATAC-Seq“ Signalen gezeigt werden, als auch durch signifikante Anreicherung von IS an HSZ-spezifischen Genen und hämatopoetischen Risiko-Genomvarianten. Zusammenfassend konnten wir durch die Integration von öffentlich verfügbaren Hochdurchsatz-Sequenzdaten (ATAC-seq, ChIP seq, HiC, GWAS-SNP) die erste genomweite Karte für aktive regulatorische Regionen in funktionell definierten humanen repopulierenden HSZ erstellen.

TABLE OF CONTENTS

1 Introduction 1

1.1 Hematopoiesis – the production of blood 1

1.1.1 From hematopoiesis to hematopoietic stem cells 1
1.1.2 Immunophenotypic isolation of murine and human hematopoietic stem cells 2
1.1.3 The hematopoietic system – discrete vs. continuum-based models 6
1.1.4 Post-transplant versus unperturbed steady-state hematopoiesis 7

1.2 Gene regulation in eukaryotes 8

1.2.1 Promoters 9
1.2.2 Enhancers 11
1.2.3 Insulators 12
1.2.4 Higher order genome structures through topologically associated domains 13

1.3 Identification of regulatory regions 16

1.3.1 Accessibility of chromatin 16
1.3.2 Identification of enhancers 17

1.3.2.1 Chromatin states and histone modifications 18

1.3.3 Medical impact of sequence alterations in regulatory regions 19

1.4 Hematopoiesis and its correction in the context of disease 20

1.4.1 Gene therapy for the treatment of Wiskott-Aldrich-Syndrome 21

1.5 Biology of retroviruses 22

1.5.1 Integration biology of γ-retroviruses 24
1.5.2 Retroviral insertional mutagenesis 26
1.5.3 Using γRV ISs as molecular tags for active regulatory regions. 27

2 Aims of the thesis 29

3 Results 31

3.1 Identification of novel key hematopoietic regulators through γ-retroviral insertion sites 31

3.1.1 Analysis of common integration sites and genes in their vicinity 32
3.1.2 Top 100 CIS are highly enriched for hematopoietic regulators 33
3.1.3 Selection of protein-coding genes as potential novel hematopoietic regulators 34
3.1.4 Most candidate genes are expressed in the hematopoietic system 36

3.2 Establishing a pooled lentiviral based screening platform 36

3.2.1 Stable overexpression of candidate genes with lentiviral vectors 37
3.2.2 Custom molecular barcoding and high-throughput sequencing cassette 38

3.2.3 Systematic one-by-one amplification of indices reveals very low index-bleeding 40
3.2.4 Titers can vary between constructs and productions and do not correlate with cDNA length 42
3.2.5 Stable GFP expression and barcode representation over time *in vitro* 43

3.3 Characterization of candidate hematopoietic regulators 44

3.3.1 Results of serial CFU re-plating assays revealed changes in BC representation over time but were limited by poor library presentation 44
3.3.2 Cell trace experiment indicated varying proliferative potential between cells transduced with different lentiviral OE constructs 47
3.3.3 *In vivo* lentiviral overexpression screen reveals potential hematopoietic regulators but is again limited by engraftment and library representation 49
3.3.4 Statistical evaluation of BC combinations indicates that changes in BC proportions are partially driven by a biological effect 52

3.4 Mapping active gene-regulatory regions using γ–retroviral integration sites 54

3.4.1 The majority of γRV ISs are located in enhancer regions 55
3.4.2 Switch from short-term to long-term hematopoiesis after transplantation occurs after 6-12 months 56
3.4.3 Integration site pattern of γRVs mark cell-type specific regulatory elements 59
3.4.4 Rainfall plots visualize commonalities and differences between patient ISs and ISs from CD34$^+$ cells 60
3.4.5 Clonal skewing of ISs pattern is restricted to few known leukemogenic loci 63
3.4.6 HSC specific ATAC-seq signal intensity but not gene expression correlates with IS pattern 65
3.4.7 Reported key hematopoietic transcription factors are efficiently marked by WAS IS 71
3.4.8 Differences between patient and CD34$^+$ ISs highlight long-term HSC specific genes 73
3.4.9 The time point of ISs detection indicates activity of enhancer modules in HSCs and progenitor populations 75
3.4.10 Tagged enhancers and disease variants can be linked to their putative target gene by integration with long-range interaction data 77
3.4.11 Clusters of ISs co-localize with hematological GWAS SNPs 78
3.4.12 Integration sites show elevated sequence conservation 81
3.4.13 Boundaries of topological associated domains but not CTCF sites show significantly more ISs than expected by chance 82
3.4.14 γRV provide a catalogue of >3,000 regulatory regions in functionally define human long-term HSCs 84

4 Discussion..85

4.1 The lentiviral overexpression-library approach is largely limited by transduction efficiencies and cell numbers 85

4.1.1 Tracking clonal dynamics *in vitro* and *in vivo* through genetic barcodes and multiplexed HT-seq is technically feasible 85
4.1.2 Titer variations in cDNA overexpression library screens 86
4.1.3 Transduction with pooled overexpression library in CFU assays is limited by colony number and transduction efficiency 87
4.1.4 Cell trace experiments showed improved library representation but insufficient timeframe 89
4.1.5 Pooled overexpression screen *in vivo* mostly suffers from low transduction efficiencies. 90
4.1.6 γRV genotoxicity is probably less universal than expected. 91

4.2 Oncogenic γ-retroviral integration events are less common than expected 93
4.3 Using γ-retroviral integration sites from long-term repopulating clones to map active gene-regulatory regions in HSCs 95

4.3.1 HSCs differ in their capacity to differentiate and self-renew 95
4.3.2 The post-transplant enrichment of HSCs can be also be visualized using ATAC-seq signal intensity and gene set enrichment analysis 98
4.3.3 γ-retroviral integration might also point towards super-enhancers 99
4.3.4 Prediction of enhancer activity during early hematopoiesis using the time point of IS detection 100

4.4 Conclusion and perspectives 101

5 Material and Methods 103

5.1 Material 103

5.1.1 Technical equipment 103
5.1.2 Commercial kits 104
5.1.3 Reagents 104
5.1.4 Consumables 106
5.1.5 Plasmids 107
5.1.6 Western blot Antibodies 107
5.1.7 Flow cytometry antibodies and staining panels 108
5.1.8 Oligonucleotides 110
5.1.9 Mouse Strains 116
5.1.10 Software 116

5.2 Methods 118

5.2.1 Molecular cell biology 118

5.2.1.1 RNA isolation 118
5.2.1.2 DNA isolation for mouse genotyping 118
5.2.1.3 Genotyping PCR 118
5.2.1.4 Reverse transcription and Quantitative real-time PCR (qRT-PCR) 119
5.2.1.5 Protein isolation and semi dry Western blot for the detection of Igf2bp2 119
5.2.1.6 Cloning of the barcode cassette containing p902 target vector 122
5.2.1.7 Synthesis of candidate gene cDNAs and cloning into p902 target vector 122

5.2.1.8 Barcode amplification followed by high-throughput sequencing 123
5.2.1.9 Cell culture 124
5.2.1.10 Virus production 125
5.2.1.11 Lentiviral titer calculation 125
5.2.1.12 Lentiviral transduction of LSK$^{Rosa26\ rtTA}$ or LSK-SLAM$^{Rosa26\ rtTA}$ cells 126
5.2.1.13 Colony forming unit assay 126
5.2.1.14 Cell trace assay 127
5.2.1.15 Fluorescence activated cell analyzing and sorting 127

5.2.2 Mouse experiments 128

5.2.2.1 Harvesting of bone marrow samples for the isolation of hematopoietic stem and progenitor cells 128
5.2.2.2 Acquiring of blood samples from transduced mice 128
5.2.2.3 Bone marrow transplantations 129

5.2.3 Patient integration site data sample collection from Wiskott-Aldrich syndrome gene therapy study 129
5.2.4 Computational analysis 130

5.2.4.1 Acquisition of datasets used in this study 130
5.2.4.2 Cluster analysis for the selection of candidate genes 131
5.2.4.3 Centered distance of ISs to TSSs and percentage of genes tagged by IS 131
5.2.4.4 Rainfall plots 131
5.2.4.5 Cluster prediction 131
5.2.4.6 Overlap of ChIP-seq signal and IS positions 131
5.2.4.7 Pearson correlation and PCA 132
5.2.4.8 Estimation of hematopoietic switch after transplantation using pairwise positive association matrix 132
5.2.4.9 Comparison of cluster dimension and sizes between IS datasets 133
5.2.4.10 Assessment of growth kinetics to address the frequency of overproportional clonal expansion 133
5.2.4.11 Fitting the mean of the $\log_2$ odds ratio using a 4-parametric log-logistic model 134
5.2.4.12 Assessment of ATAC-seq signal and gene expression at sites of integration 134
5.2.4.13 Differentially tagged Ggnes between CD34$^+$ cells and patients and gene set enrichment analysis 135
5.2.4.14 Assessment of activity of MYC enhancer modules using WAS ISs 135
5.2.4.15 Enrichment of ISs in CHi-C interaction fragments 136
5.2.4.16 GWAS and common SNP enrichment analysis 136
5.2.4.17 Circularized view of CD34$^+$ capture Hi-C long-range interactions and WAS ISs and WashU epigenome browser custom track 137
5.2.4.18 Sequence conservation at sites of integration 137
5.2.4.19 Permutation test to assess the enrichment of ISs at TAD boundaries 137
5.2.4.20 Permutation test to assess the enrichment of ISs at CTCF sites from CD34$^+$ cells 138

6 References..139

7 Contributions..155

8 Acknowledgements 157

9 Appendix.. 159

9.1 List of figures 159
9.2 List of tables 161
9.3 List of equations 162
9.4 Abbreviations 163
9.5 ClusterCount function for R 167
9.6 Condon optimized cDNA sequences of candidate genes 168

9.6.1 Amica1 168
9.6.2 Ccnd3 168
9.6.3 Evl 169
9.6.4 Fbxl18 169
9.6.5 Igf2bp2 170
9.6.6 Irf2bpl 171
9.6.7 Lair 1 long isoform 171
9.6.8 Lair1 short isoform 172
9.6.9 Lrrc33 172
9.6.10 Mbnl1 173
9.6.11 Ninj2 173
9.6.12 Nrip1 174
9.6.13 Plcb4 175
9.6.14 Prkcb 176
9.6.15 Slx4ip long isoform 176
9.6.16 Slx4ip short isoform 177
9.6.17 Swap70 177
9.6.18 Xbp1 178
9.6.19 Xbp1S 178
9.6.20 Znf217 179

1 Introduction

1.1 Hematopoiesis – the production of blood

The term hematopoiesis describes the production of the cellular components of the blood system, one of the most regenerative tissues in the body. Every day, 10^{12} - 10^{13} blood cells are produced in the human body, which resemble about 45% of the five to six liters of blood in a human adult. This so-called hematocrit can be coarsely segregated into red and white blood cells and platelets. While red blood cells were already described in the 17th century, it took until the 1840s until platelets and white blood cells (leukocytes) were discovered (Hajdu, 2003). Up until today, well over 20 differentiated blood cell types and their precursors have been described. The underlying principles of their formation however are still under heavy investigations with implications ranging from basic science to translational medicine.

1.1.1 From hematopoiesis to hematopoietic stem cells

While the existence of different blood cell types was already postulated in the 17th century, their birthplace was only discovered in 1868 by a pathologist, who reported for the first time that in mammals, blood cells are produced in the bone marrow (BM) and that mature cells exit the marrow via small blood vessels. Up until then, leukemia was considered a deficiency disease, so physicians eagerly began to test rather bizarre treatments based on these new findings. For example, patients were compelled to swallow fresh BM from juvenile cattle at equal parts with fresh orange juice (Forman et al., 2015). Obviously, such therapeutic approaches remained without success and it took until the 1950s before first allogeneic hematopoietic cell transplantations were reported. In 1954, it was noted that irradiated mice recover after infusions from spleen and BM cells, followed by experiments in 1956 that showed that transplanted recipient mice exhibit the same cytogenetic characteristics of the donor (Ford et al., 1956). In the same year, the first successful treatment for leukemia in mice was reported: high-dose full body irradiation followed by BM transplantation (Barnes et al., 1956). Although hematopoietic precursor cells were already postulated in the 19th century, the first experimental evidence for the existence of hematopoietic stem cells (HSCs) –although not defined yet – was only given with these transplantation experiments. Directed research on HSCs first began with observations made by Till and McCulloch in the 1960s, who reported that the number of colonies in the spleen are related to the number of marrow cells transplanted and that colonies contained myeloid and erythroid cells (Till and McCulloch, 1961). Later it was reported that cells within individual spleen colonies are clonal thus originate from a common precursor (Bortin, 1970; Thomas et al., 1959; Thomas et al., 1957; Wu et al., 1967) and that colonies were formed even after secondary transplantations (Siminovitch et al., 1963). Right about then, these features – multipotency and the capacity to self-renew – were the defining properties of hematopoietic stem cells (HSCs), which still hold true today. The isolation or enrichment of HSCs, however, was not possible before the development of monoclonal antibodies (ABs) against blood cell surface markers and fluorescence-activated cell sorters (FACS) in combination with new clonogenic *in vivo* and *in vitro* assays for differentiated cells (Forman et al., 2015). In the following paragraphs, the prospective isolation of HSCs as well as the hematopoietic system with its numerous cell types and hierarchical organization is discussed in more detail.

1.1.2 Immunophenotypic isolation of murine and human hematopoietic stem cells

With the knowledge of the existence of hematopoietic stem cells and the arising technological advances such as the development of monoclonal ABs and FACS, researchers eagerly sought for surface markers that would help to distinguish HSC from other populations. The first striking enrichment for murine HSCs was achieved by selecting for cells that would not express any markers that are characteristic for differentiated cells (Lineage negative; Lin$^-$), such as B220 (B cells), CD11b and Gr-1 (granulocytes) or CD4 and CD8 (T cells) (Müller-Sieburg et al., 1988; Müller-Sieburg et al., 1986). Furthermore, it was noted that mouse Lin$^-$ cells were also low for Thy-1 (CD90), a marker that in combination with stem cell antigen-1 (Sca-1) positivity enriched HSCs even further (Spangrude et al., 1988). Soon after, the stem cell factor receptor c-Kit was discovered, which coined the still widely used “LSK” marker combination: Lin$^-$ Sca-1$^+$ c-Kit$^+$ (Ikuta and Weissman, 1992; Ogawa et al., 1991).In the following years, scientist developed growing panels of surface marker combinations to increase the purity of the HSC population. For example, LSK cells being additionally negative for CD34 and CD135 (Flk2/Flt3) show even higher HSC activity, with CD34 further segregating HSCs into cells with short-term (CD34$^+$) and long-term (CD34$^-$) repopulating properties (Adolfsson et al., 2001; Christensen and Weissman, 2001; Osawa et al., 1996). LSK markers can also be combined with the signaling lymphocytic activation molecule (SLAM) markers, CD48 and CD150 with HSCs being positive for CD150 and negative for CD48 (Kiel et al., 2005). Also, the combination of LSK, CD34, CD135 and SLAM markers yields highly purified HSCs (Wilson et al., 2008). Apart from using these markers for the prospective enrichment of HSCs, other combinations of the same surface molecules can also be used for e.g. more committed progenitors (Figure 1A).

Compared to murine HSCs, identifying surface markers for human HSCs poses much more challenges. All of the above described surface marker combinations were discovered and validated through mouse transplantation experiments. Such functional experiments measure the engraftment and output of the transplanted cells, with true HSCs being capable of replenishing the entire blood system with all lineages over a long period of time. Human cells however are rejected by the murine immune system upon transplantation, raising the need for immunodeficient mice. To date, many different immunodeficient mouse models exist, such as the widely used NOD-Scid Il2γc$^{-/-}$ (NSG) mice, which lack B and T and natural killer (NK) cells. Interestingly, the very first marker for the enrichment of human HSCs was already discovered before the presence of immunodeficient mice. In 1984, Civin and colleagues discovered CD34 as a surface marker only present on histologically immature normal and leukemic BM cells (Civin et al., 1984). Later, numerous transplantation experiments in mice and patients validated CD34 as an HSC-enriching marker, which is still widely used today. Other important human HSC markers were discovered only later with the help of transplantation experiments, for example the absence of CD45R (Lansdorp et al., 1990), CD38 (Bhatia et al., 1997) and low rhodamine 123 retention (Rho^{low}) (McKenzie et al., 2007) or on the contrary expression of CD90 (Baum et al., 1992) and CD49f (Notta et al., 2011). Taking all of these markers together, CD34$^+$ CD38$^-$ CD45RA$^-$ CD90$^+$ CD49f$^+$ and Rho^{low} cells would resemble the highest possible enrichment of human HSCs to date (Figure 1B). Although about 1 in 15 cells within this pool possesses the ability of long-term blood reconstitution (Huntsman et al., 2015), one has to consider that some cells outside this definition also have the capacity for long-term engraftment, which also holds true for immunophenotypically enriched murine HSCs. For example Weksberg et al. (2008) showed, that a CD150^{-} side population also contained long-term HSCs. Moreover, recent technological advances like single cell sequencing as well as single cell transplantation experiments indicate significant heterogeneity even within highly purified populations, raising the question how well phenotype and function are really linked (Lu et al., 2011; McKenzie et al., 2006; Velten et al., 2017).

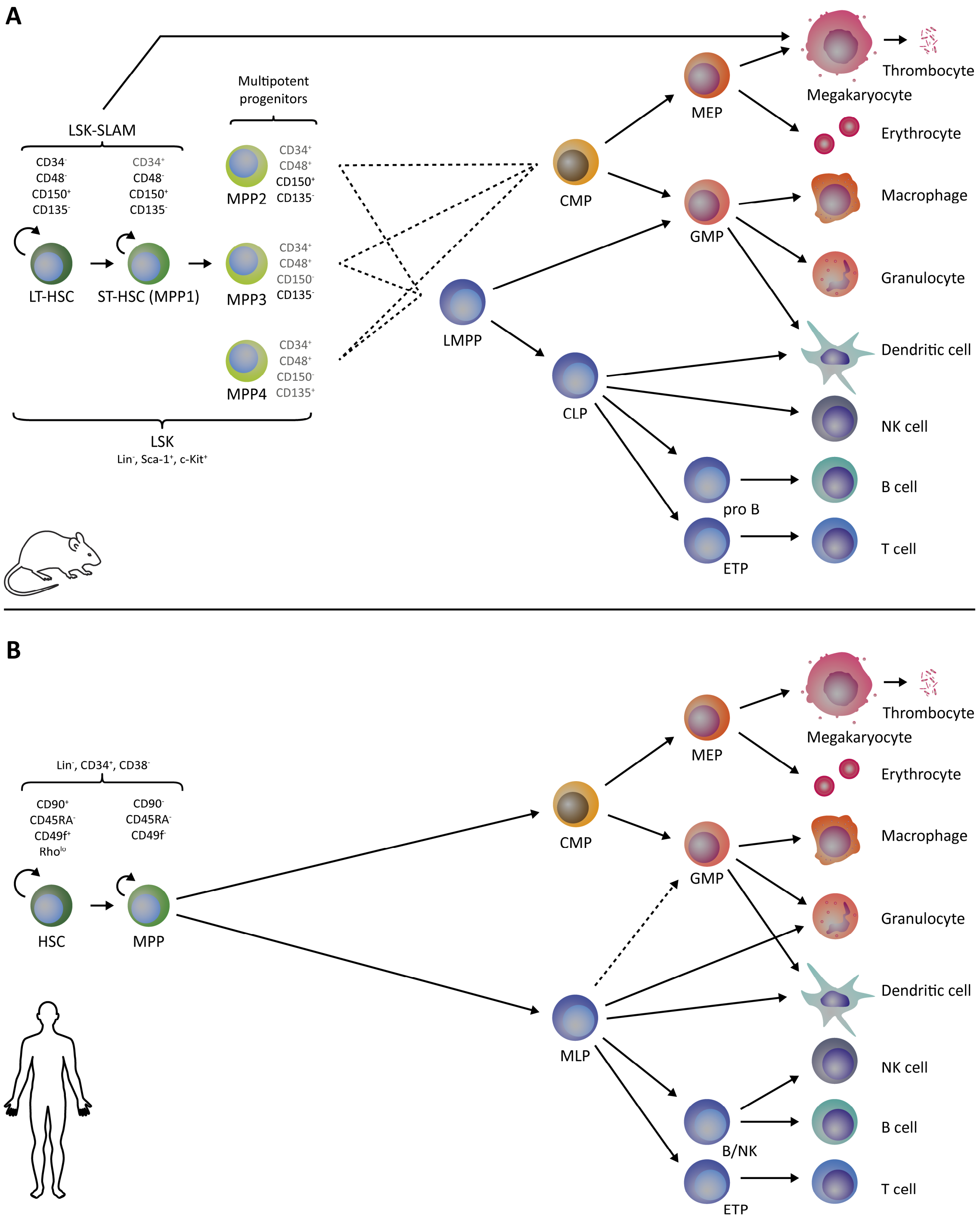


Figure 1 | **Hierarchical organization of the murine and human hematopoietic system.**

A | Simplified concept of the murine hematopoietic lineage tree showing stem and progenitor cells with their most important surface marker combinations on the left and fully differentiated blood cells on the right. **B** | Simplified version of the human hematopoiesis, again with the most relevant surface markers for stem and progenitor cells. For both panels, selected intermediate populations are not depicted for clarity. HSC, Hematopoietic stem cell; ST-HSC, Short-term hematopoietic stem cell; MPP1-4, Multipotent progenitors (1-4); CMP, Common myeloid progenitor; CLP, Common lymphoid progenitor; MEP, Megakaryocyte erythroid progenitor; GMP, Granulocyte macrophage progenitor; MLP, multilymphoid progenitor; ETP, earliest thymic progenitors; pro B, pro B cells; B/NK, B cell NK cell precursor. Adapted from Doulatov et al. (2012) and modified according to Cabezas-Wallscheid et al. (2014), Haas et al. (2015), Wilson et al. (2008) and Rieger and Schroeder (2012).

1.1.3 The hematopoietic system – discrete vs. continuum-based models

Classically, the hematopoietic system is regarded as a series of divisions and differentiation events originating from a population of homogenous multipotent HSCs that reside at the apex of a hierarchically organized branching tree. In this classical model, the maturation of a primitive precursor cell towards a fully differentiated effector cell is characterized by a compulsory stepwise progression through intermediates (Figure 1 and Figure 2A). However, it is important to note that this model has recently been challenged by studies showing significant heterogeneity of HSCs in terms of self-renewal capacity and lineage biases (Morita et al., 2010; Notta et al., 2016; Velten et al., 2017) or even direct maturation of HSCs into megakaryocytes (Haas et al., 2015). As a result, new models were proposed such as the early-split model or the continuous Waddington-like model. Evidence for early HSC lineage separation (early-split) arose from studies showing uni-lineage output of single phenotypic HSCs in transplantation experiments or that common-myeloid progenitors (CMPs) are a mixture of committed uni-lineage cells that already lost their presumed oligo-potency (Karamitros et al., 2018; Rodriguez-Fraticelli et al., 2018) (Figure 2B). Partially in line with this are observations from single cell RNA-seq experiments that showed that HSCs gradually acquire lineage biases instead of transitioning from one discrete state to another. It was also noted, that HSCs directly give rise to uni-lineage restricted cells from a so called continuum of low-primed undifferentiated hematopoietic stem and progenitor cells (CLOUD-HSPCs) (Haas et al., 2018; Velten et al., 2017) (Figure 2C).

The causes for the HSC heterogeneity are still subject of intensive investigations. Until now, several determinants have been suggested, such as the location of the stem cell in the BM niche or the genetic and epigenetic heterogeneity. However, also transcriptional and metabolic activity, segregation of cell fate determinants or simply stochasticity may also serve as distinct sources of HSC heterogeneity (Haas et al., 2018).

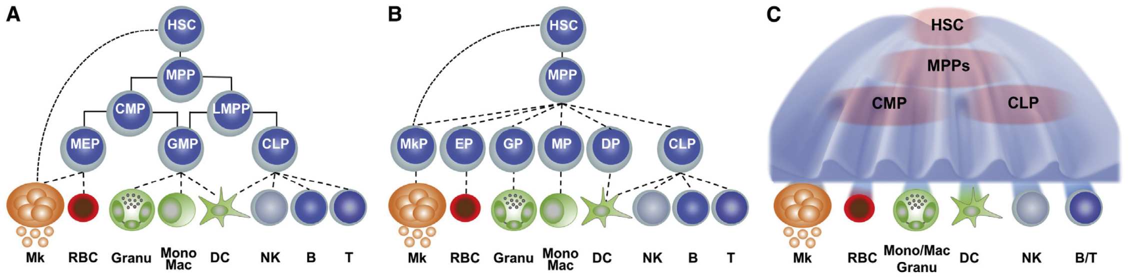


Figure 2 | **Classical and modern models of the hematopoietic tree.**
A | In the classical model, a series of division and differentiation events occur, originating from a population of homogenous multipotent HSCs that reside at the apex of a hierarchically organized branching tree. **B** | The early-split model, in which HSCs and MPP are mostly determined in their lineage potential. **C** | The continuous Waddington-like model, in which HSCs undergo a continuous lineage commitment. In this model, progenitor populations such as MPPs, CMPs or CLPs do not resemble stable cell types but rather transitory states. Reprinted from (Haas et al., 2018), copyright 2018, with permission from Elsevier.

1.1.4 Post-transplant versus unperturbed steady-state hematopoiesis

Transplantation experiments have greatly advanced our current understanding of the hematopoietic system, but their results have always been limited by the fact that it might not represent the normal physiological situation. Recently, non-invasive *in situ* fate mapping of HSCs has been successfully employed to study hematopoiesis in mice in an unperturbed setting. These new insights have uncovered major differences between normal and post-transplant hematopoiesis such as the number of actively contributing clones. During unperturbed steady-state hematopoiesis in mice, blood production is believed to be maintained by a large number of MPPs, which alternate between proliferation and dormancy. HSCs are also actively contributing to hematopoiesis, but to a much smaller degree (Busch et al., 2015; Sun et al., 2014). In contrast, post-transplant hematopoiesis is driven by a much smaller number of clones which are active over a much longer period of time. Also, the contribution of HSCs and progenitors to the blood system was found to change at different time points. After HSC or BM transplantation in humans and primates, the hematopoietic reconstitution is believed to occur in two major waves. A short-term reconstitution phase, lasting about 6-12 months and mostly driven by progenitors is followed by a long-term reconstitution phase, starting around 6-12 months after transplantation, mostly driven by HSCs but also long-term MPPs (Biasco et al., 2016; Kim et al., 2014) (Similar observations were made in mice). Moreover, due to the long-lasting contribution to the blood production, transplanted HSCs are required to have a much higher self-renewal rate compared to HSCs in steady-state. However, the major advantage of transplantation settings is that it naturally selects for self-renewing cells that have the ability to repopulate an entire organism, hence fulfill all HSC-defining criteria. In contrast, models to study unperturbed hematopoiesis again rely on phenotypic definitions of HSCs to use HSC specific loci for the transgene expression, leading to biases towards HSC subsets which might not represent the whole populations (Busch and Rodewald, 2016).

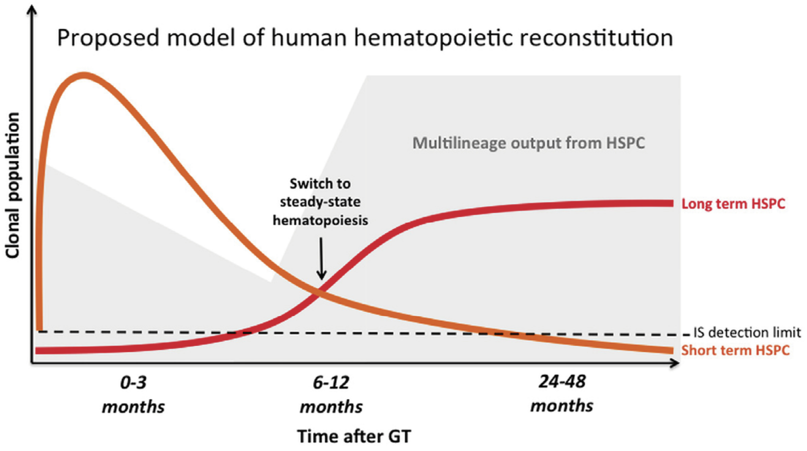


Figure 3 | **Model of clonal dynamics after HSPC transplantation.**
Lentiviral integration sites were used to track individual clones after autogenetic HSPC transplantation leading to the proposed model of human hematopoietic reconstitution. GT, gene therapy. Reprinted from Biasco et al. (2016), with permission from Elsevier.

In summary, the endeavor to maintain the cellular composition within the hematopoietic system, to respond to intrinsic and extrinsic stimuli and to repopulate an entire organism after transplantation through self-renewal, proliferation and differentiation resembles an extremely complex task for HSCs. As a consequence, HSC behavior needs to be tightly controlled, which in turn is regulated through the spatiotemporal activity of genes and their gene-regulatory regions.

1.2 Gene regulation in eukaryotes

The proportion of protein-coding genes within the genome is almost identical across all metazoans regardless of their biological complexity and constitutes only about 1.5 - 3% of the genome. In contrast, the amount of non-protein-coding DNA (ncDNA) positively correlates with the biological complexity of the organism and is nowadays appreciated as one basic prerequisite for complex life (Liu et al., 2013). Parts of the ncDNA comprise of regulatory elements and non-coding RNAs (ncRNA) required for orchestrating the spatiotemporal expression of genes during development, maintenance and homeostasis in all tissues at all times. Consequently, its understanding is fundamental to many biological processes, physiological as well as malignant. To this date, many different types of gene-regulatory regions are known, which interact with the gene promoter trough large protein complexes to modulate gene expression. In the following paragraphs, the most important regulatory elements, such as promoters, enhancers, insulators and three-dimensional chromatin organization are discussed in greater detail.

1.2.1 Promoters

The expression of genes can be regulated at various nodes, but always require a gene promoter to initiate the transcription process. This stretch of DNA is located close to the transcription start site (TSS) and contains multiple DNA consensus sequences, such as TATA-binding protein (TBP) binding sites, initiator elements (Inr), transcription factor II B (TFIIB) recognition elements, downstream core (DCE) elements or motif ten (MTW) elements. The composition of these elements varies between promoters and plays a crucial role in the assembly of the transcriptional machinery and thus gene regulation (Smale and Kadonaga, 2003). In eukaryotes, these promoter sequences are recognized by one of three structurally similar RNA polymerases (RNAP) – RNAPI, RNAPII and RNAPIII, which are responsible for the transcription of DNA to RNA. While RNAPI and RNAPIII transcribe ribosomal-, transfer-, and other small RNAs, all protein-coding genes, miRNAs and some other small RNAs are transcribed by RNAPII, which cooperates with so called general transcription factors (GTFs). Among others tasks, these GTFs are required for the precise positioning of the RNAPII at the TSS, recognizing DNA sequences such as the TATA-box or stabilizing the RNAPII interaction with TBP and TFIIB or recruiting and regulating transcription factor II H (TFIIH), which possesses DNA helicase activity to help unwinding the DNA and revealing the template strand (Alberts et al., 2008). The complex of RNAPII and a minimum of five GTFs are termed the preinitiation complex (PrIC), which on its own is not sufficient for the transcription of genes *in vivo* (Figure 4A). Due to complex chromatin structures, RNAPII also requires transcriptional activators, a mediator, histone modifying enzymes and chromatin remodeling proteins. The transcriptional activators are essential to guide the RNAPII to the desired TSS, which is followed by the interaction of RNAPII with the mediator. The mediator is a large protein complex that ensures the communication with the PrIC and activating proteins, histone modifying enzymes and chromatin remodeling complexes (Figure 4B). These activating proteins are usually TFs, which not only recognize a specific DNA sequence but also contain activation domains. Due to the size and multiple subunits of the mediator, multiple activation domains of different TFs can interact simultaneously, which facilitates enhancer-promoter gene looping (Figure 4B), transcription inhibition mediated by insulators or even the organization of DNA into topological domains (Allen and Taatjes, 2015). Given the essentiality of these three types of interactions, the following paragraphs will explain them in more detail.

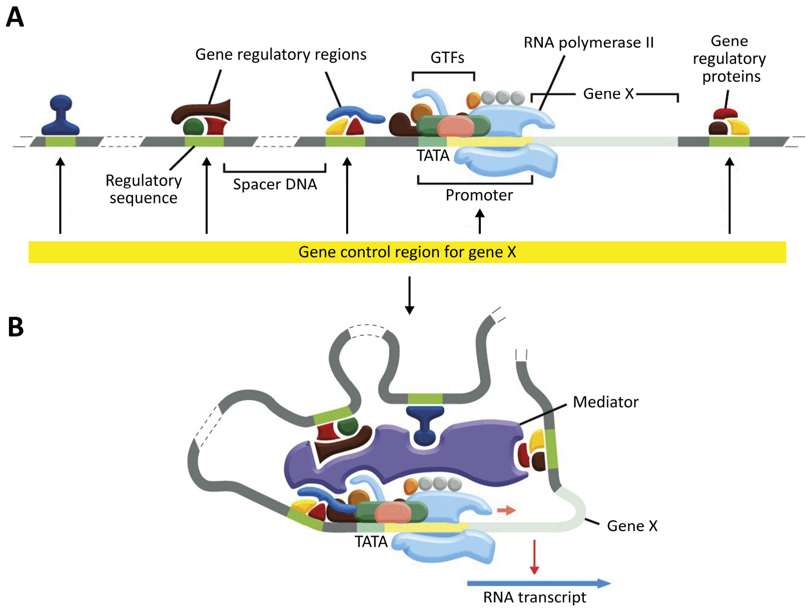


Figure 4 | **Simplified view of the transcription initiation.**
A | General assembly of the RNAPII and GTFs at the promoter region of gene and its gene control region, including regulatory sequences or regions which can be occupied by gene regulatory proteins, such as TFs.
B | Transcription is initiated by interaction with RNAPII, GTFs and the mediator, which links the regulatory regions and their activating TFs to form a DNA loop. From: Molecular Biology of the Cell by Alberts (2008), Reproduced with permission of Taylor & Francis Group in the format Thesis/Dissertation via Copyright Clearance Center.

1.2.2 Enhancers

Enhancers or *cis*-regulatory modules (CRMs) were first discovered in the SV40 virus genome more than 30 years ago (Banerji et al., 1981). Since then, enhancers have been studied extensively in multiple organisms and their understanding has helped to unravel many longstanding questions regarding the complexity of gene regulation. While enhancers were readily known for their important roles during organismal development by regulating the spatiotemporal expression of many key factors, their importance also gained increasing awareness during disease development such as cancer (Sur and Taipale, 2016). Enhancers contain short DNA consensus sequences that are recognized by sequence-specific TFs. These TFs can be repressive or activating and influence the state and activity of RNAPII and the GTFs (Figure 5A). Interestingly, some enhancer locate as far as 1Mb or even 1.7Mb away from their target gene, a distance that would never be bridged by protein assemblies alone on a linear stretch of DNA. Only the three-dimensional structure or DNA-looping of the DNA makes the physical interaction between promoters and enhancers possible (Amano et al., 2009; Bahr et al., 2018; Shlyueva et al., 2014) (Figure 5B and C). The modular nature of enhancers adds another level of complexity and fine-tunes expression through multiple TF-binding sites that can act either additively or redundantly. Here, TFs can regulate transcription on different levels, e.g. through recruitment of the transcriptional machinery, thereby initiating transcription (Figure 4B), or through regulating elongation and termination (Ong and Corces, 2011). This, in combination with tissue and/or developmental stage-dependent expression of TFs, provides the spatiotemporal regulation of gene expression that is essential to complex life. In some cases, enhancers can also actively repress gene transcription through binding of repressive TFs, mainly found during development (Shlyueva et al., 2014). However, TF-mediated repression of gene expression is classically accomplished through silencers – DNA sequences similar to enhancers but primarily bound by repressive TFs.

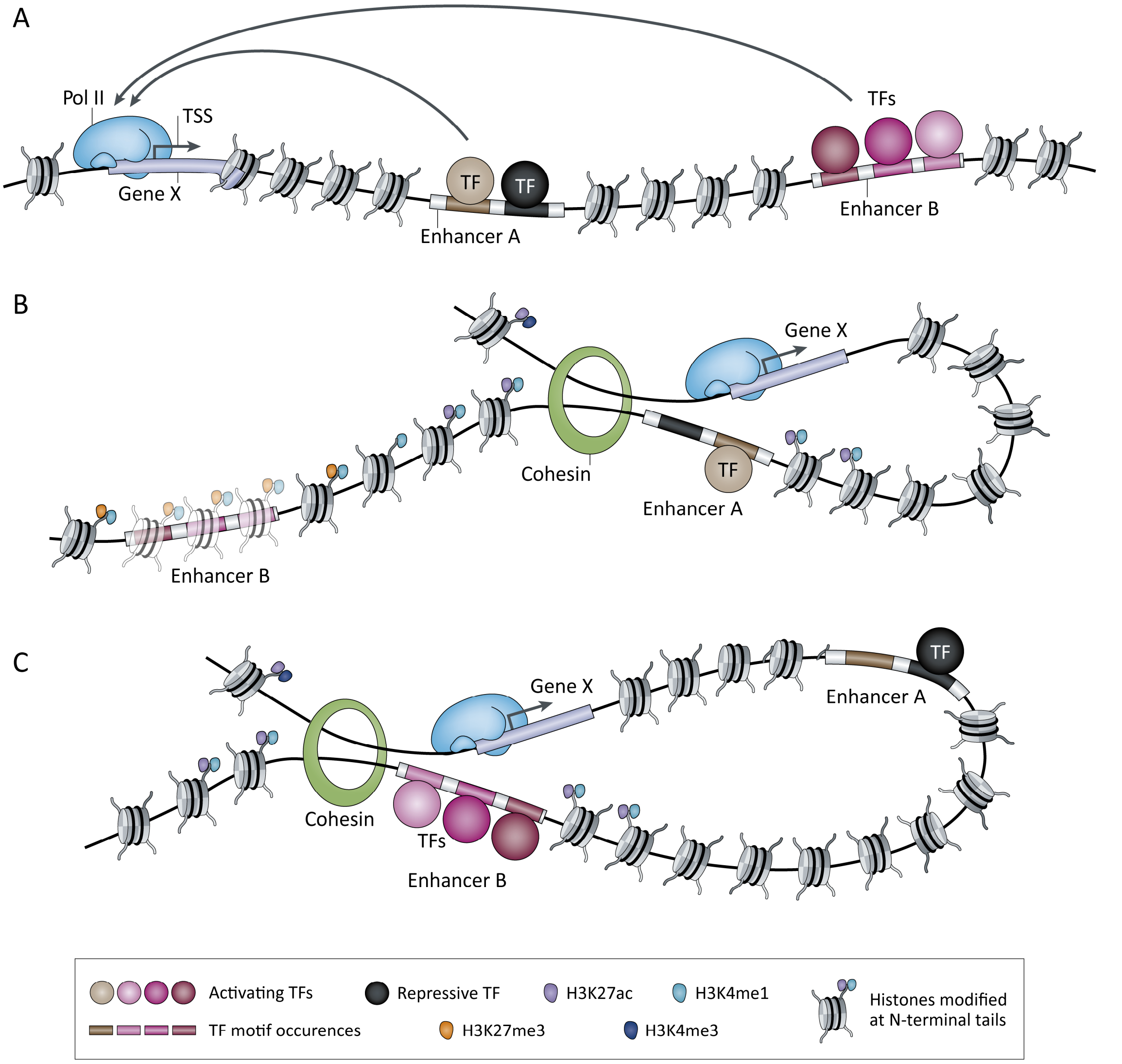


Figure 5 | **Gene regulation through enhancer mediated transcription.**
A | Genomic locus with consensus sequences for different TF that enhance or repress the transcription of gene X. **B** | Through looping of the DNA, gene X comes into close proximity to Enhancer A that is regulating its expression. The DNA-loci are kept in spatial vicinity through the restraining by cohesins. **C** | Depending on the loop size and the location of cohesin, different enhancers can regulate the same gene. In this configuration enhancer B associates with gene X while enhancer A is occupied with repressive TFs. Adapted by permission from Springer Customer Service Centre GmbH: Springer Nature, Nature Reviews Genetics, Shlyueva et al. (2014), copyright 2014.

1.2.3 Insulators

To prevent inappropriate binding of TF-bound regulatory regions to non-target genes, enhancer and promoter interactions can be controlled with so called insulators. These insulators can block the communication between genes and enhancers by e.g. maintaining the barrier between euchromatin and heterochromatin. Other mechanisms include promoter mimicking (Geyer, 1997) or acting as a physical barrier to interfere with RNAPII (Zhao and Dean, 2004). However, these mechanisms have been challenged and may not apply generally. A third mechanism by which insulators block the crosstalk between enhancers and non-target genes is the compartmentalization into discrete regulatory domains. This is largely mediated by the insulator protein or transcriptional repressor CCCTC-binding factor (CTCF) and the protein complex cohesin. DNA-bound CTCF proteins can form homodimers and this way cause the DNA to form loops. These contact points in the DNA are then reinforced by cohesin ring structures (Figure 5B). Due to the essential roles of CTCF and cohesin in mediating chromatin contact loops, their function is regarded as both, inhibiting but also facilitating the communication between enhancers and promoters (Ali et al., 2016; Hou et al., 2008). Apart from their role of forming the anchors of DNA loops, CTCF and cohesin also co-localize at boundaries of topologically associated domains (TADs) suggesting that these proteins also convey higher order genomic structures at megabase-level (Dixon et al., 2012; Rao et al., 2014). In contrast to the cell-type and developmental stage-dependent DNA loops, TADs are highly conserved and stable across cell types. The following paragraph describes TADs and the technological advances that led to their discovery in more detail.

1.2.4 Higher order genome structures through topologically associated domains

In 2002, Dekker and colleagues invented a new method called chromatin conformation capture (3C) and paved the way for studying the genomic structure in eukaryotes in a three-dimensional space (Dekker et al., 2002). In brief, the employed method uses formaldehyde cross-linking of the chromatin followed by enzymatic digestion to break down the DNA into smaller pieces, which are eventually cross-linked through ligation (Dekker, 2006). The resulting stretch of DNA contains two fragments, which might have been distant on linear DNA but came into close proximity in the cell and therefore were covalently bound during the treatment with formaldehyde. Through quantitative PCR (qPCR), the abundance of certain ligation products can be measured, which conveys direct information about the frequency with which these two loci interact. Soon after, advancements of this methods were developed, with the genome wide chromosome conformation capture-on-chip (4C) (Simonis et al., 2006) and chromosome conformation capture carbon copy (5C) (Dostie et al., 2006), eventually leading to Hi-C (Lieberman-Aiden et al., 2009) a technology that uses massive parallel sequencing to capture all genomic interactions and eventually led to the discovery of TADs. In order to further refine the method and to overcome the tremendous complexity of such libraries, Mifsud et al. (2015) developed capture Hi-C (CHi-C). This technique is based on the original Hi-C technology, but involves a solution hybridization selection step that enriches for selected genomic regions and therefore massively increases the resolution for sites of special interest e.g. promoters.

The properties of TADs are diverse and have still not been fully understood, however, some fundamental features have been unraveled (Figure 6A-E). In many regards, TADs function similar to loops, which are themselves part of TADs and make up the so-called sub-TADs. Within TADs, genes can be co-regulated by the same enhancer, while genes outside are blocked from that interaction through the insulating TAD boundary (Figure 6A and B). In fact, genes with similar functions are often found to cluster within TADs, such as olfactory receptor genes (Figure 6E). These boundaries also prevent repressive or active chromatin from spreading or even block divergent spread of transcription (Figure 6C and D) (Dixon et al., 2016).

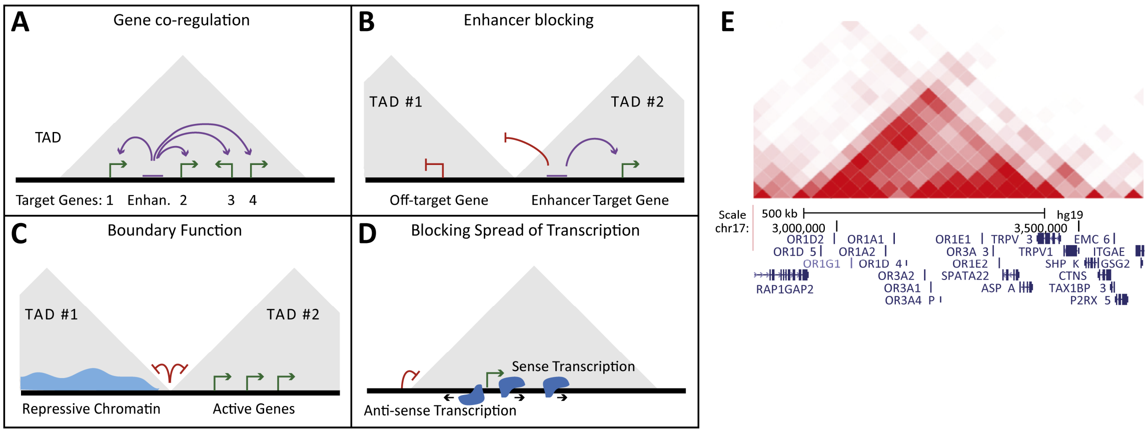


Figure 6 | **Different modes of action of TADs in genome regulation.**
A | A single enhancer co-regulates multiple target genes within the boundaries of a single TAD. **B** | Enhancers activity is restricted to genes within the same TAD and cannot influence gene expression across boundaries. **C** | TAD boundaries can prevent spreading of repressive or active chromatin to neighboring territories. **D** | TAD boundaries also serve to block divergent spread of transcription. **E** | Exemplary Hi-C data showing the interaction heatmap for an approx. 500kb wide TAD that spans around a cluster of olfactory receptor genes. Reprinted from Dixon et al. (2016), copyright 2016, with permission from Elsevier.

Interestingly, TAD boundaries are not only enriched for CTCF binding sites but also for TSS (particularly TSS of housekeeping genes), transfer RNAs and short interspersed element (SINE) retrotransposons (Dixon et al., 2012; Gorkin et al., 2014). Additionally, computational analysis revealed that TADs are also positively associated with H3K36me3 sites, TSS of mRNA and ncRNA genes, RNAPII and other specific TFs, indicating that these regions are transcriptionally active, despite being depleted from DNase I-hypersensitive sites (Figure 7).

In summary, higher order genome organization is a highly essential feature of complex gene regulation in eukaryotes and recent technological advances such as Hi-C have greatly helped to understand its core features. Nonetheless, many questions remain and require additional research to complete the picture.

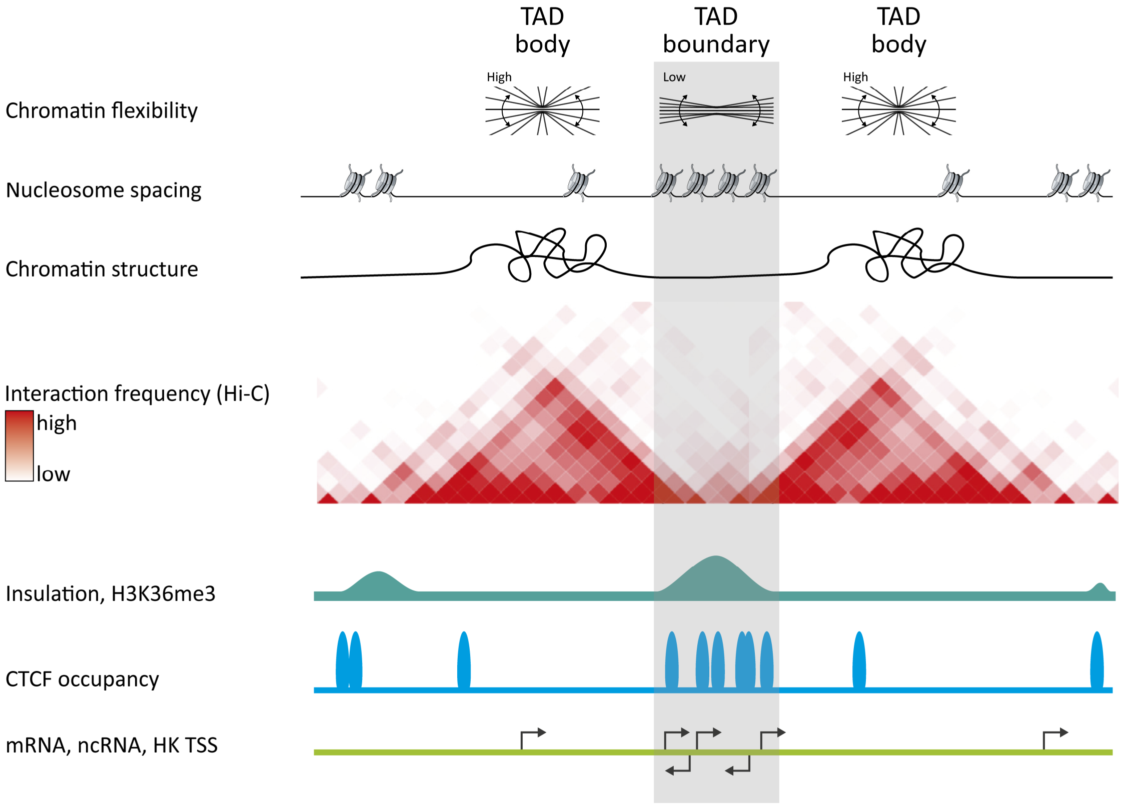


Figure 7 | **TAD and TAD boundary properties including enrichment of various genomic features.** Schematic representation of two adjacent TADs. Due to the nucleosome spacing, chromatin flexibility is low at boundaries and highest in TADs, allowing for loop formation in sub-TADs. Insulation is highest at boundaries in line with the high occupancy of CTCF binding sites which inhibit e.g. cross-talk between enhancers of different TADs. Despite high nucleosome density at boundaries, histone modification that mark active gene bodies (H3K36me3) are enriched at these sites, which is in line with enrichment of TSS for mRNAs, ncRNAs and most prevalently TSS of housekeeping genes. **HK**, housekeeping gene. Based on Dixon et al. (2016) and (Hong and Kim, 2017)

1.3 Identification of regulatory regions

The identification of regulatory regions, such as the aforementioned promoters, enhancers and insulators can be challenging. Although gene promoters can be predicted using the DNA sequence using e.g. the TATA-box sequence, their activity is cell type specific and therefore needs to be addressed in individual cell types through e.g. measuring mRNA levels or through reporter constructs. In contrast to promoters, enhancers and insulators cannot be identified using the DNA sequence alone but instead require the analysis of the epigenome. Similarly to gene promoters, their activity or even presence is cell type specific, so again requires a cell type specific analysis to map their activity. However, there is also a more general way of identifying active regulatory regions – assessing the accessibility of the genome.

1.3.1 Accessibility of chromatin

Mapping active regulatory regions using the chromatin accessibility grounds on the notion that condensed or inaccessible chromatin is associated with no transcriptional or regulatory activity, while loci that are actively transcribed are "open". For example, DHS-seq is a commonly used technique that uses the DNaseI restriction enzyme, which cleaves the DNA only at nucleosome-free regions. In combination with high-throughput sequencing (HT-seq) this reveals a genome wide map of accessible or active sites and thus putative regulatory regions (Boyle et al., 2008; Thurman et al., 2012). A newer technique called assay for transposase-accessible chromatin using sequencing (ATAC-seq) uses the hyperactive Tn5 transposase coupled with HT-seq adapters. Tn5 also integrates into accessible genomic regions, while compact DNA renders integration less probable. Compared to DHS-seq, ATAC-seq is much more sensitive as it requires a fraction of the starting material and is thus also applicable to low-input samples such as rare primary cell populations (Buenrostro et al., 2013; Buenrostro et al., 2015). Nonetheless, both of the above described methods are not capable of distinguishing between enhancers, promoters, silencers, insulators, locus control regions or any other regulatory regions, but instead only provide a broad overview of regions of open chromatin. The specific identification of enhancer can be achieved by other means as discussed below.

1.3.2 Identification of enhancers

The identification of enhancers involves many challenges and differs substantially from the identification of e.g. protein-coding genes. Despite intensive research, to date there is no single feature known that is common to all enhancers that would aid the prediction from the DNA sequence alone. Although some enhancers have been identified through sequence conservation, it cannot be applied invariably to confidently predict enhancers due to their rapid evolution or species specificity. Also, the classical mutation-phenotype approach that is still used to identify and characterize most genes has its limits: On the one hand, there are about 1 million putative enhancers in the mammalian genome, 20-fold more compared to the approx. 50,000 gene promoters. On the other hand, a lack of phenotype after genetic perturbation cannot preclude any functional relevance, as enhancers are often redundant and highly contextual (Coppola et al., 2016).

Recently, deep-sequencing approaches have been developed that exploit multiplexed reporter assays to measure transient RNA expression of tens to thousands of elements in parallel (Melnikov et al., 2012; Patwardhan et al., 2012). The self-transcribing active regulatory region sequencing (STARR-seq) method uses a genome wide library of randomly sheared genomic DNA fragments. These fragments are cloned into reporter plasmids in-between a minimal promoter and a poly-A site and transfected into the cells of interest. Fragments that contain transcriptionally active enhancer sequences self-transcribe or self-amplify inside the cells, which can be measured by extracting poly-A mRNA, reverse transcription and high-throughput paired end sequencing. This way, enhancer activity for millions of candidates can be assessed in parallel in an unbiased and quantitative way (Arnold et al., 2013). However, most of the above-mentioned methods require the manipulation of cells through e.g. CRISPR/Cas9-mediated genome editing for mutation-phenotype approaches or transient transfection and cultivation of large libraries for reporter-based assays. This largely restricts their applicability especially for rare and/or primary cell populations.

Another commonly used and well-established method is to identify putative enhancers through the assessment of TF binding or chromatin state by chromatin immunoprecipitation sequencing (ChIP-seq). Using ChIP-seq data, different chromatin states can be identified, all of which are characterized by different properties:

1.3.2.1 Chromatin states and histone modifications

In eukaryotes, DNA is wrapped around nucleosomes – a histone octamer – which resembles the basic structural unit of chromatin. Different histone types (H2A, H2B, H3 and H4) and chemical modifications of histone residues dictate the primary structure of chromatin. The development of ABs against distinct histone modifications laid the foundation for ChIP-seq experiments, which in turn have led to a tremendous gain in the understanding of how histone modifications control the activity of genomic elements like enhancers (Zhou et al., 2011). While the presence or absence of single histone modifications facilitates the prediction of chromatin states to some degree, predictions are made more reliable using combinatorial histone modification signatures. A landmark in identifying and allocating different modification combinations was the development of a generative machine-learning multivariate hidden Markov model (ChromHMM) (Ernst and Kellis, 2017; Ernst et al., 2011), leading to the definition of distinct chromatin states. Here, enhancers are categorized into genic, active and weak enhancers, all of which carry histone 3 lysine 4 monomethylations (H3K4me1). However, active enhancers additionally carry H3K27 acetylation (K3K27ac) marks while genic enhancers carry H3K27ac and H3K36 trimethylation (H3K36me3) marks (Figure 8).

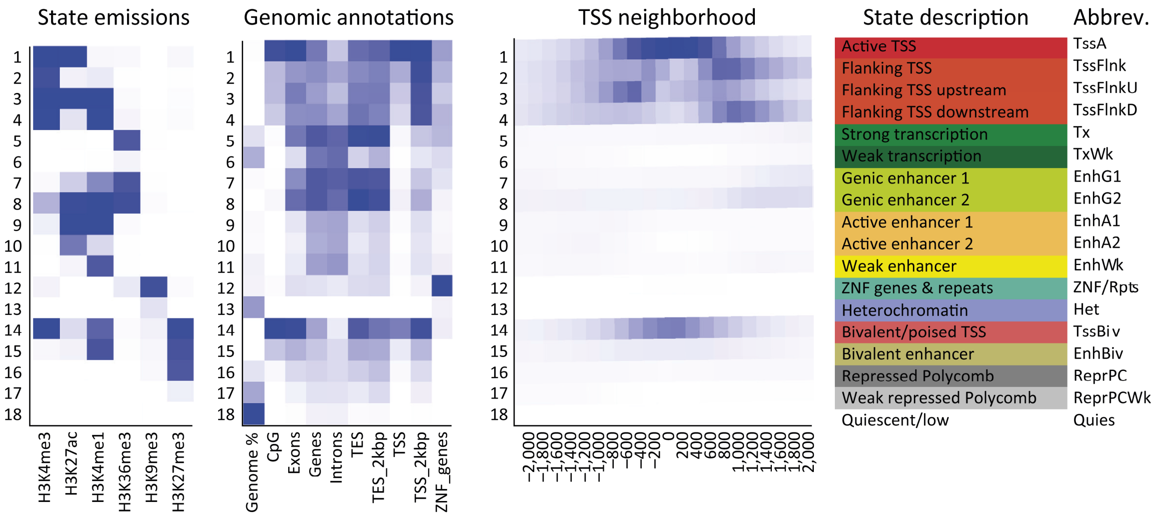


Figure 8 | **Overview of Roadmap Epigenomics 18-state expanded model ChromHMM chromatin states.** The state emissions describe the quantitative and qualitative combination of histone modifications for all 18 states. The color intensity corresponds to the probability of observing the mark in the state. The genomic annotations describe the fold enrichment of the indicated genomic annotations found in IMR90 cells with the color intensity being proportional to the fold-enrichment. The TSS neighborhood depicts the enrichment of the state in a 2kb window around a set of TSS. Darker colors correspond to a higher fold-enrichment. State descriptions for all states with commonly used abbreviations are indicated on the very right. Adapted by permission from Springer Customer Service Centre GmbH: Springer Nature, Nature Protocols, (Ernst and Kellis, 2017), copyright 2017.

1.3.3 Medical impact of sequence alterations in regulatory regions

Naturally, defining an alterations in a sequence always implies a reference sequence for comparison. In human genomics, this reference sequence is usually the average of the human population, measured through thousands of whole genome sequencing projects. Although 99.5% of the genome is identical between any two humans, the remaining 0.5% can make a huge difference, like hair color, skin tone or even medical predispositions. Alterations in a single nucleotide that occur in more than 1% of the human population are termed single nucleotide polymorphisms (SNPs) and are different to classical mutations, which occur at a frequency below 1% (Karki et al., 2015). SNPs and mutations in exons of protein-coding genes can change the amino acid sequence of a protein and consequently alter its function. The identification of the gene that is affected and the resulting change in the amino acid sequence is easy to assess and can therefore be simply linked to a disease. However, many mutations or SNPs occur in intragenic (intronic) or intergenic regions with unknown impact on gene regulation thus pathology. Genome-wide association studies (GWAS) aim to link these mutations or SNPs to diseases or traits. The experimental design involves the comparison between subjects with a given disease or trait and healthy controls. In order to assess the polymorphisms of the two groups, either whole genome sequencing data or sequence information from genotyping or SNP arrays is used. Next, frequencies of SNPs in both populations are statistically compared to assess the linkage disequilibrium (LD), a measure of non-random association of alleles at difference loci (Slatkin, 2008). In other words, a given SNP or combination of SNPs occur significantly more often in cases than controls, which implies a certain risk of developing the trait or disease when carrying the variant. As only about 2% of the human genome contains protein-coding genes it comes as no surprise that more than 85% of GWAS risk variants are located in inter- and intragenic DNA – preferably enhancers. Because of that, it remains a challenge to determine the cellular and organismal consequences these SNPs cause. It is thought that SNPs in enhancers alter e.g. DNA-protein interaction thus influencing gene expression, however the gene or genes, which are directly affected have to be identified experimentally (Corradin and Scacheri, 2014). Nonetheless, despite the lack of target gene information, enhancers that carry GWAS SNPs can at least be assigned to traits or diseases, making GWAS a powerful tool for providing new insights into mechanisms in common diseases.

1.4 Hematopoiesis and its correction in the context of disease

As described above, hematopoiesis is a fine-tuned process maintained through the interplay of HSCs, progenitors and differentiated cells. Mutation in the genome of these cells can alter or diminish their function, leading to phenotypes ranging from mild symptoms to severe defects or even to death of the affected individual. One example is the Wiskott-Aldrich-Syndrome (WAS), a recessively inherited primary immune deficiency. Diseases like WAS can be treated by allogenic BM transplantations – BM or hematopoietic stem cells, respectively, from another healthy individual. While allogenic BM transplantations are unproblematic in inbred mouse colonies, the genetic variation between humans can cause the immune system to recognize the transplant as foreign and provoke Graft versus Host Disease (GvHD) or complete graft rejection. A disparity between recipients is sensed by human leukocyte antigens (HLAs) expressed on T cells and NK cells. To prevent GvHD or rejection, donors with matching HLAs are crucial, however often hard to find (Nowak, 2008). An alternative path is to provide the patients' cells with a correct version of the malfunctioning gene – a relatively new treatment option called gene therapy (GT).

1.4.1 Gene therapy for the treatment of Wiskott-Aldrich-Syndrome

The Wiskott-Aldrich-Syndrome (WAS) is a rare primary immune deficiencie with a frequency of about 1-10 males per million and is characterized by low platelet counts (thrombocytopenia), skin rashes (eczema) and recurrent severe infections, leading to an average life expectancy of less than 10 years. The syndrome was first described in 1954 by the German physician Alfred Wiskott and the American Robert Anderson Aldrich (Aldrich et al., 1954) and later linked to a mutation in the Wiskott-Aldrich-Syndrome Protein (WASp) that renders it malfunctional (Derry et al., 1994). Expression of WASp is thought to be induced in dendritic cells via T cell receptor signaling in order to form an immunological synapse through actin cytoskeletal rearrangements, making it essential for proper immune function (Malinova et al., 2016). Apart from that, importance of WASp for the regulation of T cells, B cells, NK cells to maturation and function of myelomonocytic cells has also been reported (Ochs and Notarangelo, 2005).

As described above, WAS patients are treated by allogeneic HSC transplantations (Albert et al., 2011), given that a HLA-matched donor is available. A relatively new treatment strategy for patients without a suitable donor is GT, which describes the process of inserting DNA or RNA into body cells as a drug to treat a specific disease. While already attempted in the 80s, the first successful trial of inserting human DNA into the genome was performed in 1990 (Rosenberg et al., 1990). Following this, a large number of GT trials were carried out until today with growing success. In the present study, we used data from a GT trial on 10 WAS patients that did not have an HLA-matching donor. In this trial, a functional wild-type copy of the *WASP* gene was stably inserted into CD34$^+$ cells using γ-retroviruses (γRVs) as vectors for gene delivery. The CD34$^+$ cells were extracted from peripheral blood of the patients after mobilization with Granulocyte-colony stimulating factor (G-CSF) or G-CSF and the CXCR4 inhibitor Plerixafor. After a successful manipulation *ex vivo*, treated cells were autologously transplanted into the patient (Boztug et al., 2010). Because γRVs integrate semi-randomly into the genome, each transduced cell is characterized by a unique integration site (IS). This IS can be traced using highly sensitive Linear Amplification-Mediated

Polymerase Chain Reaction (LAM-PCR) (Schmidt et al., 2007) combined with HT-seq methodologies. Accordingly, a bulk of cells carrying an identical IS must have originated from one transduced stem cell or clone, respectively.

In order to follow stem cell engraftment and clonality during the WAS GT trial, patient blood and BM samples were taken periodically and ISs were amplified using LAM-PCR (Also see Figure 12 in the Results 3.1 section). Despite a general success of the GT by restoring WASp expression and reversing most WAS-associated symptoms, as of 2014, seven out of 10 patients showed malignant clonal expansion due to insertional mutagenesis of the viral vector (Braun et al., 2014). The following paragraphs are discussing the family or *Retroviridae* in more detail and provide information about the specific integration biology of γRVs, insertional mutagenesis and also how γRV integration sites can be exploited for the detection of epigenomic features.

1.5 Biology of retroviruses

The family of *Retroviridae* contains a total of seven different virus genera. In the context of gene therapy and laboratory use, the two most important subtypes are lenti-viruses (LV; e.g. human immunodeficiency viruses; HIV-1) and γRVs (e.g. murine leukemia virus; MLV). The retroviral positive sense RNA genome only ranges from 8 to 11kb in length, yet contains most building blocks required for its entire life cycle. Importantly, retroviruses possess the unique ability to reversely transcribe RNA into double-stranded DNA, an essential step in order to harness the eukaryotic transcriptional machinery and to integrate into the host cell DNA genome. The responsible enzyme – the reverse transcriptase – was discovered in 1970 by Baltimore, Temin and his co-worker Mizutani, a groundbreaking discovery which laid the foundation for numerous laboratory applications and therapeutic approaches (Baltimore, 1970; Temin and Mizutani, 1970). Another remarkable property of retroviruses is their ability to integrate into the host cell genome and reside as proviral DNA, this way multiplying themselves with every doubling of the host cells (Balvay et al., 2007). The proviral DNA can be transcribed again, leading to the production of new viral particles that eventually leave the cell via "budding" and thus close the life-cycle. The mechanisms behind this are very similar between the genera of Retroviridae and start with a small set of proteins encoded by four domains: Proteins of the *gag* domain are required for the viral capsid, *env* domain proteins provide the components for the viral envelope, proteins encoded by the *pol* domain perform DNA synthesis and integration and *pro* domain proteases are required for the maturation of viral proteins (Balvay et al., 2007). Importantly, the understanding of every components role in the life-cycle of the virus has enabled researchers to modify the viral genome to produce new entities for safer laboratory and gene-therapeutical use. Such research has for example led to the development of self-inactivating long-terminal repeats (SIN-LTRs), a modification that reduces the chance of undesired activation of genes in the proximity of the provirus (Dull et al., 1998; Zufferey et al., 1998).

While cell entry is very comparable between LVs and γRVs, significant differences arise during nuclear entry and integration. LVs do not require the infected cell to divide. Instead the viral capsid with the pre-integration complex (PIC) docks to the nuclear core complex (NPC) before the capsid disassembles and releases the PIC into the nucleus where it integrates into the nuclear laminar-associated DNA (Figure 9A and B). In contrast, the capsid enclosed γ-retroviral PIC first associates with the viral p12 protein, which is stabilizing the complex. Only during mitosis and concomitant nuclear membrane break down, p12 can tether the capsid to the chromosome where it is segregated into the daughter cell nucleus, before the PIC is finally released during mitotic exit and the viral DNA integrates into the DNA (Figure 9A and C) (Demeulemeester et al., 2015).

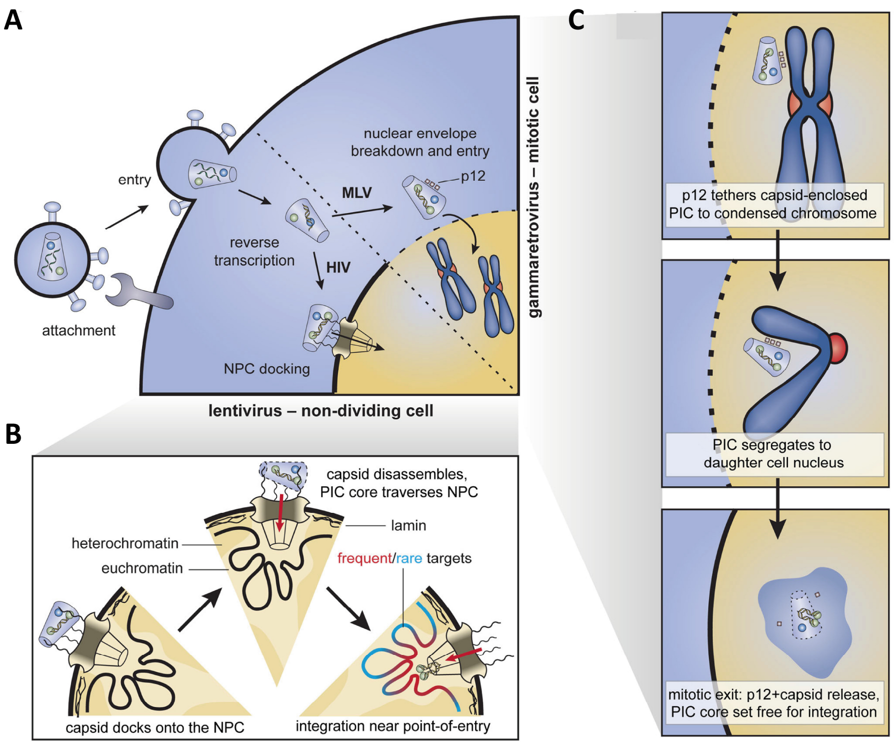
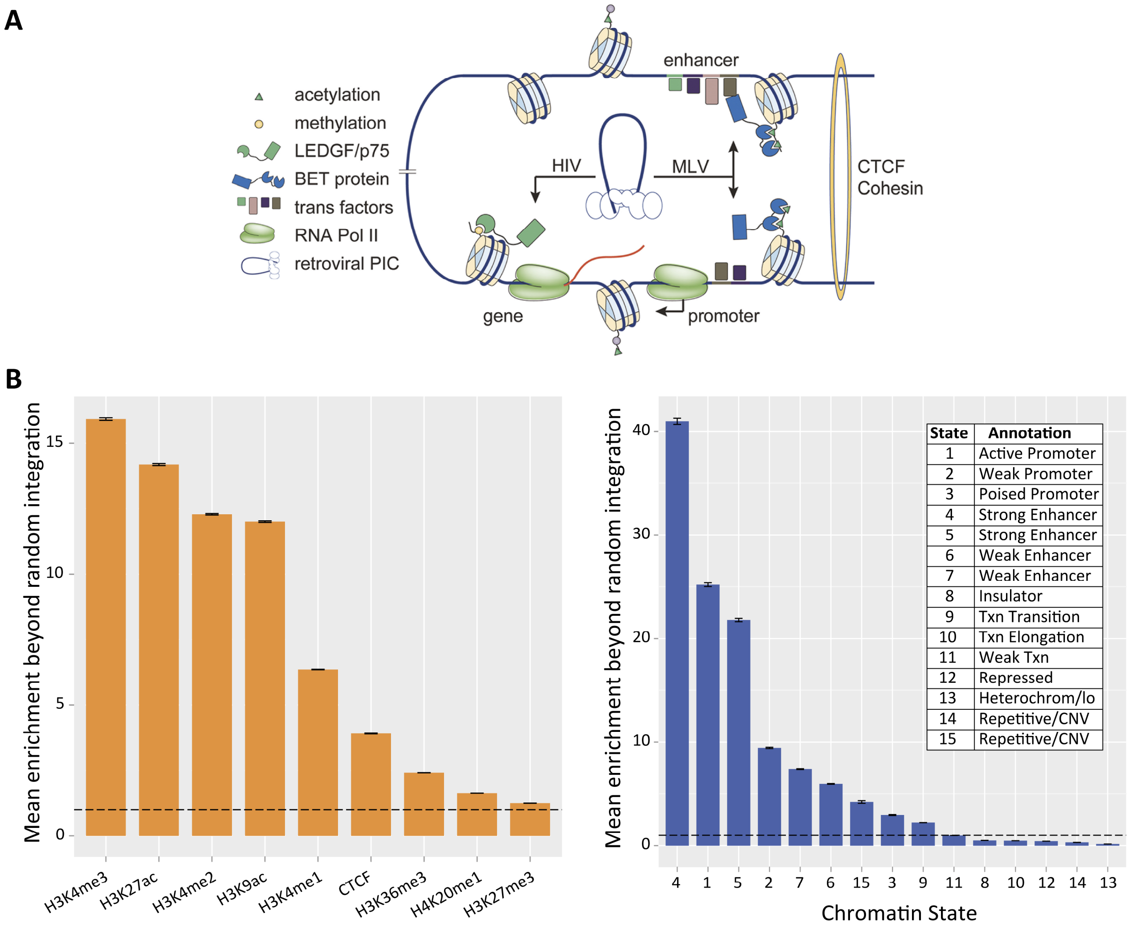

A
entry
attachment
reverse
transcription
nuclear envelope
breakdown and entry
p12
MLV
HIV
NPC docking
lentivirus – non-dividing cell
gammaretrovirus – mitotic cell
B
capsid disassembles,
PIC core traverses NPC
lamin
heterochromatin
euchromatin
frequent/rare targets
capsid docks onto the NPC
integration near point-of-entry
C
p12 tethers capsid-enclosed
PIC to condensed chromosome
PIC segregates to
daughter cell nucleus
mitotic exit: p12+capsid release,
PIC core set free for integration

Figure 9 | **Cell and nuclear entry paths of LV and γRV.**
A | The fusion of the virus and the cell membrane delivers viral proteins and RNA into the host cell were it is reversely transcribed while shuttled to the nucleus. LV capsids are shuffled towards nuclear pores, while γRV capsids associate with the viral p12 protein and await the disintegration of the nuclear membrane during mitosis. **B** | Once the reverse transcription has finished, the LV capsid disassembles and releases the PIC core, which traverses through the NPC and integrates the viral DNA into the outer perimeter DNA. **C** | During mitosis, p12 and tethers the capsid enclosed PIC to condensed chromosomes, is transported to the nucleus of the daughter cell. Finally, p12 and the capsid are released during mitotic exit and the PIC is set free for integration. **PIC,** pre-integration complex; **NPC**, nuclear core complex; Adapted with permission from (Demeulemeester et al., 2015)

1.5.1 Integration biology of γ-retroviruses

In the past decade, γRV integration biology has been studied extensively, not alone in the context of gene therapy trials. Early analysis showed that γ-retroviral proviruses are often located near TSS of active genes. However these analysis were based on either very few ISs or were derived from mutagenesis screens that suffer from substantial IS pattern-skewing (Wu et al., 2006). Only a few years later, after profound advancements in sequencing technologies and more efficient amplification of viral ISs, the understanding of γRV integration became more comprehensive. Many groups reported that active gene-regulatory regions such as enhancers were even preferred over active TSS. This preference is mediated by the interaction of the viral PIC with BET family proteins such as BRD2, 3, and 4 (Cattoglio et al., 2010; De Ravin et al., 2014; Deichmann et al., 2011; LaFave et al., 2014). BET proteins are transcriptional co-regulators and contain two N-terminal bromodomains. The bromodomain modules usually target hyper-acetylated tails of histone H3 and H4, while an extraterminal (ET) domain takes care of the interaction with other cofactors and the γRV PIC. This stands in stark contrast to LVs. Here, the PIC mainly cooperates with Lens Epithelium-Derived Growth Factor/p75 (LEDGF/p75), which specifically binds to H3K36me3-modified nucleosomes, thus active gene bodies (Demeulemeester et al., 2015).

In summary, the recent findings of three independent groups (De Rijck et al., 2013; Gupta et al., 2013; Sharma et al., 2013) soundly establish BET proteins as direct mediators of γ-retroviral target site selection, directing them reliably towards strong enhancers and active promoters.



State	Annotation
1	Active Promoter
2	Weak Promoter
3	Poised Promoter
4	Strong Enhancer
5	Strong Enhancer
6	Weak Enhancer
7	Weak Enhancer
8	Insulator
9	Txn Transition
10	Txn Elongation
11	Weak Txn
12	Repressed
13	Heterochrom/lo
14	Repetitive/CNV
15	Repetitive/CNV

Figure 10 | **Specific chromatin states are deterministic of the location of viral ISs (Legend continued on next page).**

A | The PIC of LVs or γRVs, respectively, hijack intracellular proteins to gain access to specific chromatin environments. The LV PIC mainly cooperates with LEDGF/p75 which steers it towards active gene bodies mostly displaying H3K36me3 marks. In contrast, PICs from γRVs interact with BET family proteins, which in turn read hyper-acetylated histones H3 and H4, thus delegate γRV ISs towards strong enhancers or active promoters. Adapted with permission from (Demeulemeester et al., 2015) **B** | Mean enrichment of γRV ISs at indicated ChIP-seq peaks or chromatin states beyond expected by chance. H3K4me3 and H3K27ac but also H3K4me1 marks are highly enriched at γRV IS, which is reflected by the most enriched chromatin states – Strong enhancers (4) and active promoters (1). (LaFave et al., 2014), by permission of Oxford University Press.

1.5.2 Retroviral insertional mutagenesis

The notion that retroviruses can cause malignant transformation of cells has celebrated its 100th anniversary already some years ago. In 1911, Peyton Rous discovered that a cell-free extract from chicken tumors can induce the same type of tumor in healthy chicken. As the filters were too fine for bacteria or cells to pass, he postulated that the tumor causing reagent had to be a virus – later known as the Rous Sarcoma Virus (RSV) (Rous, 1910, 1911). Subsequent research and the development of a quantitative *in vitro* bioassay for RSV in the 1950s led to the identification of oncogenes, i.e. genes whose enforced expression induce cancer, such as the *src* gene in the RSV genome. Other examples of oncogenes discovered in viruses long before their discovery in humans are *myc* in the avian myelocytoma virus genome or *ras*, first discovered in the rat sarcoma virus (Weiss and Vogt, 2011).

Retroviruses, such as the RSV can transform cells not only by expression of oncogenes from their own viral genome, but also by altering the expression or structure of genes in the host cell genome. This discovery led to the definition of proto-oncogenes, normal regulatory genes that act as oncogenes when overly expressed or mutated by the viral integrate (Bishop, 1983). The mechanisms by which retroviruses induce insertional mutagenesis are manifold and include but are not limited to overexpression of proto-oncogenes by viral enhancer elements, structural alteration of proto-oncogenes through spliced and un-spliced retroviral/cellular fusion transcripts, premature polyadenylation or aberrant splicing of mRNAs or even down-regulation of gene expression (Figure 11A-G) (Knight et al., 2013). However, the frequency at which such events occur remains speculative. In experimental setups, vector integration usually occurs in millions of cells in parallel, making an oncogenic event stochastically very probable. Due to the growth advantage of transformed cells, oncogenic events are naturally selected for, which in turn generates an impression that these events are very common, while the actual rate might be very low.

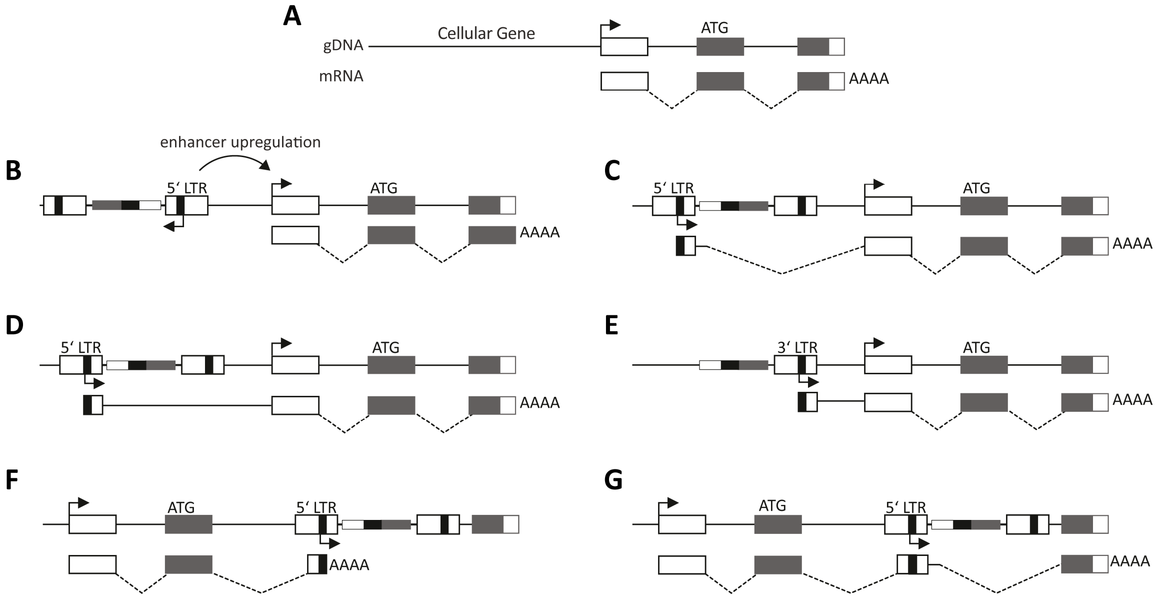


Figure 11 | **Overview of retroviral insertional mutagenesis mechanisms.**
A | Classical gene configuration with TSS (arrow), untranslated regions (white boxes), protein-coding regions with ATG start codon (grey boxes) and polyadenylated (polyA) tail (indicated by AAAA). gDNA (top) and resulting mRNA (bottom) are indicated. **B** | Retroviral enhancer elements in LTR region upregulate the expression of neighboring genes. **C** | Overexpression of the cellular gene due to mRNA fusion transcript. The 5′ LTR is fused via a vector splice site to an exonic splice acceptor. **D** | Overexpression of the cellular gene by fusion after read through of the 3′ LTR. **E** | Fusion of vector and cellular gene initiated by 3′ LTR after deletion of 5′ LTR. **F** | Generation of premature polyA tail after intronic vector integration. **G** | Aberrant splicing after intronic vector integration can lead to fusion transcripts. (Knight et al., 2013) Reproduced with permission of BENTHAM SCIENCE PUBLISHERS LTD. in the format Thesis/Dissertation via Copyright Clearance Center.

1.5.3 Using γRV ISs as molecular tags for active regulatory regions.

Due to the specific target site bias of *Retroviridae*, it seems plausible to utilize vector integrations as molecular tags for certain genomic features. In fact, in a recent study Romano and colleagues used retroviral integration signatures to identify regulatory regions. The authors integrated Cap Analysis of Gene Expression (CAGE), ChIP-seq and Moloney leukemia virus (MLV) integration site mapping in human HSPCs and committed erythroid and myeloid progenitors/precursors (EPP and MPP) to profile the transcriptional and epigenetic changes associated with HSPC lineage commitment. Interestingly, MLV clusters were significantly enriched at super-enhancers (SE) in comparison to normal active enhancers, suggesting that MLV integration sites could even be specifically used for the detection of SEs (Romano et al., 2016).

2 Aims of the thesis

Hematopoietic stem cells (HSCs) resemble a small population of cells with a wide range of properties. To reconstitute an entire blood system, HSCs need to self-renew, proliferate and differentiate, a complex endeavor orchestrated by the genetic and epigenetic landscape, which in turn regulates gene expression. Despite the substantial progress that has been made in understanding these regulatory circuits, most if not all studies on human HSC regulation rely on an immunophenotypic definition of hematopoietic stem and progenitor populations. Consequently, transcriptomic or epigenetic data are derived from probably impure or heterogeneous populations. In line with this, phenotypic HSC definitions might miss cells, which are truly functional but do not fulfill surface marker-based selection criteria. To overcome the restriction of phenotypic HSC definition, we used a large dataset of γ-retroviral integration sites (γRV IS) from a gene therapy trial on 10 Wiskott-Aldrich-Syndrome (WAS) patients. To date, it is well established that γRV ISs can be exploited for tracking clonal dynamics and utilized as molecular tags for active enhancers and promoters – their preferred integration environment (see 1.5.1 and 1.5.3). Consequently, we hypothesized that 1) γRV ISs that are detected in the peripheral blood or BM of patients during long-term reconstitution have originated from true, functionally defined human HSCs and 2) that we can use γRV ISs to identify new regulators of HSCs and hematopoiesis as well as map the regulatory landscape that is influencing their spatiotemporal expression. Thereof, the following main aims were derived:

Aim 1: Utilize the WAS patient IS repertoire to identify and select protein-coding candidate genes with undescribed roles during hematopoiesis.

Aim 2: Establish a medium throughput lentiviral overexpression pool to examine the influence of the candidate genes on proliferation, self-renewal and differentiation of hematopoietic stem and progenitor cells.

Aim 3: Combine γRV ISs with publicly available datasets to create a genome-wide resource for active regulatory regions in functionally defined human long-term repopulating HSCs.

3 RESULTS

3.1 Identification of novel key hematopoietic regulators through γ-retroviral insertion sites

The basis of the present study is laid on a large collection of γ-retroviral (γRV) integration sites that were originally acquired (prior to this study) for biosafety reasons during a gene therapy trial including 10 patients with Wiskott-Aldrich-Syndrome (WAS) (Boztug et al., 2010; Braun et al., 2014). During this trial, CD34^{+} cells were mobilized using either Granulocyte-colony stimulating factor (G-CSF) alone or a combination of G-CSF and the CXCR4 inhibitor Plerixafor and extracted from the patients' blood using leukopheresis and magnetic cell separation (CliniMACS system). Next, a functional copy of the *WASP* gene was introduced into the CD34^{+} cells using γRV vectors, and finally CD34^{+} cells were re-infused as an autologous bone-marrow (BM) transplant. Throughout the follow-up of the study, whole blood, sorted blood-cell populations and BM samples were collected from the patients, genomic DNA was extracted, and the location of the ISs were determined using linear-amplified PCR (LAM-PCR) and high-throughput sequencing (HT-seq; Figure 12). Because γRVs stably integrate into the hosts' cell genome, each transduced cell is characterized by a unique integration site (IS). Accordingly, a bulk of cells carrying an identical IS must have originated from a common ancestor. Moreover, the HT-seq read counts for each IS can to some degree also convey information about the clone size. These parameters – the clonality of the sample (number of unique ISs) and the approx. clone sizes (% of total read counts) – can be used to characterize the patients' blood reconstitution after transplantation. These parameters are particularly crucial for the detection of neoplastic growth of transformed clones. However, in the present study we did not focus on single oncogenic integration events and their associated genes but instead on the complete picture of γRV ISs in all patients collectively, the so called integrome. In the past it has been shown that γRV ISs preferentially target active transcription start sites and active enhancer elements (Aker et al., 2006; Cattoglio et al., 2010; Deichmann et al., 2011; LaFave et al., 2014) and thus are not evenly spread out across the entire genome but almost always occur in clusters, so called common integration sites (CIS, also see Figure 12B). These CIS can in turn be used as indicators of genes that are active during the transduction of CD34$^+$ cells and hence might play a role during hematopoiesis.

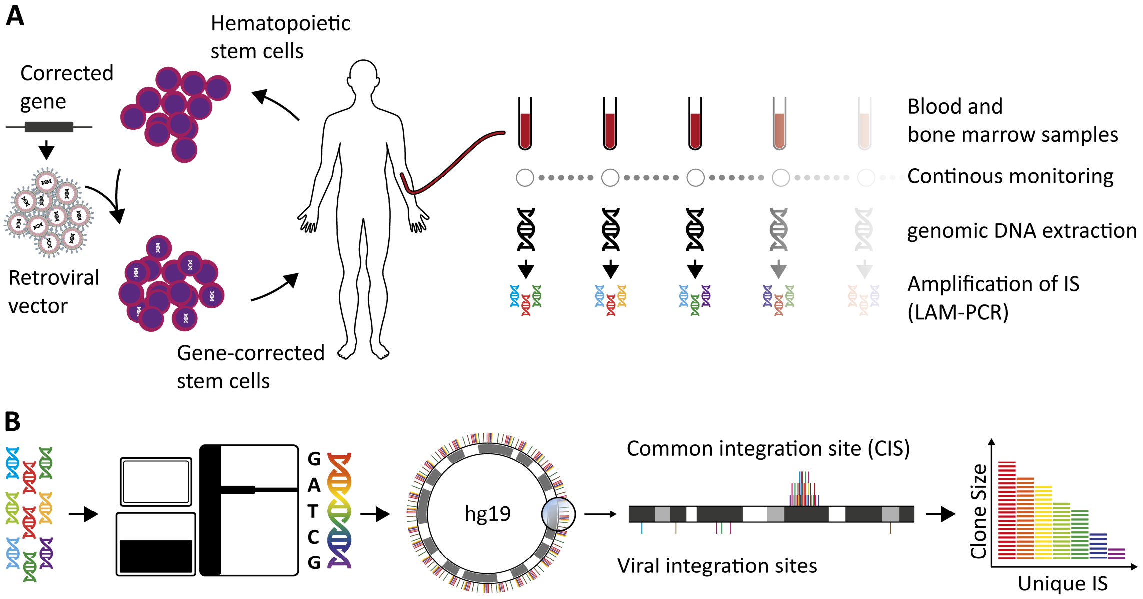


Figure 12 | **Genetic correction of diseased hematopoietic stem cells and subsequent monitoring of the patients in a clinical gene therapy trial.**
A | After mobilization of the patients' HSCs, CD34$^+$ cells were isolated and genetically engineered *ex vivo* using γRV vectors. Following gene correction, stem cells were transplanted back and patients were monitored for up to six years. Blood and BM samples were collected periodically and were either left unsorted or sorted for various cell populations, respectively. Genomic DNA was extracted and amplification of the viral integration site was performed (LAM-PCR). **B** | LAM PCR fragments were sequenced and adjacent genomic regions were mapped to the human hg19 reference genome for localization of the integration site at 1bp resolution. γRV integration into the genome occurs non-random, leading to the local accumulation of clusters of IS, called common integration sites (CIS). Additionally, read counts of unique IS provide an indirect measure for the clone size and appearance of the same clone in various cell populations at different time points.

3.1.1 Analysis of common integration sites and genes in their vicinity

To filter for genes that possess a higher likelihood to play a role during hematopoiesis we first developed criteria to weight the importance of CIS. Here, we hypothesized that a greater number of ISs close to a given transcription start site (TSS) or a smaller distance between ISs (higher density) would point towards regions that are either more active or regions that are a preferred integration target in a higher percentage of cells during the initial rounds of transduction or during engraftment of the cells. Thresholds for optimal prediction of CIS were established prior to this study and were set to a maximum distance between two ISs of the same CIS of 10kb and a maximum distance of the CIS boundary to the nearest TSS of 50kb (Figure 13A). After CIS prediction and assignment of TSS, three additional parameters were obtained – the degree (number of ISs around the TSS ±50kb), the CIS order (number of ISs within cluster) and the CIS dimension (genomic length of CIS, Figure 12B).

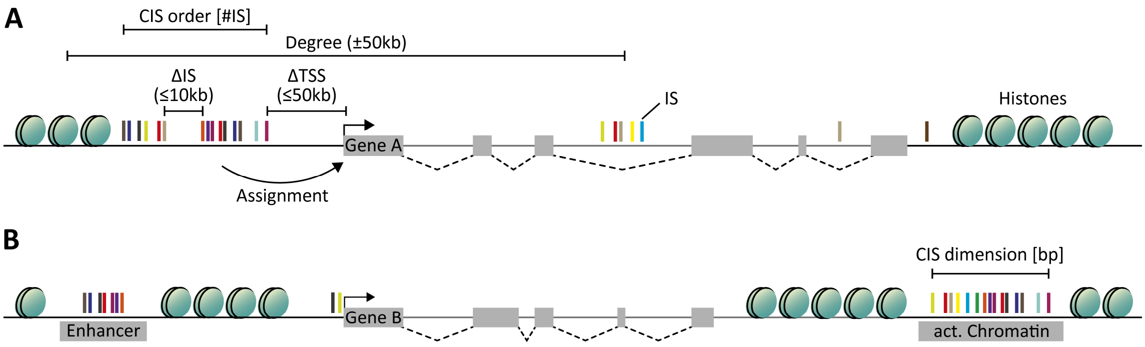


Figure 13 | **Schematic overview of genetic loci containing γRV ISs and terminology/parameters used for the characterization of clusters and statistic filtering for potential regulatory genes.**
A | ΔIS depicts the distance between two ISs and must not exceed 10kb in order to consider two neighboring IS to be present in the same cluster. ΔTSS depicts the distance between CIS and transcription start site (TSS) and must not exceed 50kb for the gene to be considered in the proximity of the cluster. The degree depicts the number of ISs in a window of ±50kb around a TSS, while the CIS order depicts the number of ISs in a given cluster. **B** | CIS dimension depicts the genomic size in bp a given cluster has.

3.1.2 Top 100 CIS are highly enriched for hematopoietic regulators

After allocation of CIS to their closest TSS, genes were ranked by CIS order. As described previously, the three top ranked genes (MECOM, LMO2 and HMGA2) are known proto-oncogenes and were previously linked to the development of malignancies and clonal expansion in γRV-driven gene therapy trials (Braun et al., 2014; Cavazzana-Calvo et al., 2010; Hacein-Bey-Abina et al., 2008; Hacein-Bey-Abina et al., 2003a; Hacein-Bey-Abina et al., 2003b; Ott et al., 2006). However, these genes as well as many other hematopoietic malignancy-related genes are also known to play essential roles during physiological hematopoiesis (Copley et al., 2013; Kataoka et al., 2011; Yamada et al., 1998). Strikingly, among the largest 100 CIS about 50% (49 genes) were reported to be linked to hematopoiesis, indicating that CIS can indeed be used as genetic marks to identify hematopoietic regulators (Figure 14A).

3.1.3 Selection of protein-coding genes as potential novel hematopoietic regulators

First, genes were ranked according to the order of their associated CIS and filtered for protein-coding genes. Next, only genes without a reported role or function in the hematopoietic system were selected. Finally, we performed extensive literature research to filter for genes with a mouse homolog, a maximum of two major isoforms as well as clonability, e.g. maximum length of mRNA of ~3.5kb. In total, we selected 17 genes of which three genes had two major isoforms (*Lair1, Slx4ip* and *Xbp1*; Figure 14B and Table 1).

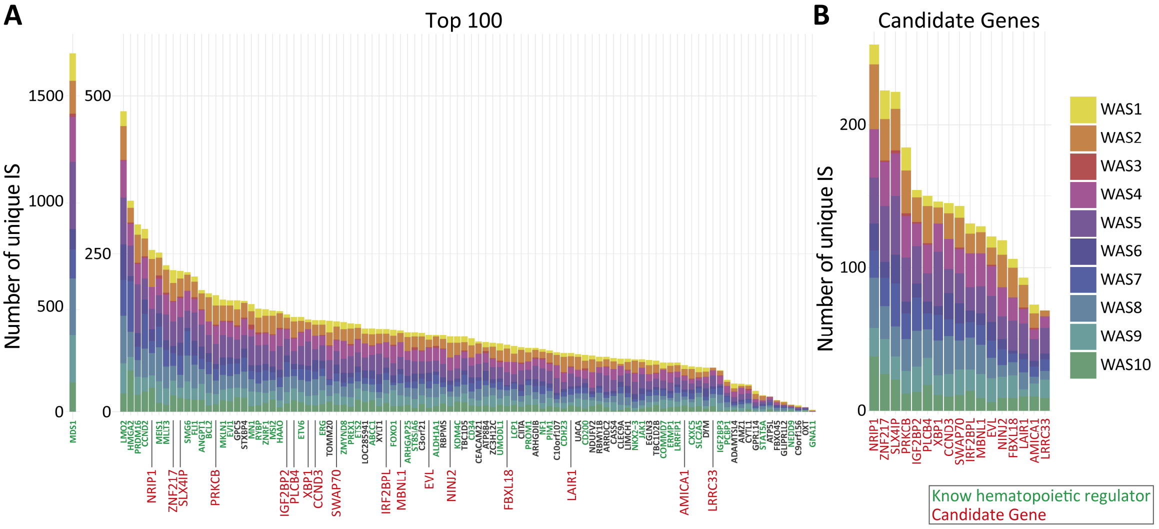


Figure 14 | **Top 100 largest CIS and their associated gene with indicated proportion of every patient.**
A | Top 100 largest clusters ranked for number of unique IS. MDS1 (MECOM) is scaled to an independent y-axis. Known hematopoietic regulators are indicated green, selected candidate genes are indicated red. **B** | Candidate genes are listed separately and ranked for number of unique IS.

Table 1 | **Overview of protein-coding candidate genes in this study**

Location	CIS Order	Degree	hGene	Full Name	Location of CIS	mGene	Size
chr20: 10485470	304	34	*SLX4IP*	SLX4 Interacting Protein	Mostly Intron 2, also Intron 1	*Slx4ip* (long) *Slx4ip* (short)	1,262 1,052
chr6: 41973175	268	122	*CCND3*	G1/S-Specific Cyclin D3	Mostly CIS>TSS or Intron 1	*Ccnd3*	899
chr3: 185475315	238	138	*IGF2BP2*	Insulin-Like Growth Factor 2 MRNA Binding Protein 2	Almost all Intron 1	*Igf2bp2*	1,799
chr20: 52251594	232	156	*ZNF217*	Zinc Finger Protein 217	CIS>TSS	*Znf217*	3,146
chr20: 9146426	226	121	*PLCB4*	Phospholipase C, Beta 4	Mostly CIS>TSS or Intron 1	*Plcb4*	3,548
chr22: 29208888	224	144	*XBP1*	X-box binding protein 1	CIS>TSS	*Xbp1* *Xbp1S*	824 1,136
chr11: 118104203	200	156	*AMICA1*	Adhesion Molecule, Interacts With CXADR Antigen 1	Mostly CIS>TSS or Intron 1	*Amica1*	1,160
chr11: 9743162	193	138	*SWAP70*	SWAP switching B-cell complex 70kDa subunit	Mostly Intron 2,3, also Intron 1	*Swap70*	1,778
chr21: 16611961	179	74	*NRIP1*	Nuclear Receptor Interacting Protein 1	CIS>TSS (far from TSS)	*Nrip1*	3,506
chr3: 151935060	173	92	*MBNL1*	Muscleblind-Like Splicing Regulator 1	CIS>TSS	*Mbnl1*	1,166
chr7: 5509323	169	103	*FBXL18*	F-Box And Leucine-Rich Repeat Protein 18	3'UTR>CIS	*Fbxl18*	2,177
chr12: 727092	164	107	*NINJ2*	Ninjurin 2	Almost all Intron 1, intronic lncRNA	*Ninj2*	452
chr14: 100536162	162	109	*EVL*	Enah/Vasp-like	Almost all Intron 1	*Evl*	1,265
chr16: 23892933	161	57	*PRKCB*	Protein Kinase C, Beta	Almost all Intron 1	*Prkcb*	2,042
chr19: 54887664	153	95	*LAIR1*	Leukocyte-Associated Immunoglobulin-Like Receptor 1	Mostly CIS>TSS	*Lair1* (long) *Lair1* (short)	812 482
chr14:77507619	126	94	*IRF2BPL*	Interferon regulatory factor 2 binding protein-like	CIS>TSS	*Irf2bpl*	2,345
chr3: 196353363	106	87	*LRRC33*	Leucine Rich Repeat Containing 33	CIS>TSS	*Lrrc33*	2,186

Location, Center of the CIS; **CIS Order**, Number of ISs in a CIS; **Degree**, Number of insertion sites within 10 kb in each direction of the transcription start-site of individual candidates; **hGene**, Name of the human Gene, **Location of CIS**, Describes where most of the ISs site are located in relation to the genes TSS; **CIS>TSS**, ISs are located upstream of the TSS; **3'UTR>CIS**, ISs are located downstream of the 3'UTR; **mGene**, Name of the corresponding mouse gene. "Long" and "short" indicated different splice variants. Xpb1S represents a splice variant of Xbp1; **Size**, Size of the cDNA in bp.

3.1.4 Most candidate genes are expressed in the hematopoietic system

After selecting the candidate genes, we investigated the expression pattern of our candidate genes throughout the human and mouse hematopoietic system by screening publicly available RNA-seq expression data (Cabezas-Wallscheid et al., 2014; Corces et al., 2016). Importantly, almost all genes showed a slight tendency to be higher expressed in stem and progenitor cells compared to more mature blood cells. In humans, *AMICA1* was the only gene that was not detected across hematopoietic cell populations, which was in line with the very low expression detected in mice. In contrast, *Znf217* was not detected in mice despite its relatively high expression in humans. In summary, the majority of the genes showed high to medium expression levels in stem and progenitor cells and was similarly expressed across species (Figure 15).

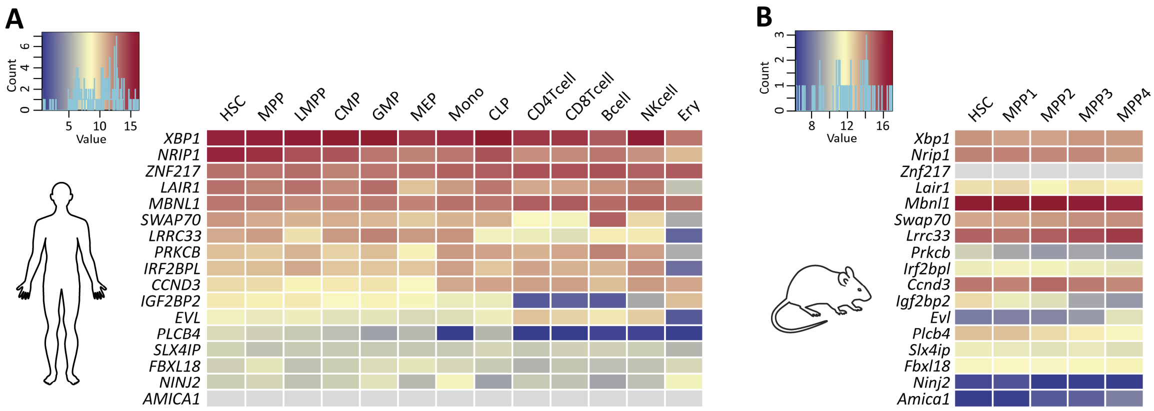


Figure 15 | **Expression pattern of candidate genes.**
A | Relative expression in the human hematopoietic system. Genes were ranked according to their expression in HSCs. **B** | Relative expression in mouse hematopoietic stem and progenitor cells. Genes are ranked according to order in Figure 15A.

3.2 Establishing a pooled lentiviral based screening platform

To study gene functions in a specific cellular context, several approaches are applicable. Classically, overexpression and knockdown *in vivo* and *in vitro* are the most commonly used techniques. In this study, we aimed to investigate the gene function in murine cells by genetically modifying murine hematopoietic stem and progenitor cells through lentiviral transduction. Because of the high number of genes that were intended to be investigated, we aimed to design a pooled lentiviral-based overexpression approach that will allow us to study the phenotype associated with the candidate genes in a parallel fashion. This step helps to gain functional data for many genes simultaneously and to re-evaluate the ranking for further downstream investigations. Important considerations for such a screen include the principal study concept, basic requirements for the library design as well as essential initial tests to validate applicability. The following subchapters of 3.2 address the establishment and testing of the screen, while the consecutive subchapters of 3.3 refer to the results generated *in vitro* and *in vivo* with this library. More information on the functional principle of the library-based screen are provided with the subchapters of 3.3.

3.2.1 Stable overexpression of candidate genes with lentiviral vectors

The pooled approach is based on lentiviral overexpression constructs that consist of an HT-seq compatible 18nt barcode (BC) and an inducible promoter, which initiates the transcription of the gene of interest (GOI) and GFP as a marker protein (Figure 16A). After synthesis of the candidate cDNAs and cloning into the target vector (see 5.2.1.6 and 5.2.1.7), we produced GFP only (control) and single candidate gene virus supernatants to test GFP stability and mRNA expression *in vitro*. To this end, we transduced human HL60 cells that express the reverse tetracycline-controlled transactivator (rtTA) and split the cells after a short recovery period into Doxycycline containing (+DOX) or control wells (-DOX). Next, +DOX cells were sorted for GFP to increase the purity, using fluorescence-activated cell sorting (FACS). While GFP control cells and most other constructs showed stable GFP expression over a period of at least one month (data not shown), some constructs showed declining GFP levels over time, indicating a greater survival fitness of untransduced cells. Next, mRNA levels of candidate genes were measured in GFP enriched candidate- or control GFP overexpressing cells, using quantitative PCR (qPCR). Primers specific to the murine codon optimized sequence were chosen to exclude unwanted amplification of the human endogenous mRNAs. Unfortunately, cells expressing *Fbxl18*, *Mbnl1*, *Plcb4*, *Xpb1S* and *Znf217* could not be cultured for an extended period of time without loss of GFP positivity, possibly due to greater fitness of non-transduced cells compared to transduced cells. Consequently, these mRNA expression levels could not be detected. All other constructs however showed stable GFP and detectable mRNA levels, which varied between approx. 10% and 100% of the levels of *GAPDH* (Figure 16B). Moreover, we exemplarily measured the protein levels of Igf2bp2 using western blot, which indicated sufficient translation of the mRNA into protein (Figure 16C). In summary, the lentiviral transduction and expression worked on both, mRNA and protein level in a doxycycline-dependent manner. Some constructs appeared to compromise the fitness of the cells; however, at this point it was not clear whether this resembles a cell type-specific biological effect or a general toxicity of the gene product itself.

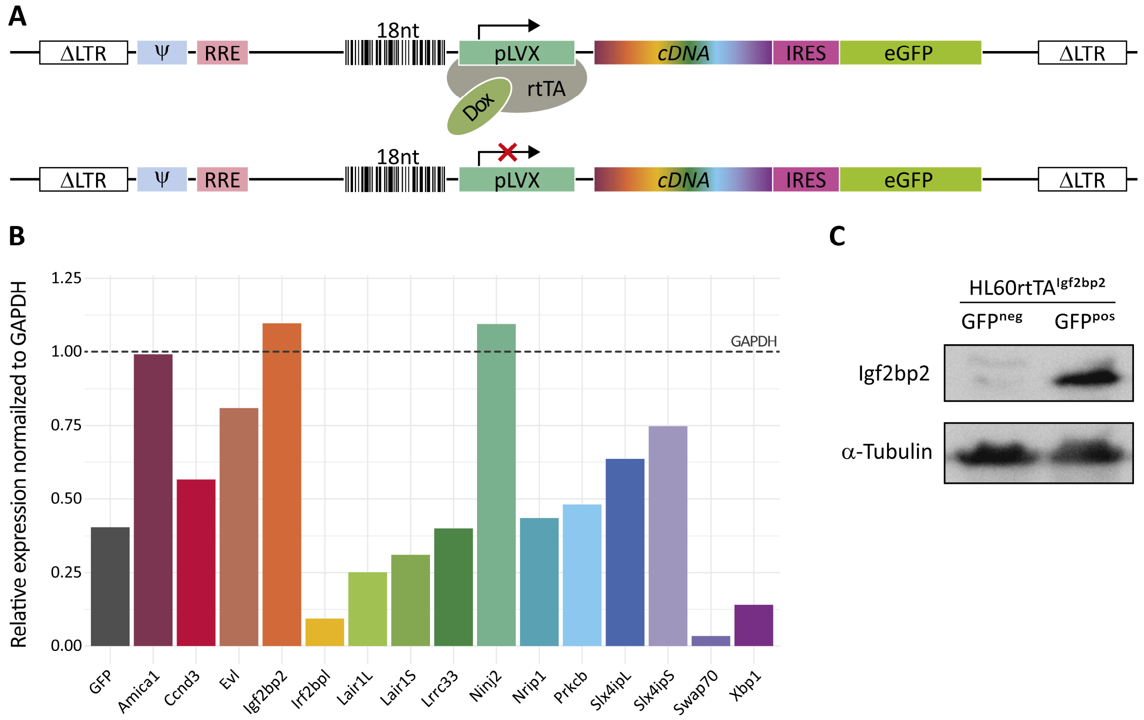


Figure 16 | **Lentiviral-mediated candidate gene overexpression.**
A | Schematic representation of the essential components of the lentiviral overexpression construct in active (with DOX, rtTA bound) and inactive (no DOX, rtTA unbound) conformation. **ΔLTR**, long terminal repeat with deletions; **Ψ**; Retroviral Psi packaging element ;**RRE**, Rev-responsive element; **18nt**, molecular BC 18nt sequence that is unique for every gene or construct, respectively (for more detail see section 3.2.2); **pLVX**, Tet-inducible promoter; **IRES**; internal ribosomal entry site; **eGFP**, enhanced green fluorescent protein; **B** | Relative mRNA expression of candidate gene normalized to *GAPDH*. **C** | Igf2bp2 protein expression in *Igf2bp2* overexpressing HL60^{rtTA} cells, segregated in GFP negative and GFP positive cells. Housekeeping gene α-tubulin indicates equal loading.

3.2.2 Custom molecular barcoding and high-throughput sequencing cassette

To trace the differentiation and proliferation behavior of transduced cells *in vitro* and *in vivo*, we equipped the lentiviral OE constructs with a ~250bp barcode cassette, harboring primer binding sites for a nested PCR in order to amplify the cassette including a 18nt BC from genomic DNA (gDNA) as well as an HT-seq primer binding site. The cassette was adapted from the well-established Cellecta shRNA library, which contains more than 27,000 unique 18nt BCs. In total, we randomly selected 96 unique BCs from this pool in order to provide two BCs for every GOI and 10 BCs for the controls. The remainder of the 96 BCs was used for related projects on miRNA overexpression (Elias Eckert, data contained in Wünsche et al. (2018)) or tests for amplification efficiency and accuracy (3.2.3).

Upon detection of BCs in transduced cells, the cassette is amplified by nested PCR with the first PCR cycle amplifying the 18nt BC and the P7 sequence (important for hybridizing with the chip during sequencing), while eliminating the endogenous P5 sequence. Because of the anticipated large number of sequencing samples, we re-designed the reverse primer for the second PCR to harbor one of 96 possible 8nt index sequences. This allows for multiplexing the samples and thus for simultaneous sequencing of up to 96 samples per Illumina HiSeq lane. Moreover, the reverse primer during the second PCR also introduces the binding site for the Illumina index primer as well as a new P5 sequence (sequence identical to original P5 sequence, important for hybridizing with the chip during sequencing). This way, the generated PCR products are ready for sequencing and do not require adapter ligation or library preparation. Finally, the sequencing run consists of two steps, the first read provides the barcode sequence and thus allows for allocation of the GOI followed by the second read, which provides the index and thus allows for the allocation of the sample (Figure 17).

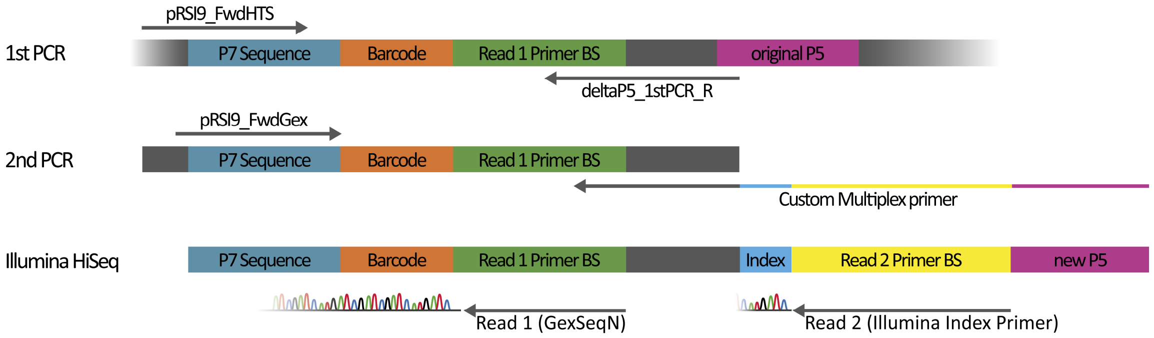


Figure 17 | **Detailed representation of the HT-seq barcode cassette with multiplexing PCR step.** During the 1st PCR the barcode cassette is amplified from gDNA, eliminating the endogenous P5 sequence. In a 2nd PCR the barcode is further amplified and reactions can be by one out of 96 8nt indices while also introducing a new P5 sequence. Finally, the PCR products are sequenced in a 50bp single read Illumnia HiSeq 2000/2500 run in which the 1st read identifies the 18nt barcode (overexpressed gene) and the 2nd read the 8nt index (sample)

3.2.3 Systematic one-by-one amplification of indices reveals very low index-bleeding

Due to the relatively short sequence of the index (8nt), it is conceivable that technically unavoidable sequencing errors could cause the misallocation of samples. To minimize this effect, we used the pre-designed NuGene index library, which was designed to vary between indices in at least two positions, ensuring accurate indexing. To minimize cross-contamination during multiplex primer (Figure 17) synthesis, primers were ordered individually and timely spaced. To assure that plasmid stocks were not contaminated with other BC containing plasmids, we next transformed bacteria with plasmids and picked single colonies for barcode amplification during 1st PCR. The 2nd indexing PCR was inoculated using 5µL of PCR product from the 1st PCR. After additional quality checks, PCR products were pooled and sequenced using the standard HiSeq 2000 50bp single read protocol (more detail is provided with Figure 18 and 5.2.1.8).

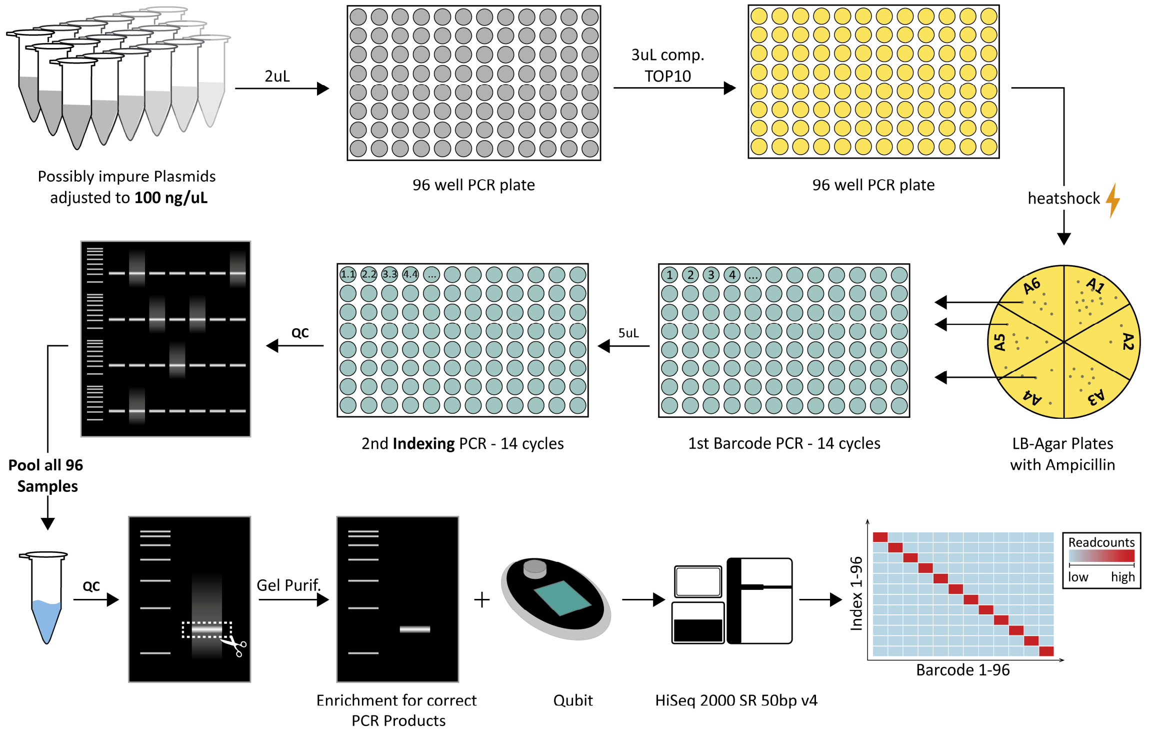


Figure 18 | **Matrix experiment to evaluate any potential index bleeding (Legend continued on next page).**
After adjusting the concentration of all 96 plasmids, plasmids were aliquoted into a 96 well plate and supplemented with competent TOP10 bacteria. After heat shock and incubation, reactions were applied to LB-Agar plates to grow single clone colonies. Next, 1st PCR was conducted as a colony PCR amplifying the barcode directly from bacterial plasmids to ensure maximum purity of the barcode, followed by the 2nd PCR, in which every barcode was amplified with one unique index primer (also see Figure 17). Before

pooling, every reaction was loaded on an agarose gel to ensure successful amplification. After pooling of all reactions, correctly sized PCR products were enriched using gel purification, followed by concentration adjustment, sequencing and data analysis.

After sequencing, raw data were de-multiplexed and reads were trimmed and counted (also see 5.2.1.8). All barcodes showed very low cross-contamination or index-bleeding with a mean of 734 wrongly allocated BCs per sample at an average of 1.24 x 10^6 reads per sample (0.059%), which equals a mere 7.7 incorrectly allocated reads per BC per sample (0.0006%, Figure 19). The only exception of relatively high cross-contamination levels (approx. 20,000 wrongly allocated reads per sample) were found in samples D4 and D10. However, due to the indistinguishable nature of these indices compared to other indices, these contaminations might have occurred during the preparation rather than the sequencing steps. Taken together, this test experiment proofs the feasibility to pool up 96 samples on one HiSeq lane and thus provides the basis for all following *in vitro* and *in vivo* screening experiments.

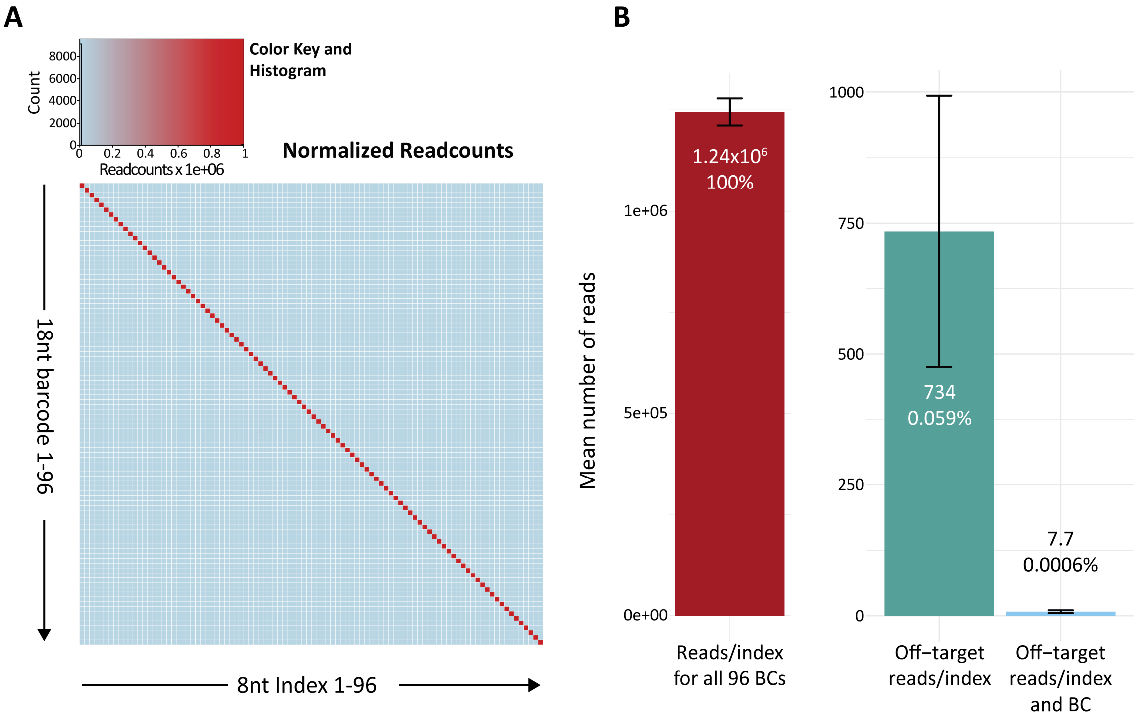


Figure 19 | **Evaluation of index bleeding.**
A | Heatmap of read counts normalized to 1 x 10^6 per barcode with color being proportional to number of reads. **B** | Average number of reads per barcode before normalization and quantification of off-targets per index and per barcode. **BC**, barcode; Data jointly produced with Elias S. P. Eckert.

3.2.4 Titers can vary between constructs and productions and do not correlate with cDNA length

Generally, it is believed that bigger plasmids are less efficiently packed during virus production or transcribed less abundantly and hence result in lower virus titers (Kumar et al., 2001). Because we aimed to produce the virus in a single reaction using the pooled plasmid library that contains constructs of varying sizes, we assessed the titer for all constructs for comparison. Virus supernatants were produced simultaneously in 6-well plates to avoid batch effects and titrated on HL60^{rtTA} cells. Interestingly, two independent virus productions showed varying results. The 1st production showed a relatively tight range of titers varying between approx. 7,000 to 37,000 infectious units (IU) per mL, while the 2nd production varied substantially more (100 - 40,000 UI/mL; Figure 20A). Although the slope of the linear regression of the 1st production indicated that smaller constructs yielded a slightly higher titer, the overall fit was very poor and hence suggested that the construct size cannot be the only titer-influencing factor (Figure 20B). In summary, all constructs gave rise to functional viruses and titers indicated by measurable GFP expression, which were sufficiently comparable for following *in vitro* and *in vivo* experiments with the pooled lentiviral OE library.

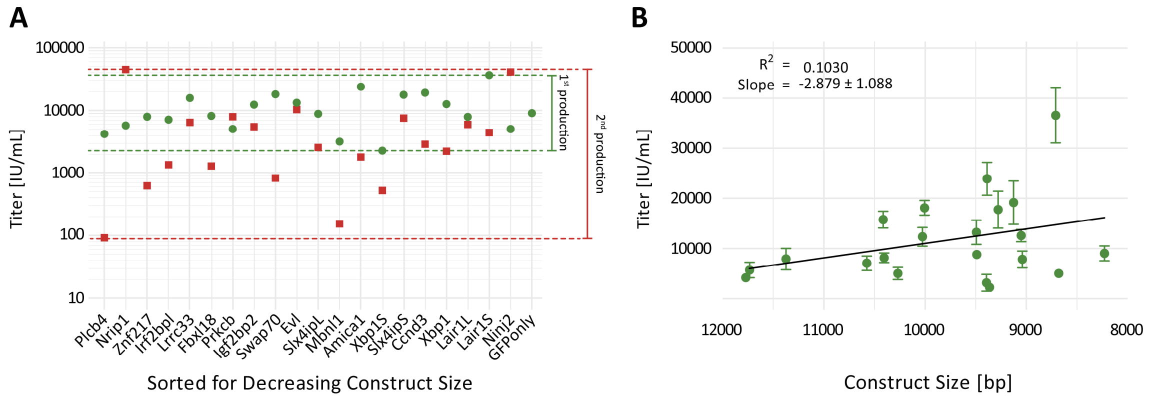


Figure 20 | **Relation between construct size and virus titer.**
A | Titer for every construct after two independent virus productions. Productions and their outer boundaries of titer variations are indicated by colors and dashed lines. Constructs on x-axis are sorted for decreasing construct sizes with the biggest construct on the left and the smallest on the right hand site. **B** | Titers of 1st production plotted against actual size in bp. Fit and slope of the linear regression are indicated. Note that the x-axis is reversed. **IU**, infectious units.

3.2.5 Stable GFP expression and barcode representation over time *in vitro*

Next, we investigated the stability of GFP and BC representation *in vitro*. To this end, we used a pool of 25 constructs equal in size (GFP only) but with unique BCs. Plasmids were pooled at equal amounts with the exception of two BCs, which were spiked in at a 5-fold and 2-fold overrepresentation. After virus production, HL60^{rtTA} cells were transduced with two different multiplicities of infection (MOI, 0.13 and 1.3) and GFP expression was measured periodically for 41 days. Percentage of GFP positive cells was stable across the entire time independent of the MOI (Figure 21A). Additionally, genomic DNA (gDNA) was extracted from cell aliquots at day 1, 3, 10, 21 and 28 and sequenced after BC amplification. This revealed stable BC representation until day 28. In summary, constructs only overexpressing GFP but no cDNA did not encounter positive or negative selection *in vitro* and therefore are suitable for the pooled LV OE library.

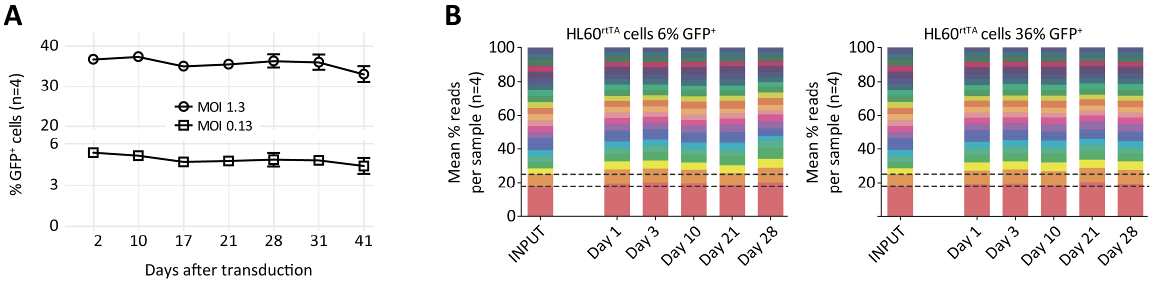


Figure 21 | **Assessment of GFP and BC stability over time *in vitro*.**
A | Cells were transduced with a MOI of 0.13 or 1.3, respectively, and GFP was measured at indicated days. **B** | BC representation at indicated time points in comparison to the initial plasmid mix (INPUT). Dashed lines indicate the input boundaries for the 5-fold and 2-fold overrepresented BCs.

3.3 Characterization of candidate hematopoietic regulators

After establishing the platform required for the simultaneous characterization of our candidate genes, we proceeded with *in vitro* serial colony forming unit (CFU) and cell trace assays as well as *in vivo* transplantation experiments. At this stage of the study, we speculated that cells transduced with our candidate genes might have a competitive advantage over untransduced or GFP only transduced cells. This hypothesis was based on the widely postulated concept that large clusters of ISs near a given gene occur due γRV-mediated activation of this gene, which in turn leads to the clonal expansion. Due to the expected differences in proliferation changes upon transduction with either only GFP or GOI expressing constructs, two independent lentiviral OE pools were produced:

1) A “GFP pool”, consisting of 10 plasmids only expressing GFP but each harboring a unique BC.
2) A “GOI pool”, consisting of 20 protein-coding genes each represented by two unique BCs, resulting in a total of 40 BCs.

3.3.1 Results of serial CFU re-plating assays revealed changes in BC representation over time but were limited by poor library presentation

In order to assess the effect of our candidate genes on proliferation, differentiation and self-renewal, we sorted Lineage negative (Lin$^-$), Sca-1$^+$, c-Kit$^+$, CD48$^-$ and CD150$^+$ (LSK-SLAM) cells from Rosa26 rtTA mouse bone marrow (BM). After 48h recovery time, cells were split and either transduced with the GFP pool or the GOI pool. Next, DOX was added to the medium in order to induce GFP and/or GOI expression. 72h later cells were sorted for GFP followed by seeding into semi-solid medium. For all assays, the transduction efficiency was kept below 25%, ensuring a maximum of one vector copy per cell. In total, 1,200 cells for both, the control as well as the GOI OE group were plated (120-fold / 30-fold barcode representation). Cells were re-sorted for GFP after seven days and re-plated again after 15 and 28 days. An aliquot of cells was kept for all re-plating time points before and after DOX administration and finally gDNA was extracted, barcodes were amplified and sequenced (Figure 22A).

Besides flow-cytometric analysis of GFP levels at every sort or re-plating time point, GFP positivity was also assessed by fluorescence microscopy and revealed bright and homogenous GFP levels throughout colonies (Figure 22B). The GOI pool BC analysis revealed drastic changes in relative percentages with the majority of BCs declining over time in favor of very few BCs, which showed a corresponding increase (*Ccnd3*, *Lair1L*, *Lair1S* and *Xbp1*; Figure 22C). Expectedly, GFP pool BCs appeared much more stable over time except for one BC, which also declined over time. However, this BC showed a much lower initial representation in the pool compared to other BCs, indicating a technical rather than biological effect (Figure 22C, yellow BC, indicated by arrow). In three following CFU assays, similar effects were observed with relatively stable GFP pool BCs but drastically changing BC proportions in the GOI pool cohort. However, colony numbers and transduction efficiencies for all following CFU assays suggested an underrepresentation of the GOI pool library, increasing the chance that BC changes occur for stochastic rather than biological reasons. Interestingly however, on average GOI pool transduced cells did not show increased colony numbers, total cell counts or replating efficiencies compared to GFP pool transduced cells, indicating that the overexpression of candidate genes did not result in drastic changes in proliferation or self-renewal capacity as initially suspected. In summary, the approach appeared to be feasible, however, showed limitations due to transduction efficiency and GFP$^+$ colony numbers, rendering biological conclusions for individual candidates difficult. Nevertheless, the CFU experiments indicated thus far that our candidate genes do not massively alter the proliferation or self-renewal capacity of GOI pool transduced cells.

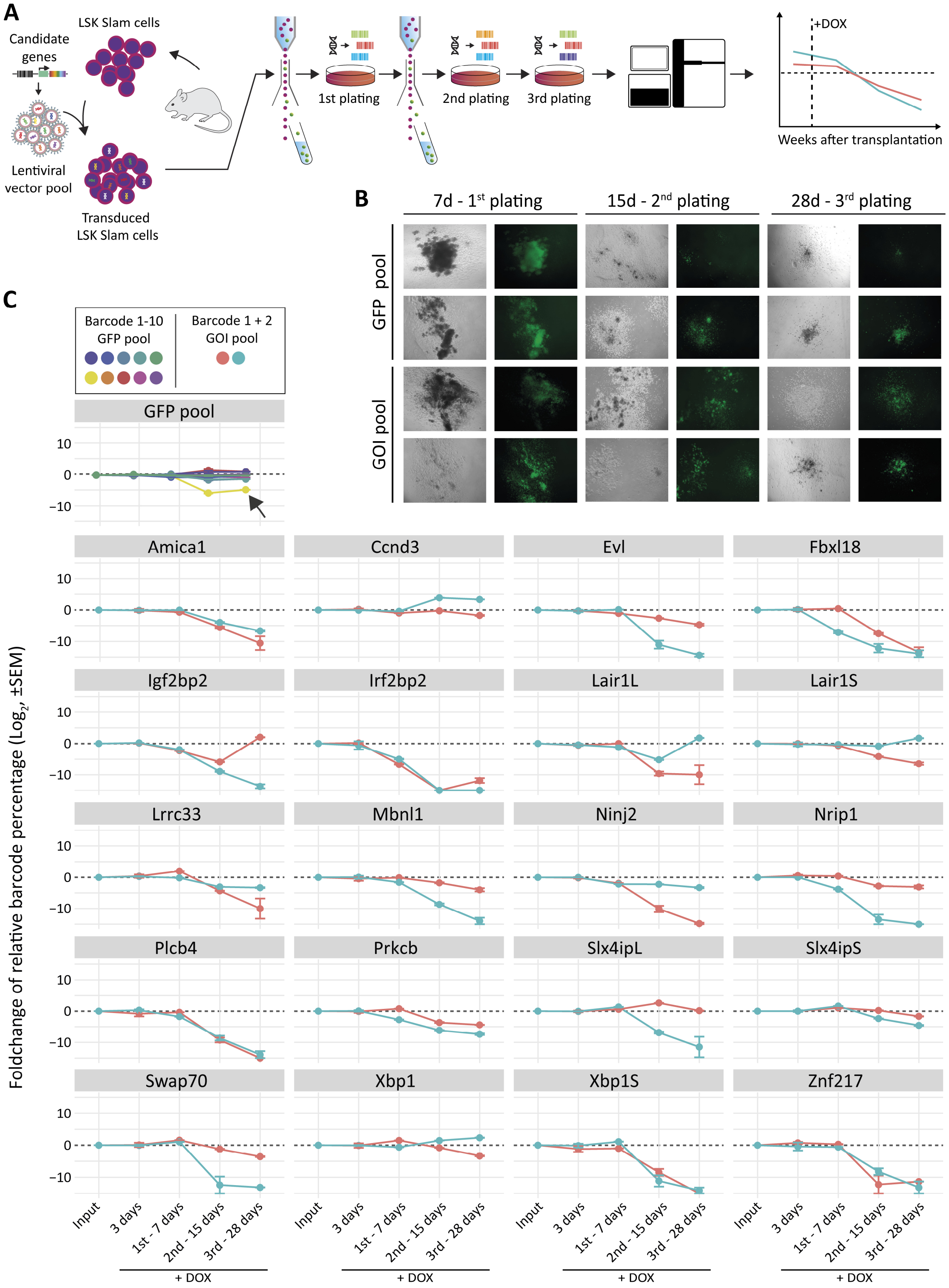


Figure 22 | **Serial replating of transduced LSK-SLAM cells in semisolid medium (Legend continued on next page).**

A | Experimental workflow. LSK-SLAM cells were isolated from Rosa26 rtTA mouse BM, transduced with the lentiviral pool and sorted for GFP 72h after DOX induction. Cells were re-sorted for GFP after 1st re-plating and subsequently plated two more times. gDNA was extracted from spare cells at every time point in order to allow for BC amplification followed by HT-seq. **B** | Representative fluorescent microscope pictures of colonies at 40x magnification with bright field (left columns) and GFP channel pictures (right columns). **C** | Log$_2$ of BC counts normalized to initial LSK-SLAM representation before DOX administration (±SEM). The BC with very low initial representation (yellow points) in GFP-control group is indicated by an arrow.

3.3.2 Cell trace experiment indicated varying proliferative potential between cells transduced with different lentiviral OE constructs

Due to the limitations encountered in the CFU assay, we developed another assay that is scalable to larger numbers of cells. To this end, approx. 4 x 10^5 LSK cells from donor Rosa26 rtTA mice were split into $^1/_3$ GFP pool and $^2/_3$ GOI pool cells and transduced as described above (3.3.1). Next, transgene expression was initiated using DOX and cells were stained with CellTrace™ violet, a FACS compatible membrane bound dye which signal intensity is approx. halved with every cell division. Stained cells were cultured for 3 or 5 days, respectively, and finally sorted into fast (weakest signal intensity), intermediate fast, intermediate slow and slow (highest signal intensity) cycling cells. Lastly, gDNA was extracted and BCs were amplified and sequenced (Figure 24A and B). In general, most of the genes appeared to have only little effect on the proliferation of LSK cells *in vitro*, which was reproducible in both independent experiments. The only two genes which showed a mild effect on proliferation were *Irf2bp2*, which was enriched in the fast cycling fraction and the long splicing form of *Lair1* (*Lair1L*), which was enriched in the slow cycling fraction. Taken together, the two experiments showed comparable results, indicating that the approach is feasible and less prone to stochastic effects due to the higher initial number of cells. However, the relatively short culture time of 3 to 5 days requires profound changes in proliferation to become apparent in this assay. Our candidate genes however, appeared to only mildly affect proliferation *in vitro*, which is partially in line with the observations made in the CFU assay.

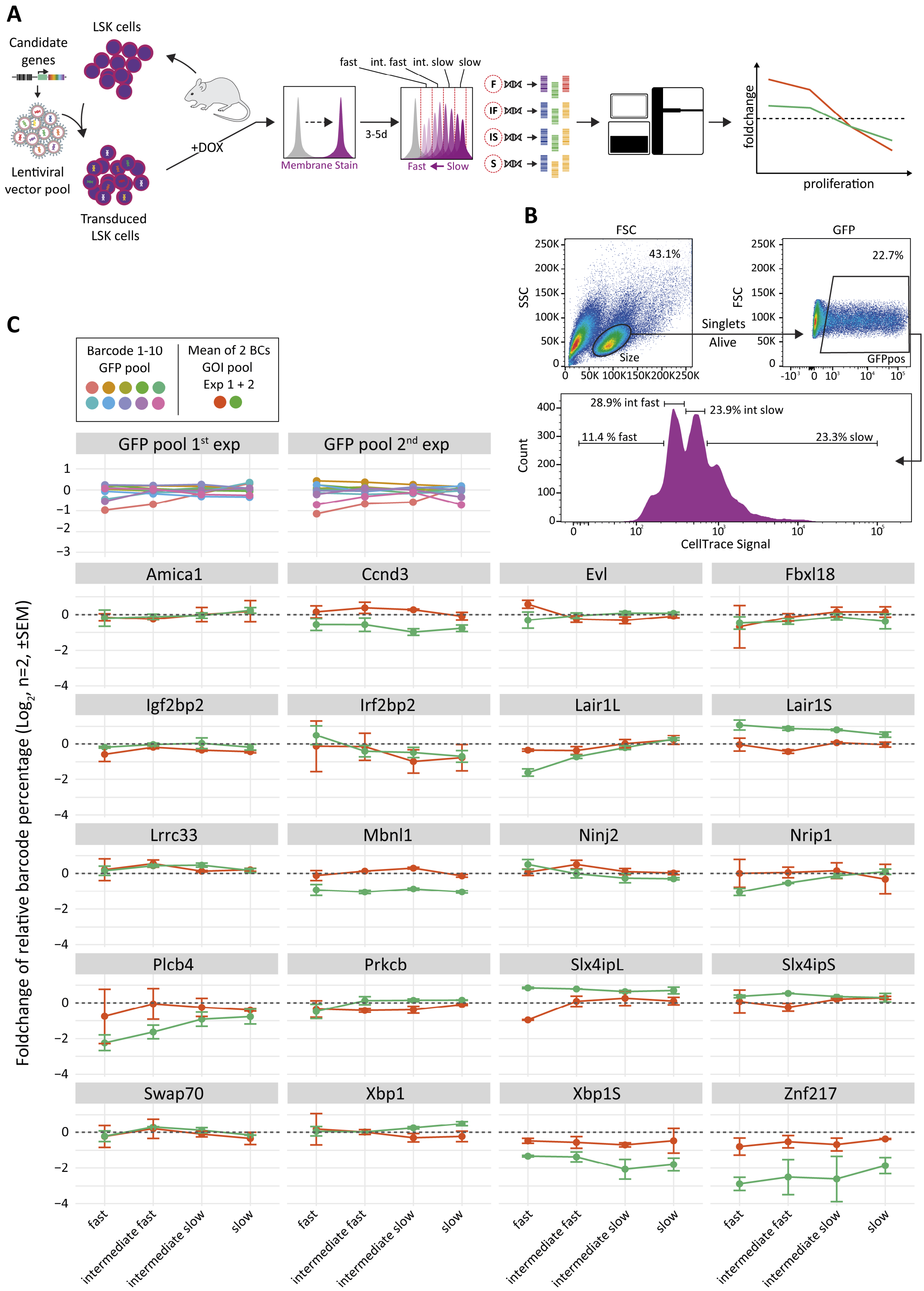


Figure 23 | **CellTrace assay for the detection of changes in proliferation upon GOI OE (Legend continued on next page).**

A | Experimental workflow. LSK cells were isolated from Rosa26 rtTA mouse BM and transduced with either the GFP or GOI pool. Transgene expression was induced using DOX, cells were stained with CellTrace^{TM} and cultured for 3 or 5 days, respectively, before sorting fast, intermediate fast, intermediate slow and slow cycling cells. Finally, gDNA was extracted from sorted fractions, followed by BC amplification and HT-seq. **B** | Exemplary FACS plots showing the gating strategy and generational peaks. **C** | Log_2 of the mean counts of two BCs per gene normalized to initial LSK representation (grey dashed line) before DOX administration (±SEM). Results of two independent experiments are plotted in one plot (orange and green dots and lines), while GFP control experiments are plotted individually. **F,** fast; **IF**, intermediate fast; **IS**, intermediate slow; **S**, slow; **FSC**, forward scatter; **SSC**, side scatter.

3.3.3 *In vivo* lentiviral overexpression screen reveals potential hematopoietic regulators but is again limited by engraftment and library representation

After gathering initial results through *in vitro* assays we continued characterizing the candidate genes through mouse transplantation experiments. To this end, $\mathrm{LSK}^{\mathrm{Rosa26\ rtTA}}$ cells were harvested and transduced as described before (3.3.1). After transduction, cells were kept in culture for 48h-72h and finally transplanted into lethally irradiated recipient mice. For all experiments, peripheral blood (PB) was harvested for the first time four weeks after transplantation, followed by addition of DOX to the drinking water to induce expression of GFP or GFP and GOI, respectively. Next, PB was harvested every four weeks until week 20. At all bleeding time points, barcodes were amplified from whole blood samples as well as sorted fractions of myeloid cells, B cells or T cells (myeloid cells: Ly-6G$^+$ CD11b$^+$, T cells: CD3$^+$, B cells: CD45R$^+$; lineage data not shown). Collectively, out of 156 transplanted mice, 11 mice from the GFP pool group (approx. 25-28% GFP$^+$ cells and 37-51% engraftment) and 11 mice from the GOI pool group (approx. 20% GFP$^+$ cells and 75% engraftment) exhibited sufficient GFP positive cells and engraftment for BC analysis (for an overview of conducted experiments see Table 2). All mice revealed significant proportional changes over time in both groups, with control mice appearing more stable, compared to mice transduced with the GOI pool (Figure 24B and Figure 25). Comparable to the results from the CFU assay, many BCs showed a decline over time with *Nrip1* declining most significantly. Importantly however, *Nrip1* also showed the lowest initial BC representation, indicating a comparably low titer and consequently an initial lower number of transduced cells. Interestingly, all genes that showed a significant change in BC representation (Friedman-Test) showed declining BCs, while no gene showed a consistent and significant increase in BC representation. Unexpectedly, control mice transplanted with GFP pool transduced $\mathrm{LSK}^{\mathrm{Rosa26\ rtTA}}$ cells also showed significant changes in relative BC proportion (both, increasing and decreasing BC representation; Figure 25). However, here every BC was tracked on its own, while the two BCs for every gene in the GOI pool were averaged first before statistical evaluation, making a direct comparison difficult. To account for the differences in the number of BCs, we performed an independent statistical comparison between GFP and GOI pool, which is explained in more detail below (see 3.3.4).

Although the interpretation of the *in vivo* data is again challenging due to a relatively low library representation and possibly small number of engrafted and transduced clones, we strikingly never observed clonal expansion or signs of leukemia or neoplastic growth in any of the 156 transplanted mice. This again shows that the effect on proliferation or self-renewal of transduced cells is much smaller than initially anticipated, questioning whether the clusters of γRV ISs that were originally used for the selection of candidate genes appeared due to enhanced expansion or self-renewal or rather marked large regulatory sites in the genome of transduced cells.

Table 2 | **Overview of performed transplantation experiments**

Exp.	**# LSK cells per mouse**	**Mean % GFP GFPp | GOIp**	**# Mice GFPp | GOIp**	**# Mice survived (4w) GFPp | GOIp**	**Mean % engraftment (4w) GFPp | GOIp**
TX V01	5,000	0% | 0%	6 | 12	6 | 12	69% | 46%
TX V02	15,000	45% | 20%	6 | 12	0 | 12	0% | 75%
TX V03	15,000	25% | 17%	6 | 12	5 | 10	51% | 53%
TX V04	15,000	28% | -	6 | 12	6 | 2	37% | 52%
TX V05	5,000	0% | 22%	6 | 12	6 | 8	0% | 0%
TX V06	15,000	- | -	6 | 12	5 | 9	0-35%
TX V07	40,000	0-1%	24 (Mix)	9	60-70%
TX V08	20,000	0-1%	24 (Mix)	18	70-80%

Exp., name of experiment; **#**, number of; **GFPp**, mice transduced with GFP pool; **GOIp**, mice transduced with GOI pool; **Mix**, mice transduced with lentiviral pool containing both GFP control and GOI constructs; **4w**, four weeks after transplantation.

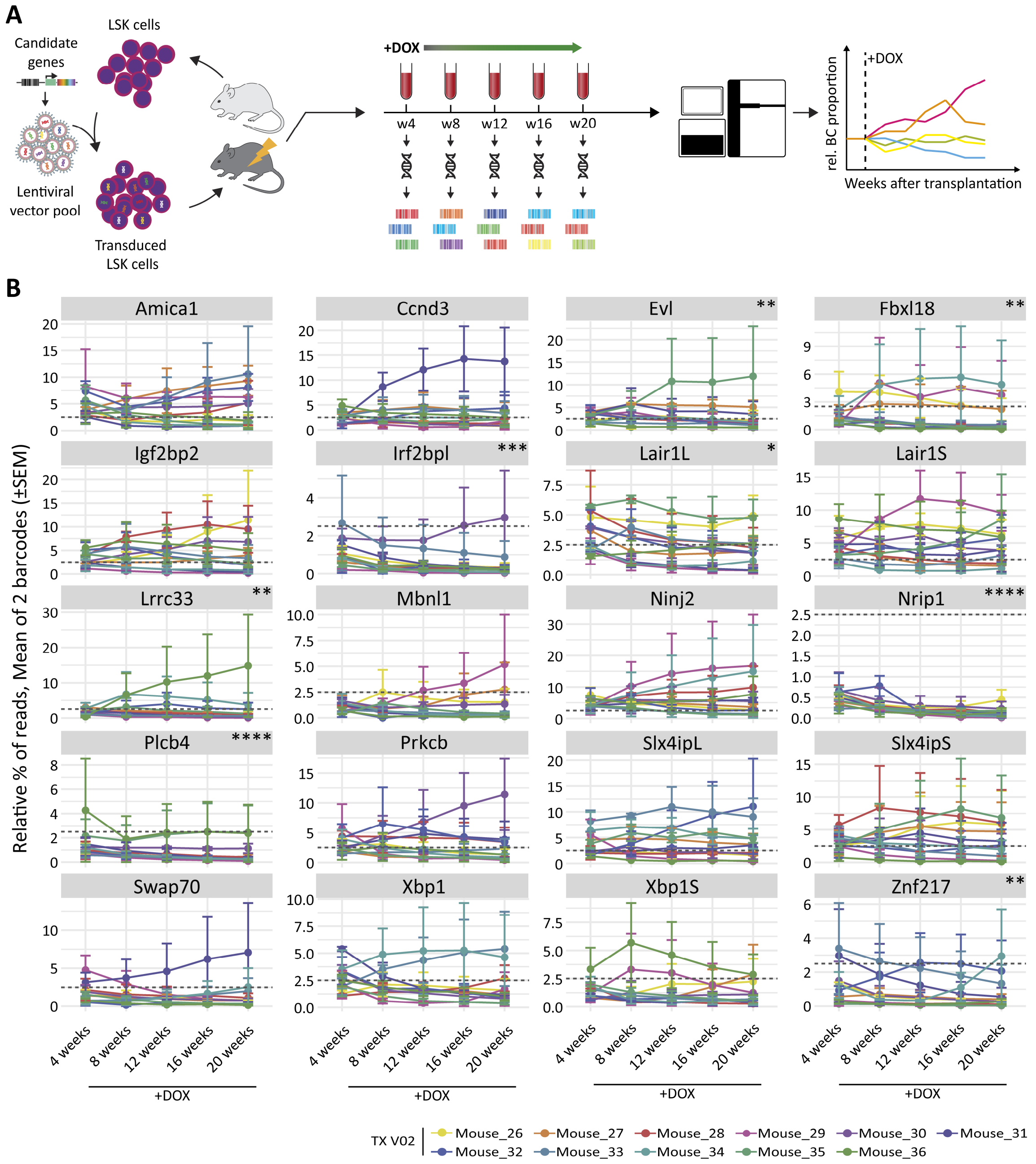


Figure 24 | **Relative proportions of read counts for each gene (mean of two BCs) per mouse over time.** **A** | Experimental workflow. LSK cells were isolated from Rosa26 rtTA mouse BM and transduced with either the GFP or GOI pool. Transgene expression was induced using DOX 4 weeks after transplantation and mice were bled every four weeks. After lysis of erythrocytes, gDNA was extracted from peripheral blood samples fractions, followed by BC amplification and HT-seq. **B** | Relative % of BC counts over time (mean of 2 BCs that depict the same gene). Data was not normalized to 4 weeks after transplantation to indicate initial BC representation. **TX V02**, Transplantation cohort No. 2 (Table 2). Gray dashed lines indicate the theoretical mean proportion for each barcode (2.5%). Friedman Test: * = p ≤ .05, ** = p ≤ .01, *** = p ≤ .001, **** = p ≤ .0001, ns = not significant.

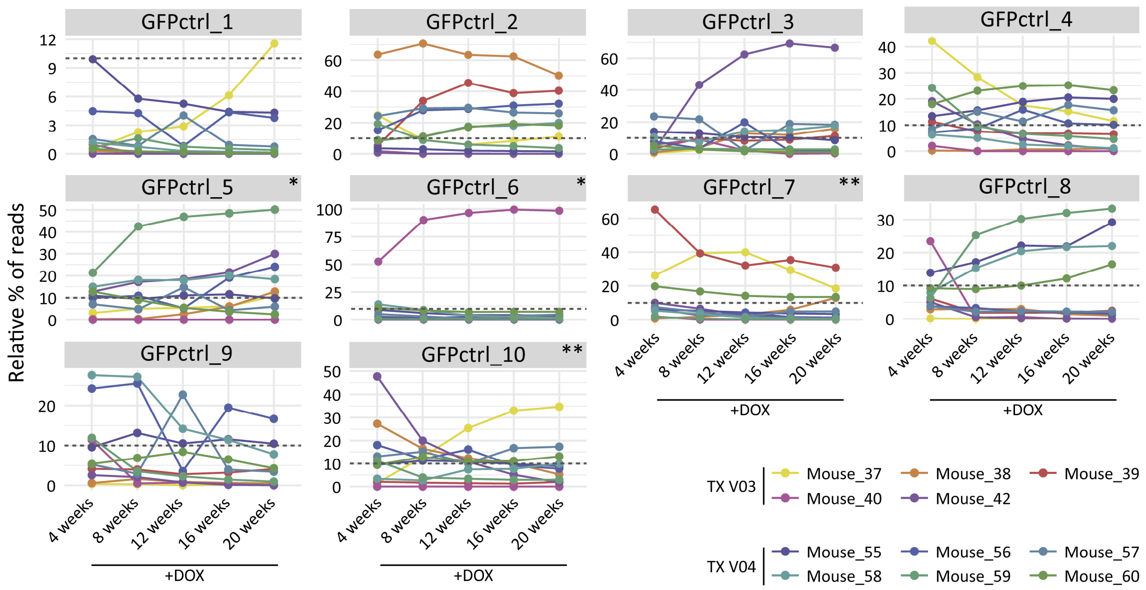


Figure 25 | **Relative proportions of GFP control BCs per mouse over time, related to Figure 24.** Relative % of BC counts over time Data was not normalized to 4 weeks after transplantation to indicate initial BC representation. **TX V03, TX V04**, Transplantation cohort No. 3 + 4, respectively (Table 2). Gray dashed lines indicate theoretical mean proportion for each barcode (2.5%). Friedman Test: * = p ≤ .05, ** = p ≤ .01, *** = p ≤ .001, **** = p ≤ .0001, ns = not significant.

3.3.4 Statistical evaluation of BC combinations indicates that changes in BC proportions are partially driven by a biological effect

Due to the experimental design with 10 BCs representing empty GFP OE constructs and only 2 BCs representing one GOI, a direct comparison using the average of BCs depicting the same gene would be biased. On the one hand, averaging over all 10 BCs in the GFP pool would result in an artificially stable BC representation, as effects would cancel each other out. On the other hand, creating the average of pairs of two GFP pool BCs can be biased, depending on which pairs are created. To circumvent this problem, we calculated the p-values using a Wilcoxon-test for all 45 possible BC combinations of the GFP pool comparing 4 weeks with 8-20 weeks after transplantation and determined the percentage of significant BC combinations (Figure 26A and B). For comparison, we also calculated the p-values for all 780 possible BC combinations from the GOI pool as well as the p-values for the true BC combinations. We hypothesized that BC combinations that depict the same gene should behave similarly, while BC combination of different genes would only behave similarly by chance. In fact, only approx. 5% of all BC combinations from the GFP pool exhibit a p<0.05, which corresponds to the 5% false discovery rate (FDR), hence the amount of significant combinations one would expect by chance. In contrast, about 15% of BC combinations are significant when combining GOI pool BCs. This indicated that most genes or BCs, respectively, behave similarly (most of BCs decline over time) and thus result in more significant combinations than expected. However, the percentage of significant combinations is highest (approx. 25%) when comparing true BC combinations, indicating that BCs behave most similarly when allocated to the same gene (Figure 26B). In summary, this analysis demonstrates that the over expression of our GOIs might indeed show a biological effect. Nonetheless, given the large differences and BC deviations between individual mice, higher transduction efficiencies and better engraftment or combined GFP and GOI pools are needed draw definite conclusions.

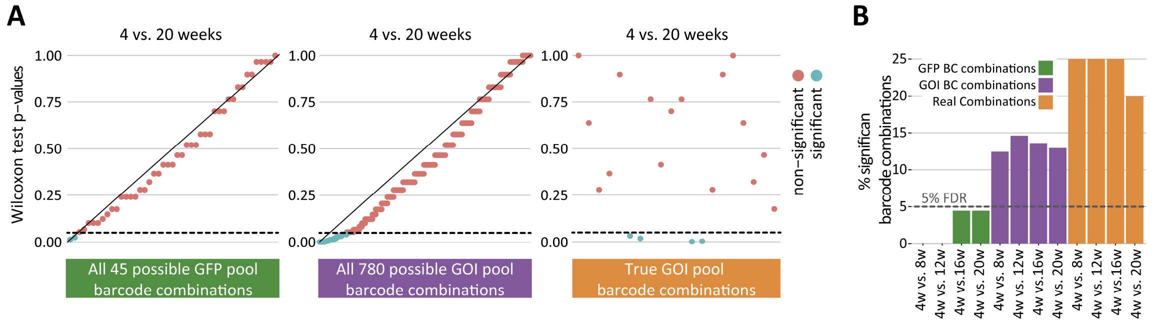


Figure 26 | **Wilcoxon-test for all possible barcode combinations.**

A | p-values of Wilcoxon-test comparing all possible combinations of the mean of 2 BCs at 4 and 20 weeks after transplantation. p-values are sorted for decreasing significance along the x-axis except for true GOI combinations. Here p-values are sorted according to gene name (alphabetically). Grey dashed line indicates p-value of 0.05. Black diagonal represents the theoretical distributions of p-values of infinite tests of random data-pairs. **B** | Percent significant combinations for 4 vs. 8, 12, 16 and 20 weeks.

The following paragraphs of the results contain text sections that have been taken from Wünsche et al., (2018) and have been originally written by myself. All literal quotes are indicated by quotation marks (" … "), following the guidelines of good scientific practice of the Ruperto-Carola University of Heidelberg.
Reprinted or adapted figures and tables from Wünsche et al., (2018) are indicated as such either in the figure legend or table header.

3.4 Mapping active gene-regulatory regions using γ–retroviral integration sites

During the last decade, intensive investigations and technological advances have reshaped the understanding of insertional preference of viruses. Instead of being scattered randomly across the genome, γ-retroviral integration sites (γRV ISs) have been found to almost exclusively accumulate in so called common integration sites (CIS), a phenomenon that primarily occurs due to insertional preference for active transcription start sites (TSS) and active strong enhancers (Cattoglio et al., 2007; Cattoglio et al., 2010; De Ravin et al., 2014; Deichmann et al., 2011; LaFave et al., 2014; Sharma et al., 2013; Sultana et al., 2017; Suzuki et al., 2002; Wu et al., 2003). As a consequence, γRV ISs can be used as molecular tags to map the aforementioned active regulatory regions in the genome of transduced cells (Romano et al., 2016). Due to the stable integration into the genome, transduced cells inherit the ISs to all offspring, which enables to derive information about the regulatory landscape of the cell of origin through sequencing and mapping of ISs in their descendants. Thus, we hypothesized that γRV ISs sequenced during steady-state hematopoiesis in differentiated blood and BM cells from patients that were transplanted with transduced HSPCs would point towards regulatory regions in long-term reconstituting HSCs. To this end, we re-analyzed the complete repertoire of 181,055 ISs or 130,637 unique ISs, respectively, and combined these data with an array of publicly available datasets (Table 3) for meta-analysis and validation purposes.

Table 3 | **Overview of γRV ISs and complementary datasets used in this study. Taken from Wünsche et al. (2018)**

Cell Type	Type of Data	Quantity / Type	Source	Build
CD34$^+$ cells post-transpl.	γRV ISs	130,637 IS	This study	GRCh37/hg19
CD34$^+$ cells pre-transpl.	γRV ISs	1,014,151 IS	De Ravin et al., 2014	
CD34$^+$ cells xenotranspl.	γRV ISs	16,288 IS	De Ravin et al., 2014	
CD34$^+$ cells pre-transpl.	γRV ISs	209 IS	Aiuti et al., 2007	
CD34$^+$ cells post-transpl.	γRV ISs	484 IS	Aiuti et al., 2007	
HepG2 cells	γRV ISs	2,620, 137 IS	LaFave et al., 2014	
K562 cells	γRV ISs	230,950 IS	LaFave et al., 2014	
13 primary blood cell types	Fast ATAC-seq	5,000 cells	Corces et al., 2016	
13 primary blood cell types	RNA-seq	1,000-100,000 cells	Corces et al., 2016	
CD34$^+$ cells	Capture Hi-C	418,037 interactions	Mifsud et al., 2015	
CD34$^+$ cells	ChIP-seq	(H3) K4me1, K36me3, K27ac, K9me3, K27me3, K4me3	Bernstein et al., 2010	
CD34$^+$ cells	ChIP-seq	CTCF binding sites	Jeong et al., 2017	
CD34$^+$ cells	Hi-C	TAD boundaries	Rao et al., 2014	
GWAS SNPs	Medical impact	24,435 GWAS SNPs	NHGRI-EBI Catalog	
SNPs	No known impact	38,128,476 SNPs	NCBI human variants	

xenotranspl, human CD34$^+$ cells transplanted into NSG mice; **GWAS**, genome wide association study; **SNP**, single nucleotide polymorphism.

3.4.1 The majority of γRV ISs are located in enhancer regions

Because we aimed to use γRV ISs as molecular tags to mark active genes and gene-regulatory regions, we first sought to confirm the reported integration preference in our patient cohort. Indeed, we found that ISs sharply peak around TSS as reported in the earliest studies that investigated γRV insertion biology. However, in line with more recent studies, we found that 70% of ISs locate further away from TSS than ±5kb, indicating that ISs predominantly mark non-promoter regulatory regions (Figure 27A). ISs that locate in a 10kb window around TSS can be further segregated into gene classes, showing that protein-coding genes are by far the most common class, followed to a much lesser extend by lincRNAs and lastly miRNAs (Figure 27B). However, one has to note that this is not corrected for the abundance of the different gene classes. The notion that ISs preferentially target active regions was also further supported by a 20-fold enrichment of active histone marks (H3K4me1, H3K27ac, H3K4me3) compared to repressive marks (H3K9me3, H3K27me3, H3K36me3). Particularly co-occurrence of H3K4me1 and H3K27ac modifications are a surrogate for strong active enhancers, while H3K4me1 are usually absent from promoters (Zhou et al., 2011) (Figure 8 and Figure 27C).

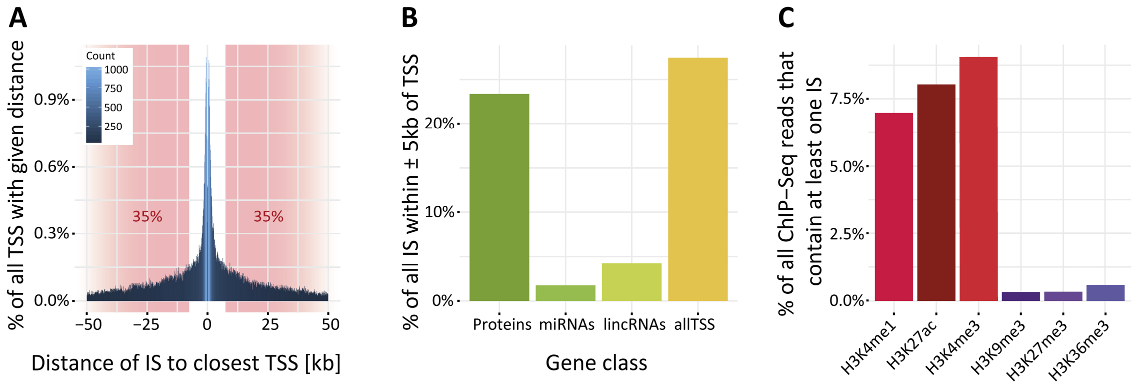


Figure 27 | **ISs mainly cluster around protein-coding gene TSS and histone marks associated with active promoters or enhancers**
A | Distance of ISs to closest TSS. Red areas with percentages depict ISs further away than ±5kb. **B** | Percent of all genes that are marked by ISs closer than ±5kb segregated by gene classes. **C** | Association of ISs positions and ChIP-seq signal for major active (red shades) and repressive (violet shades) marks. Subpanel A and C are adapted from Wünsche et al. (2018)

3.4.2 Switch from short-term to long-term hematopoiesis after transplantation occurs after 6-12 months

In order to map regulatory regions in long-term contributing HSCs, we sought to determine the transition of short-term (transient) to long-term hematopoiesis for all patients. After BM transplantation an early transient reconstitution phase mostly driven by highly proliferative progenitor cells occurs, quickly supplying the organism with new blood cells followed by a long-term hematopoiesis phase, driven by HSCs (Busch and Rodewald, 2016). While studied extensively in mice, the dynamics of this process have only recently been revealed for non-human primates and humans (Biasco et al., 2016; Kim et al., 2014). The switch between these two phases of reconstitution is characterized by a change in clonal association between time points, with low association during the early phase and higher association during the stable long-term phase. To estimate the switch in our patient cohort, we calculated the pairwise positive association (odds ratio) between all samples and time points for all patients with sufficient sequencing depth and time points (Figure 28). Moreover, we developed a mathematical model-fit that describes the increase in association for patients objectively based on the mean $\log_2$ odds ratio for all sequencing time points. After modeling the data, the switch was defined as 30% of the functions maximum (Figure 28, horizontal dashed lines). In line with the study from Biasco et al., (2016), we detected a noticeable change in association (onset of long-term hematopoiesis) between 6-20 months after transplantation. Unfortunately, patients 3, 6, 7, and 10 had insufficient data for a robust determination of the switch, so instead we used the median of 404 days as a stringent cut off – calculated from the remaining patients. These results illustrate not only the transient phase of hematopoiesis, but also the stable long-term commitment of HSCs, indicated by a continuous positive association starting from 6 months up to more than 6 years post transplantation. These findings are also in line with blood reconstitution experiments after irradiation in non-human primates, where clones appeared transiently for 6-12 months after transplantation, which eventually were replaced by long-term repopulation HSC whose progeny were still detectable after 12 years (Kim et al., 2014). Based on the patient specific cut off, we excluded ISs sequenced before the switch in order to prevent contamination with progenitor-derived IS positions, yielding 79,424 unique IS. A detailed overview of ISs per patient, separated into total, early and late sequenced ISs according to the patient-specific switch is provided with Table 4. Hereinafter, only ISs sequenced after the switch to long-term hematopoiesis (79,424 IS) were used for analysis, unless stated otherwise.

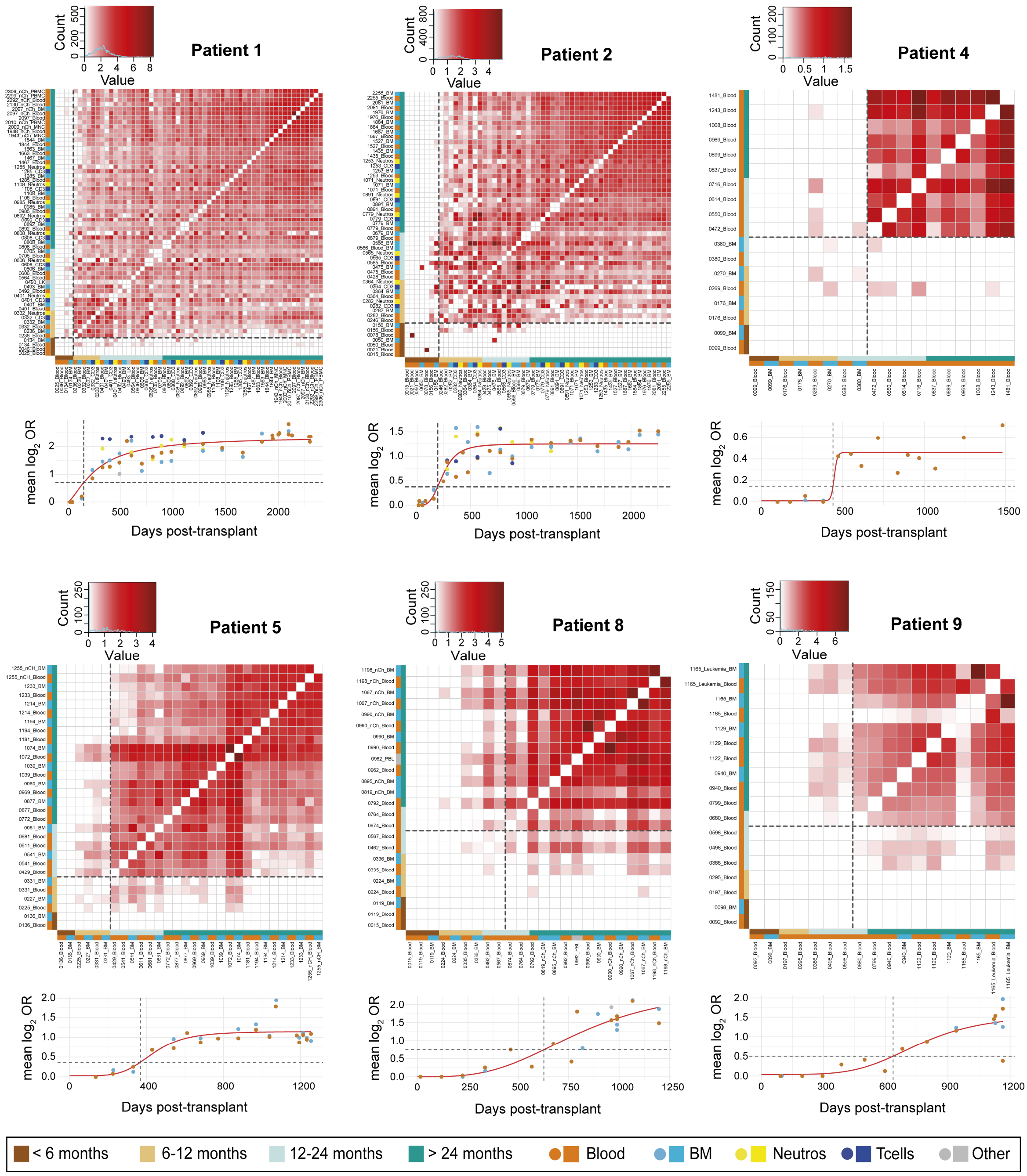


Figure 28 | **Positive association matrices from patient 1, 2, 4, 5, 8, and 9 indicate a switch from short-term to long-term hematopoiesis after 6-20 months.**

Positive association was plotted with the color intensity being proportional to the $\log_2$ odds ratios (OR). Samples are ranked for time-point of sequencing and color coded. Sequencing time points are categorized into < 6 months, 6-12 months, 12-24 months and > 24 months and color coded. For every patient, a log-logistic model is used to fit the mean $\log_2$ odds ratio of the pairwise positive association (red curve). The switch from early to stable (late) hematopoiesis was defined as 30 % of the functions maximum. Adapted from Wünsche et al. (2018).

Table 4 | **Overview of ISs per patient, separated into total, early and late sequenced ISs according to indicated switch. Taken from Wünsche et al. (2018)**

Patient	#samples	#total ISs	#early ISs	#late ISs	%early ISs	%late ISs	Switch [d]
1	60	8,137	866	7,271	10.64	89.36	153
2	55	19,124	1,348	17,776	7.05	92.95	192
3	6	928	928	-	100.00	-	*404**
4	18	17,951	9,439	8,512	52.58	47.42	440
5	30	23,897	9,334	14,563	39.06	60.94	367
6	24	10,229	8,194	2,035	80.11	19.89	*404**
7	22	12,212	2,368	9,844	19.39	80.61	*404**
8	24	16,294	10,937	5,357	67.12	32.88	627
9	18	15,886	9,551	6,335	60.12	39.88	635
10	13	11,032	3,302	7,730	29.93	70.07	*404**

#, number of; **d**, days.

3.4.3 Integration site pattern of γRVs mark cell-type specific regulatory elements

Because the activity of enhancers and promoters is highly cell type specific (Heinz et al., 2015), we investigated the IS pattern across all 10 patients including four additional public datasets, which contain γRV ISs from human CD34$^+$ (pre and post-transplantation in NSG mice), myelogenous leukemia K562, and hepatocellular cancer HepG2 cells. We expected that the IS pattern are similar between patients, as both, the initial population of CD34$^+$ cells before transplantation as well as the resulting population of HSCs after engraftment should be comparable. To this end, we performed Pearson correlation as well as Principle Component Analysis (PCA) using the relative percentage of all ISs in genome-wide 10kb bins and adjusted for differently sized datasets by random sampling (Figure 29A and B). Pearson correlation showed a high similarity between all patients except patient 3, who suffered from engraftment failure, which was also reflected by PCA. Interestingly, while ISs from CD34$^+$ cells showed a high Pearson correlation with patient IS, PCA revealed a distinct difference in the 2nd component. This difference decreased two months after transplantation of CD34$^+$ cells into NSG mice (xenotransplant), indicating a converging of the IS patterns from patients and CD34$^+$ cells during engraftment. Expectedly, ISs from K562 or HepG2 cells showed a poor Pearson correlation with patient ISs and located far outside the patient cluster in the PCA (Figure 29A and B). These results underline the non-random and cell type specific integration nature of γRV and indicate that transplantation and engraftment renders the regulatory landscape highly similar among patients and most importantly different from the initial state in CD34⁺ cells.

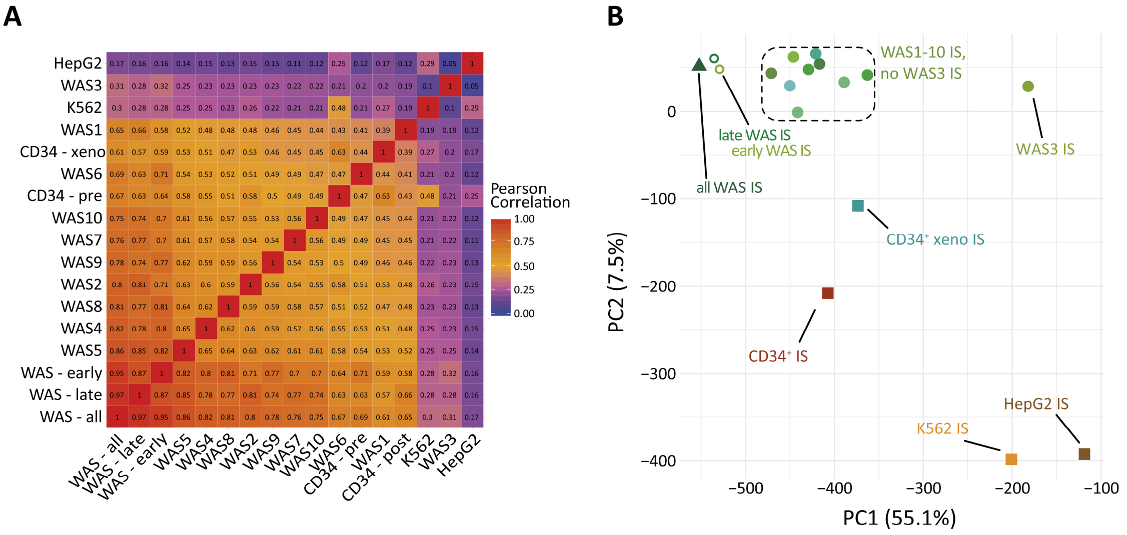


Figure 29 | **Similarity between IS pattern is highest among patients**
A | Pearson correlation matrix of all patient ISs and *in vitro* IS data sets ranked according to principle component 1 in PCA. Color intensity is proportional to correlation coefficient, which are depicted in the squares. **B** | PCA on the same datasets as used for the Pearson correlation matrix. Percentages in brackets represent the proportion of variance explained by this component. **all WAS ISs**, IsS pooled from all patients; **early/late WAS ISs**, ISs occurring either during short-term (early) or long-term (late) hematopoiesis; using the patient specific cut off; **CD34⁺ ISs**, ISs from CD34⁺ cells before transplantation; **CD34⁺ xeno ISs**, after transplantation of human CD34⁺ cells into NSG mice; **K562 ISs/HepG2 ISs**, ISs from K562 or HepG2 cells, respectively. Subpanel B is adapted from Wünsche et al. (2018)

3.4.4 Rainfall plots visualize commonalities and differences between patient ISs and ISs from CD34⁺ cells

Next, we visualized the differences between patients IS and a number matched representative sampling of ISs from CD34⁺ cells in greater detail by plotting the genomic distances from one IS to its consecutive IS (inter-IS distance, Figure 30). Interestingly, the fundamental properties of both datasets did not appear substantially different, as both datasets showed the same characteristic gap of inter-IS distance at approx. 10 kb. This indicates that most ISs are contained within clusters and only the minority has a greater distance to the neighboring IS of more than 10 kb (Figure 30). Moreover, the distribution of ISs along the genomic scale is also similar between the two datasets, again indicating an absence of major clonal skewing events, which would have led to the appearance of very few large clusters instead of thousands of small ones. Nonetheless, there are noticeable differences in the distribution of the top 100 biggest clusters. For example, WAS patients exhibit few very large clusters (e.g. Chr3 – *MECOM*, Chr11 - *LMO2*), which have indeed occurred due to insertional mutagenesis, as reported previously (Boztug et al., 2010; Braun et al., 2014). However, most of the remaining clusters appear at a similar size compared to those observed for CD34$^+$ cells (Figure 30). This indicates an overall unchanged clonality again arguing against a frequent outgrowth of clones due to insertional mutagenesis (with the exception of the aforementioned oncogenic examples). Yet, the positions of the top 100 clusters are considerably different, suggesting an underlying dissimilarity between cells in patients and CD34$^+$ cells. As the initial population of cells in the gene therapy study was also CD34$^+$, it is conceivable that the observed differences have occurred post transplantation.

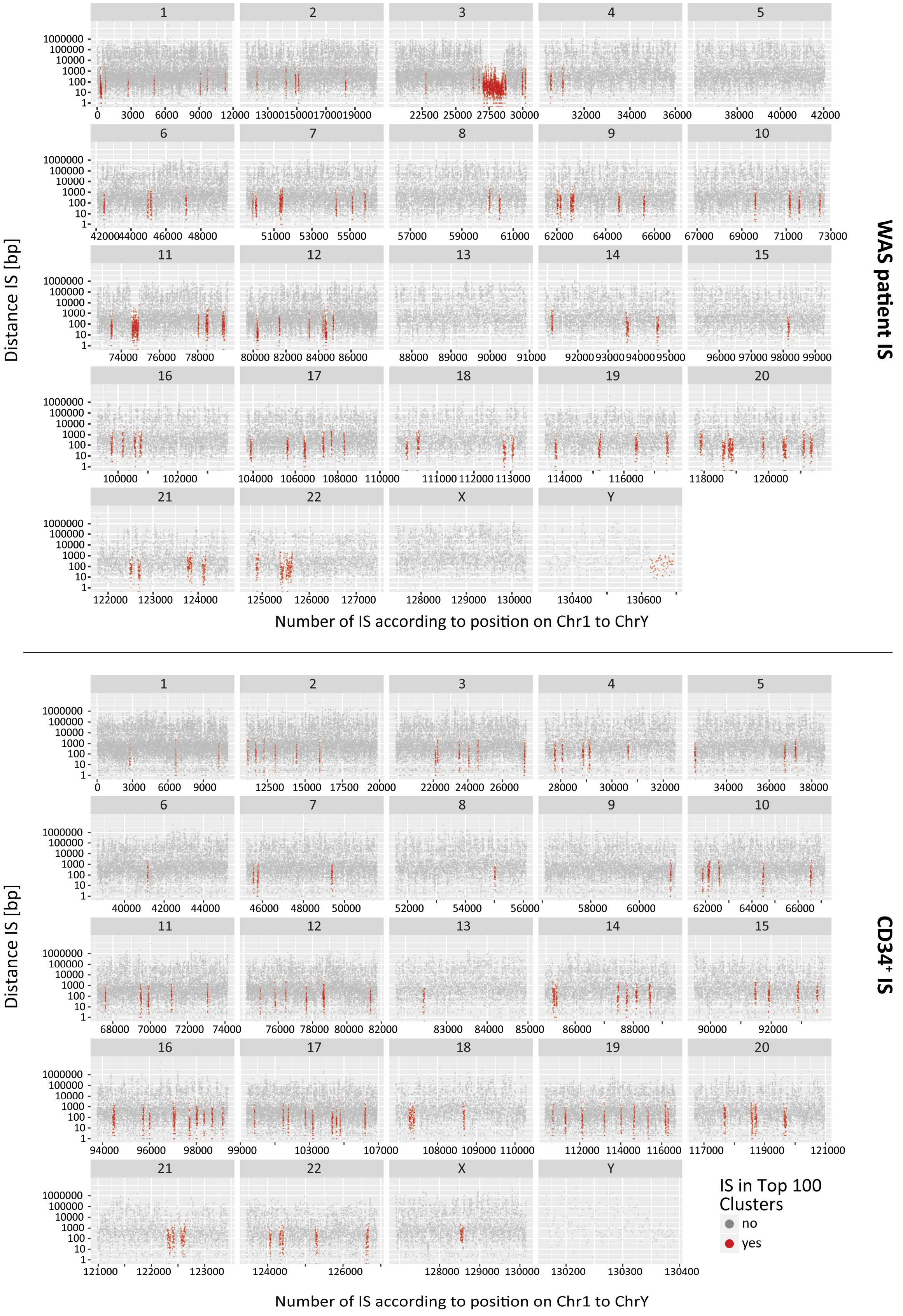


Figure 30 | **Rainfall plots of ISs from WAS patients and CD34+ cells visualize commonalities and difference between the two datasets (Legend continued on next page).**

IS are shown as grey or red dots and are numbered and ordered on the x-axis according to their position in the genome, segregated by chromosome. The position on the $\log_{10}$ y-axis corresponds to the distance in bp to the subsequent IS, calculated using the *imd* function from the ClusteredMutations package for R. ISs contained within the top 100 clusters (according to number of ISs per cluster) are marked red.

3.4.5 Clonal skewing of ISs pattern is restricted to few known leukemogenic loci

It has been known for over a century that γRV insertions have a mutagenic or transforming potential via activation of proto-oncogenes (Ellermann and Bang, 1908; Rous, 1911). In fact, for many years γRVs were used to screen for oncogenes in a variety of mouse tissues (Kool and Berns, 2009). Here γRV ISs were used as genetic molecular tags pointing towards proto-oncogenes due to the outgrowth of malignant clones, resulting in a shift from a polyclonal to an oligo- or monoclonal pattern within the studied system – an unfavorable side effect when highly complex clonality is desired. Consequently, we sought to estimate the impact of clonal skewing in the patients on the complexity of our dataset. As of 2014, seven out of ten patients within the WAS gene therapy cohort developed either AML or T-ALL. However, ISs in dominant clones or blast cells detected in patients were found almost exclusively in the vicinity of *MDS1*/*EVI1* (*MECOM*) or *LMO2*, followed to a lesser extent by *MN1* (secondary AML), *SETBP1*, *PRDM16* and *CCND2*, all of which were observed previously during benign or malignant expansion of clones in other γRV gene therapy trials (Braun et al., 2014; Hacein-Bey-Abina et al., 2003a; Hacein-Bey-Abina et al., 2003b; Howe et al., 2008; Ott et al., 2006; Stein et al., 2010). However, only clones carrying ISs in the proximity of MECOM and LMO2 were commonly found across all patients, suggesting a driver role of these genes only. To address whether other unexplored oncogenic events led to clonal skewing of our dataset, we compared the quantity of ISs per cluster as well as the cluster sizes found in patients with three *in vitro* IS datasets (Figure 31A). Importantly, clonal skewing of the *in vitro* datasets can be excluded as all three *in vitro* IS datasets were sequenced 2-4 days after transduction. Although there was a statistical significant difference detected in cluster dimension and size between patient ISs and most of the *in vitro* datasets, the magnitude was neglectable and comparable to those between the individual *in vitro* datasets themselves (Figure 31A). As neoplastic cell growth after transformation usually occurs with some latency, we also compared early and late occurring cluster dimensions and sizes in patients (Figure 31B). Strikingly, we did not observe any significant differences, again indicating that clonal skewing only occurred for very few and already reported loci and does not resemble a general phenomenon applicable to all IS.

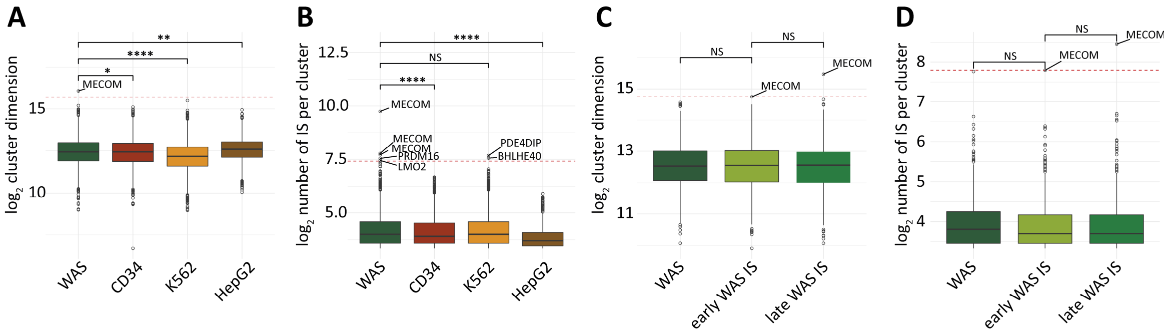


Figure 31 **| Mean cluster dimensions [bp] and sizes of patient and *in vitro* IS datasets as well as early and late patient ISs are highly similar.**

A | Comparison of cluster dimensions [bp] and size [number of ISs per cluster] of patient ISs and three *in vitro* IS datasets. Plots show one representative sampling of 1,000 samplings of the *in vitro* IS datasets to match the number of patient IS. Genes closest to clusters, which exceed the mean (red dashed line) of the maximum cluster size or dimensions of 1,000 samplings from all *in vitro* datasets are labeled. **B** | Comparison of cluster dimensions [bp] and size [number of ISs per cluster] of early and late occurring ISs as described above. Genes closest to clusters, which exceed the maximum cluster size or dimensions of early WAS ISs are labeled. **NS**, not significant; *p < 0.05; **p < 0.01; ***p < 0.001; ****p < 0.0001. Reprinted from Wünsche et al. (2018).

To further expand on this, we measured the cumulative number of ISs that mark the same gene over time as a surrogate for proliferation and calculated the area under the curve (AUC, Figure 32A). We hypothesized that clones marked by unique ISs which show accelerated proliferation should be detected at a higher rate as compared to normally behaving clones. Confirmatively, only 13 genes, most of which are established drivers of clonal expansion, were detected to exhibit an AUC greater than the 95% confidence interval (CI, Figure 32B) and were thus statistically suspicious. The vast majority (96.8%) however shows an accumulation of ISs over time at a slow but constant rate, again leading to the conclusion, that the majority of ISs are of inert behavior and do not lead to clonal expansion.

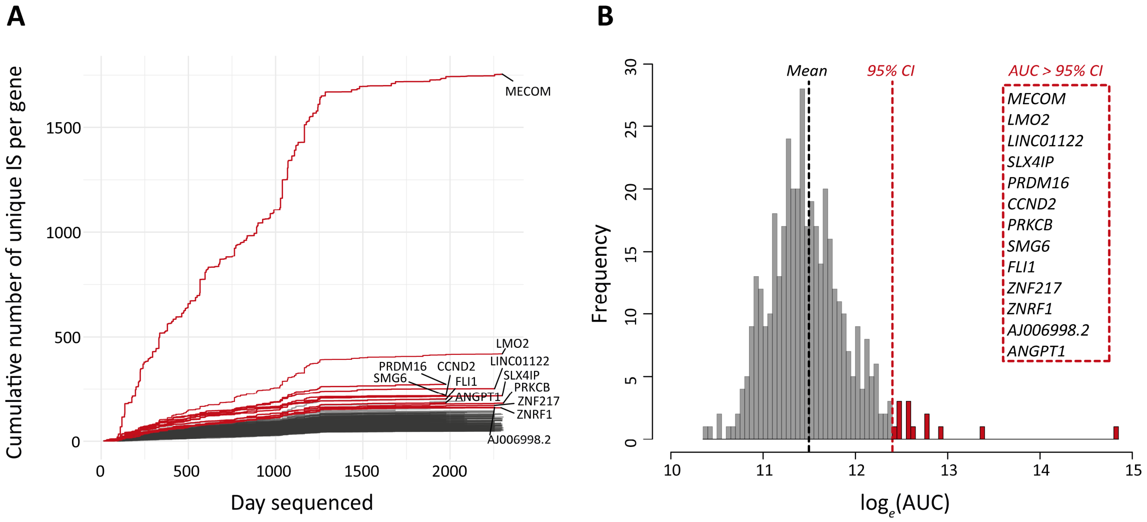


Figure 32 | **Statistical analysis of occurrence of ISs over time highlights known leukemogenic drivers.**
A | Cumulative number of unique ISs per gene with time point of sequencing. Genes with a greater $\log_e$(AUC) than the 95% CI are labeled red. **B** | Frequencies of the $\log_e$ area under the curve (AUC) for all genes calculated from the cumulative amount of ISs depicted in Figure 32A. The grey dashed line indicates the mean, the red dashed line indicates the 95% CI. Genes that show a higher $\log_e$(AUC) than the 95% CI are labeled red. **CI**, confidence interval. Adapted from Wünsche et al. (2018).

3.4.6 HSC specific ATAC-seq signal intensity but not gene expression correlates with IS pattern

Because we hypothesized that γRV ISs from patients sequenced during long-term hematopoiesis originate from long-term HSCs, we matched RNA-seq and ATAC-seq data from 13 primary human blood cell types with our IS data. As both, ATAC-seq signal – a measure for accessible chromatin – as well as gene expression signatures are cell type dependent, we aimed to identify the cell population that resembles the CIS positions best. To this end, we correlated either the expression level of genes with the number of ISs in their vicinity or the ATAC-seq signal intensity at CIS, respectively (Figure 33).

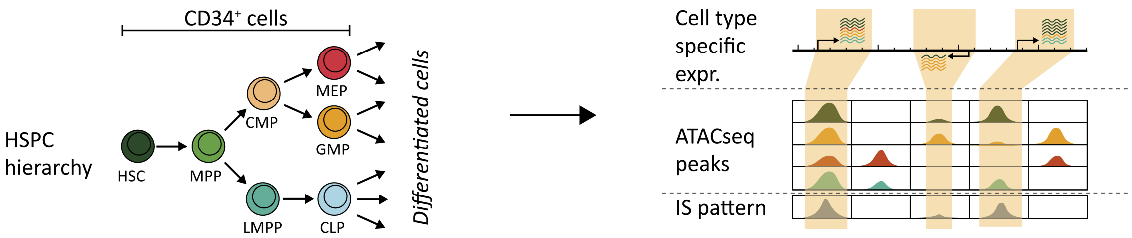


Figure 33 | **Conceptual outline.**
Correlation of the cell type specific gene expression with the number of ISs in their vicinity and intensity of ATAC-seq signals at sites of IS. Related to Figure 34, Figure 35 and Figure 36. Adapted from Wünsche et al. (2018).

In contrast to the findings of De Ravin et al. (2014), the expression of genes in HSCs but also in any other cell type did not correlate with the number of ISs when considering all ISs to its closest gene regardless of the distance (Figure 34A). “Accordingly, the median expression of IS-tagged genes in HSCs, although ranked highest of all blood cell populations, was not significantly different compared to their expression in most progenitor populations (Figure 34B).” Even testing various genomic windows around TSS did not improve correlation (data not shown). Importantly however, De Ravin et al. only considered ISs in a narrow 2 kb window around TSS and averaged the expression levels for a limited number of bins separated by relative activity. This stands in contrast to our approach, which compares all ISs per gene regardless of the distance with their individual expression level. Therefore, our analysis demonstrates that e.g. the number of ISs which can be used as a surrogate for enhancers and their accessibility may not necessarily correlate with gene expression. On the other hand, enhancers can regulate distal genes while skipping genes in their vicinity, rendering proximity-based approaches very difficult. In summary, correlation of patient IS data and gene expression did not show striking differences between all 13 primary cell types. In contrast, the median ATAC-seq signal intensity at ISs was significantly higher for HSCs compared to all downstream progenitors except for closely related multipotent progenitors (MPPs, Figure 35A), indicating that the WAS ISs positions recapitulate the chromatin configuration of HSCs best. As a control, we also performed the same analysis using ISs derived from untransplanted CD34$^+$ cells, which resemble the IS pattern before transplantation. Unfortunately, we do not have IS data from pre-infusion CD34$^+$ cells that were used in this WAS GT trial. Instead, we used ISs from CD34$^+$ cells from De Ravin et al. (2014), which were mobilized, culture and transduced in a very comparable fashion (Table 5). “Interestingly, ISs from CD34$^+$ cells resembled best the ATAC-seq peaks from common myeloid progenitors (CMPs) followed by MPPs and megakaryocyte-erythroid progenitors (MEP, Figure 35B).”

Table 5 | **Comparison of purification, cultivation and transduction between different studies**

	This study - Boztug et al. (2010), Braun et al. (2014)	**De Ravin et al., (2014)**	**Aiuti et al. (2002), Aiuti et al. (2007)**
Mobilization	G-CSF or G-CSF and CXCR4 inhibitor plerixafor	G-CSF	Not mobilized
Purification	Leukopheresis and CliniMACS system	Apheresis and MACS	collected from bone marrow
Culture conditions, Cytokines	X-VIVO 10 medium with 2mM L-glutamine, 60ng/ml IL-3, 300ng/ml SCF, 300ng/ml FLT3L, and 100ng/ml TPO	X-VIVO 10 medium with 1% HAS, 10ng/ml IL-3, 50ng/ml SCF, 50ng/ml FLT3L, and 50ng/ml TPO	IL-3, SCF, FLT3L, TPO
Vector	CMMP-WASP gRV vector	MFGS-gp91 MLV gRV vector	GIADAl MLV gRV vector
MOI	approx. 5	ND	ND
Rounds of transduction	2 x transduction every 24h	Spinoculation, 3 x transduction every 24h	3 x transduction every 24h
Flask treatment	retronectin coated	retronectin coated	ND

G-CSF, granulocyte-colony stimulation factor; **CliniMACS/MACS**, Magnetic Cell Separation; **IL-3**, Interleukin-3; **SCF**, Stem Cell Factor; **FLT3L**, FLT3-ligand; **TPO**, Thrombopoietin; **ND**, not disclosed

Due to the lack of own ISs data from pre-infusion CD34$^+$ cells, we next investigated whether we observe the same change in ATAC-seq signal intensity after transplantation using ISs "[…] from an independent dataset from a γRV ADA-SCID gene therapy study (Aiuti et al., 2007)" and ISs from CD34$^+$ cells that were transplanted in immunodeficient mice (De Ravin et al., 2014). "Importantly, all 5 patients in the ADA-SCID study lacked any sign of clonal expansion up to 47 months after transplantation. Despite the small number of ISs available (209 ISs pre-transplantation and 484 ISs post-transplantation), we observed a similar change in ATAC-seq signal intensity, comparing pre and post-transplantational IS, thus validating our findings in an independent cohort (Figure 36A and B). Strikingly, ISs from CD34$^+$ cells that were transplanted into NSG mice also showed significant enrichment of ISs at HSC specific ATAC-seq peaks (Figure 36C and D)." Alongside with the differences between WAS ISs and CD34$^+$ ISs observed in the PCA (Figure 29B) and the rainfall plots (Figure 30), these results strongly suggest a post-transplant effect that has led to the enrichment of HSC specific IS.

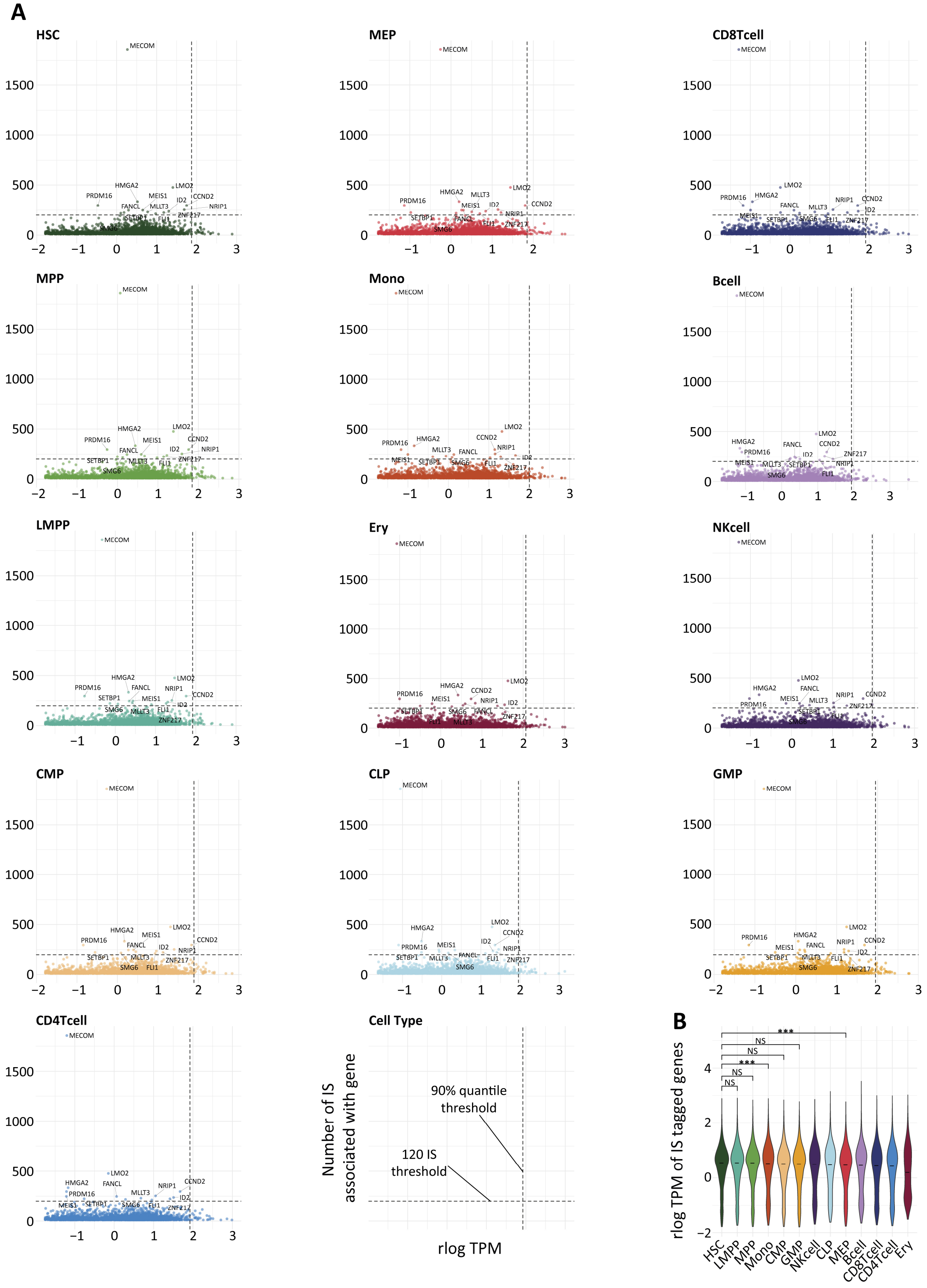


Figure 34 | **Correlation between WAS ISs and gene expression in 13 primary cell types.**

A | Correlation analysis of expression of IS-tagged genes in 13 primary cell types and their corresponding cluster size (number of IS) in patients. Genes that are tagged by a cluster greater than 120 ISs but with an expression below the 90% quantile (grey dashed lines) are labeled. **B** | Violin plot of expression of all

IS-tagged genes in transcripts per million (TPM) ranked for median expression from left to right. Subpanel A “HSC” is adapted from Wünsche et al. (2018).

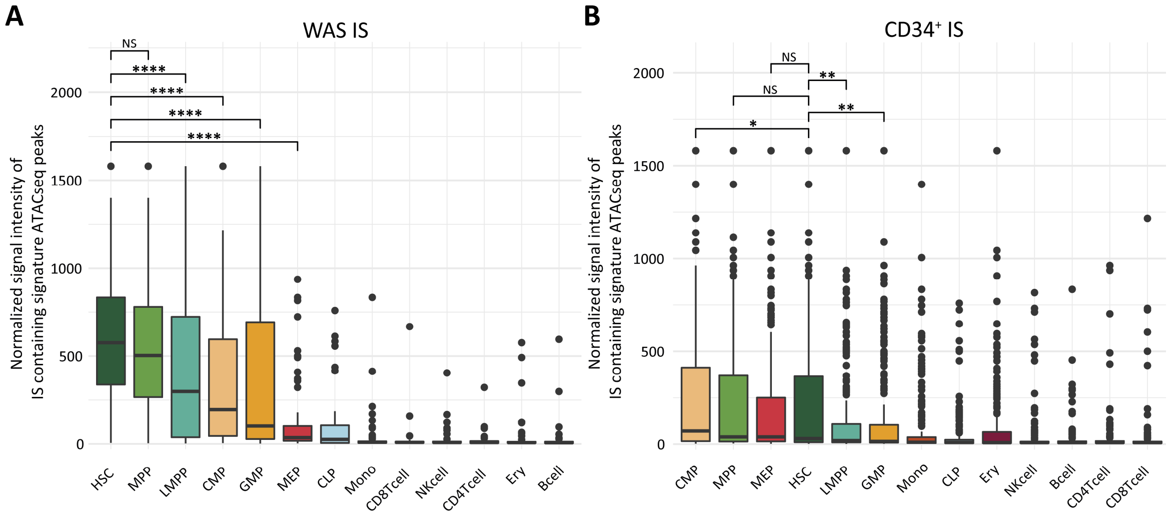


Figure 35 | **Signature ATAC-seq signal intensity at ISs from patients is highest in HSCs, while ISs from CD34$^+$ cells mostly enrich for CMP specific ATAC-seq peaks.**

A | Signal intensity of ATAC-seq signature peaks at sites of WAS ISs ranked for median signal intensity. **B** | For comparison, signal intensity of ATAC-seq peaks was also measured at ISs in CD34$^+$ cells and ranked for median signal intensity. For abbreviation of cell type see 9.1. **NS**, not significant; *****p < 0.05; ******p < 0.01; *******p < 0.001; ********p < 0.0001. Reprinted from Wünsche et al. (2018).

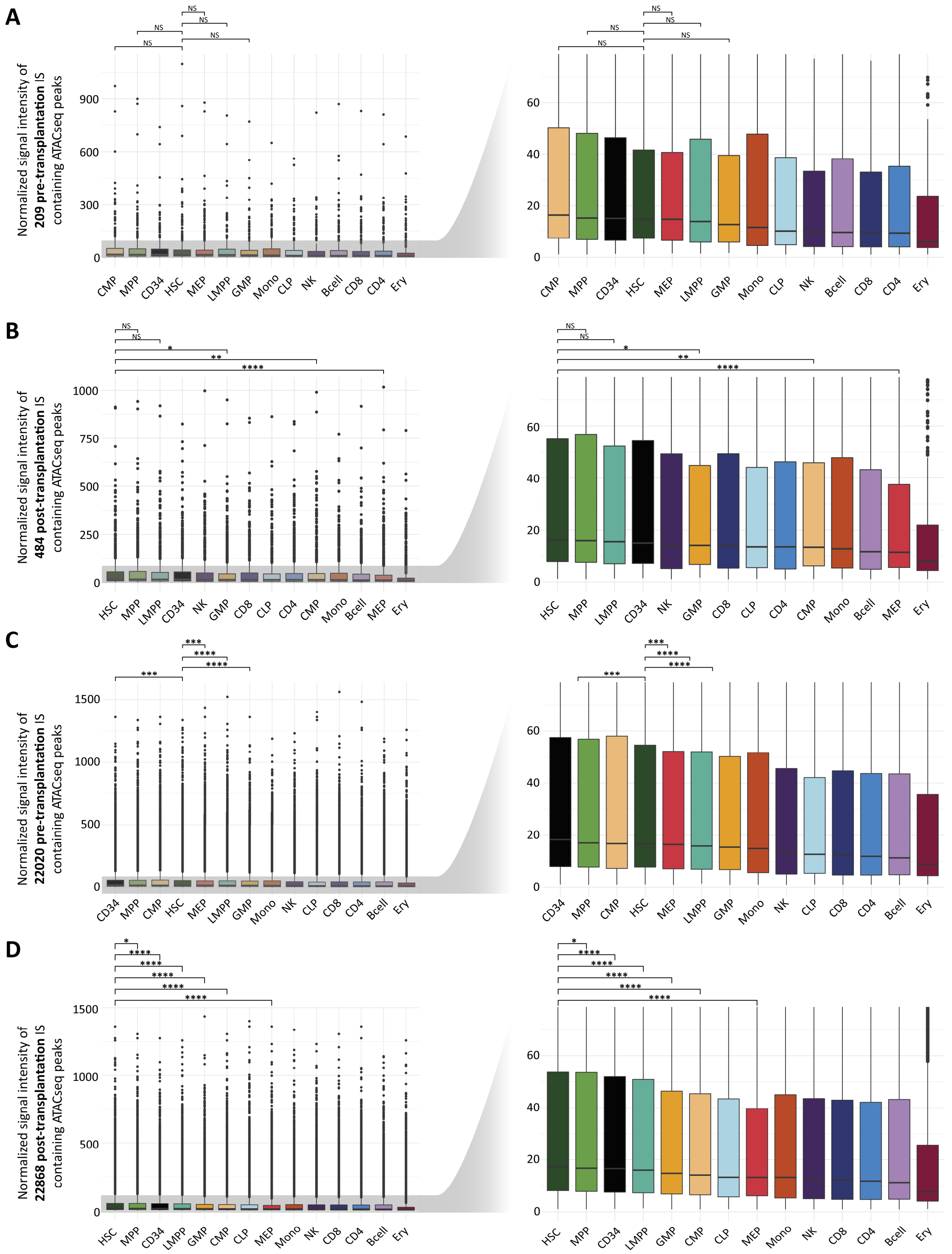


Figure 36 | **ATAC-seq signal intensity at ISs from two independent studies is also highest in HSCs after transplantation. (Legend continued on next page)**

A | Signal intensity of ATAC-seq peaks at sites of 209 pre-transplant ISs from Aiuti et al. (2007) ranked for median signal intensity. **B** | Signal intensity of ATAC-seq peaks at sites of 484 post-transplant ISs from Aiuti et al. (2007) ranked for median signal intensity. C | Signal intensity of ATAC-seq peaks at sites of 22020 pre-transplant ISs from De Ravin et al. (2014) ranked for median signal intensity. **D** | Signal intensity of ATAC-seq peaks at sites of 22868 post-transplant ISs from De Ravin et al. (2014) ranked for median signal intensity. For abbreviation of cell type see 9.1. **NS**, not significant; *****p < 0.05; ******p < 0.01; *******p < 0.001; ********p < 0.0001. Subpanel A, B and D have been reprinted from Wünsche et al. (2018).

3.4.7 Reported key hematopoietic transcription factors are efficiently marked by WAS IS

To further expand on the enrichment of HSC specific ISs after transplantation, we next analyzed the genetic loci of ten previously described key hematopoietic stem and progenitor cell (HSPC) transcription factors (TFs) (Wilson et al., 2010) for their presence of patient ISs (Figure 37A). Moreover, we also included promoter capture Hi-C (CHi-C) data from CD34$^+$ cells (Mifsud et al., 2015), in order to detect interactions of IS-tagged regions with the gene promoters of the TFs. Although derived from CD34$^+$ cells, incorporating the chromatin conformation information through CHI-C data allows to predict putative regulatory regions of GOI. Reassuringly, all ten loci showed intronic ISs or ISs in close vicinity, many of which located in regions that were found to physically interact with the promoter of interest in CD34$^+$ cells. Particularly, the intronic eR1 sub-module of the RUNX1 super-enhancer was efficiently tagged, which has been reported to be specifically active in HSCs and MPPs (Ng et al., 2010). Interestingly, this enhancer module showed a balanced occurrence of early and late IS, probably suggesting equal activity in HSCs and MPPs (Figure 37B).

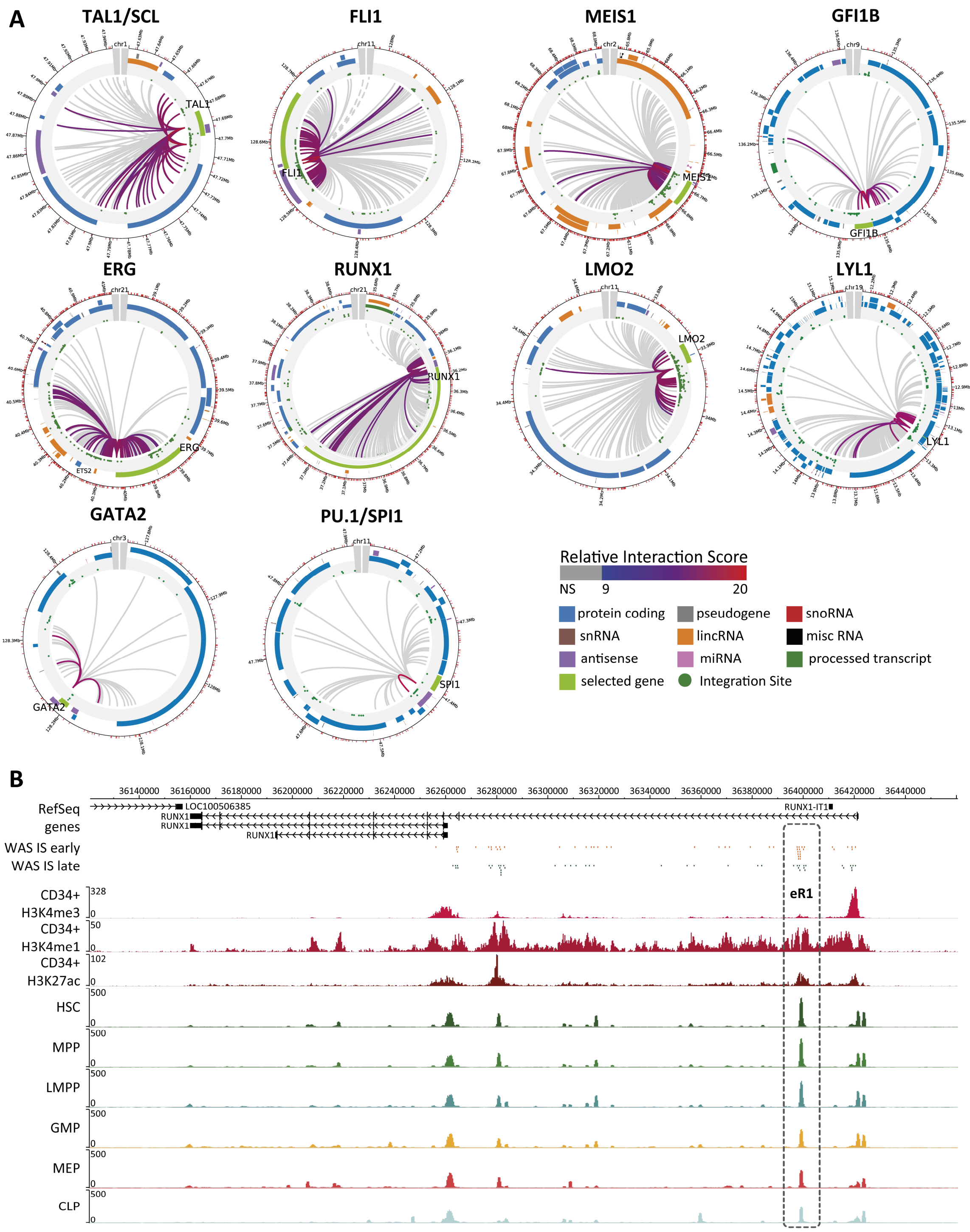


Figure 37 | “**Circular plots showing loci of reported HSPC regulators with CHi-C interaction and WAS IS.**” **“A** | Circular plots of known HSC regulators (Wilson et al., 2010). Shown are significant interactions of promoter regions of selected genes (light green). Patient ISs are depicted as green dots. For clarity reasons, some genes are not shown. **B** | WashU Epigenome Browser view of the RUNX1 locus with ATAC-seq and ChIP-seq signal intensities as well as patient IS cluster. eR1 super enhancer sub module is indicated by the grey dashed line.” Reprinted from Wünsche et al. (2018).

3.4.8 Differences between patient and CD34$^+$ ISs highlight long-term HSC specific genes

As many of the reported hematopoietic key TFs are active in both, HSCs and downstream progenitor cells, we next aimed to further enrich for HSC specific IS, by subtracting the IS signal before transplantation from the signal found during stable long-term hematopoiesis after transplantation. To this end, we down-sampled ISs from CD34$^+$ cells (De Ravin et al., 2014) 1,000 times to match the number of late WAS ISs and subtracted the average number of ISs per gene in CD34$^+$ cells from the number of WAS ISs for the same gene (Figure 38A). Because CD34$^+$ cells mostly consist of progenitors, genes with more ISs in CD34$^+$ cells vs. patients ($\Delta IS < 0$) should highlight progenitor specific regions, whereas regions with more ISs in patients vs. CD34$^+$ cells ($\Delta IS > 0$) highlight HSC specific regions. To test our hypothesis, we performed pairwise comparisons for all IS-tagged genes (Figure 38B) and performed gene set enrichment analysis (GSEA) "[…] on 4,731 curated gene sets (C2) from the MSigDB Collections complemented with 20 custom gene sets from Cabezas-Wallscheid et al. (2014) that specifically compare murine long-term HSCs (LT-HSCs) with short-term HSCs (MPP1) and three other multipotent progenitor types (MPP2-4). Strikingly, 7 out of 9 significantly enriched gene sets ($FDR < 0.1$) were HSC related along with another HSC gene set, ranking 11th, while only one gene set showed significant enrichment ($FDR < 0.1$) of genes that have more ISs in CD34$^+$ cells. Interestingly, mouse gene sets containing genes that are significantly higher in LT-HSCs compared to ST-HSCs and other MPP populations showed the highest adjusted p-values (Figure 38C). Furthermore, we analyzed the overlap between HSC relevant gene sets to address similarity and robustness of the different sets (Figure 38D and E). In summary, differential analysis of CD34$^+$ ISs vs. patient ISs demonstrated that the IS pattern after transplantation changes towards HSC specific genes in line with the enrichment of HSC-specific ATAC-seq peaks after transplantation."

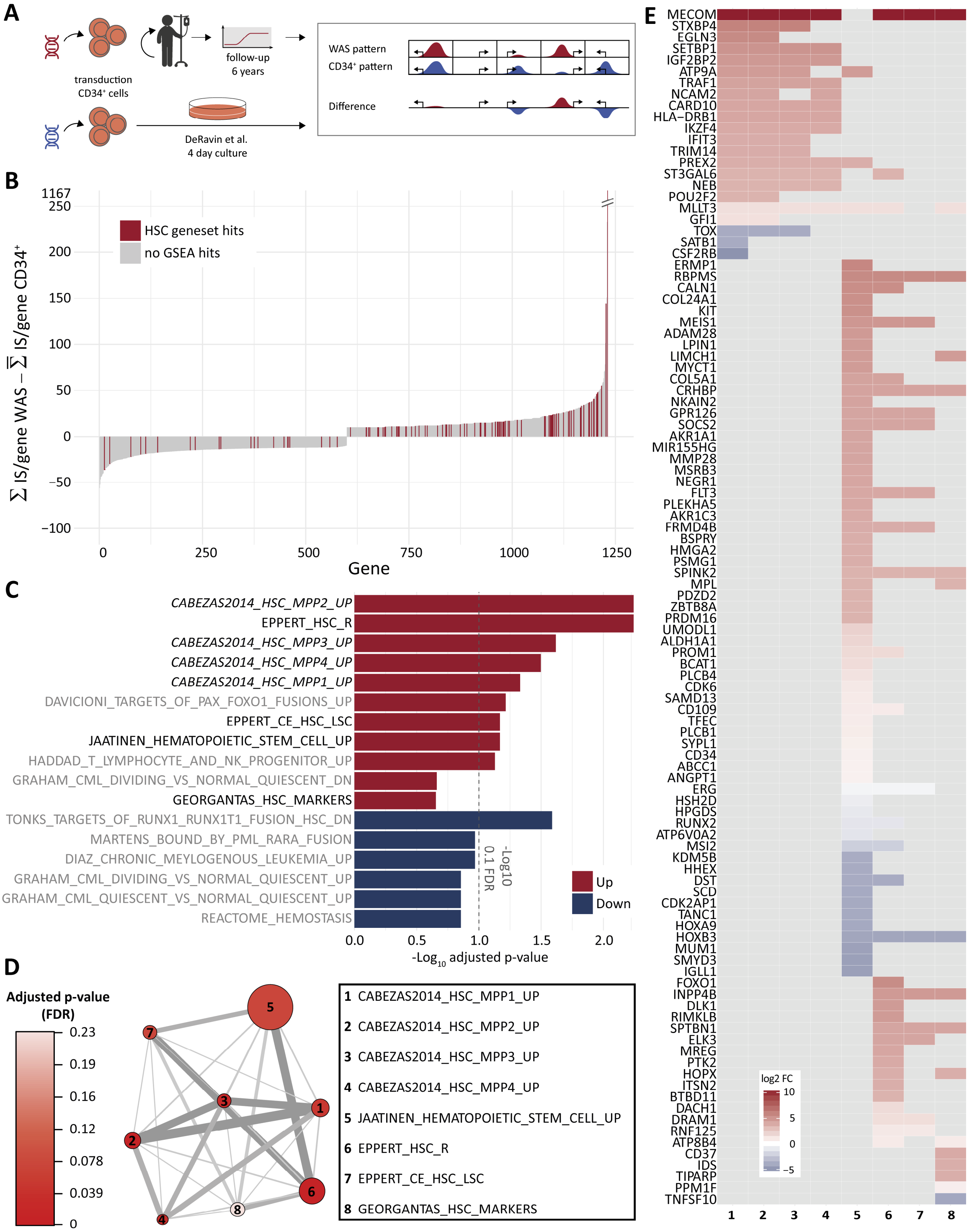


Figure 38 | “**Gene set enrichment of differentially tagged genes in WAS patients vs. un-selected ISs from CD34⁺ cells highlight long-term HSC specific genes (Legend continued on next page).”**

“**A** | Conceptual outline. Difference was measured by subtracting number matched samplings of CD34⁺ ISs from late WAS ISs that are associated with the same gene. **B** | Histogram showing the differences in number of late ISs from WAS patients compared to CD34⁺ cells for each gene. Differences were calculated as the sum of WAS ISs per gene subtracted by the mean sum of CD34⁺ ISs for the same gene from 1,000 random down-samplings to match the WAS data set. Genes that scored in any of the 9 HSC related gene sets listed under C are marked red. **C** | Barplot of gene sets enriched for genes with positive fold-change in red (FDR < 0.2) and genes with negative fold-change (FDR < 0.2) are displayed. Grey dashed line indicates FDR –log10 (0.1). In total, 4,751 gene sets were analyzed. Murine gene sets are italicized. **D** | Network plot illustrates overlap between HSC relevant gene sets. Circles are numbered according to gene set. Node size is proportional to number of genes within gene set while edge width corresponds to number of shared genes. **E** | Heatmap of scoring genes from HSC relevant gene sets. Number on x-axis correspond to number in Figure 38D. Gene sets are sorted from left to right according to p-value, genes are sorted for increasing fold-change. Color intensity reflects $\log_2$ fold change; grey = gene not present in gene set. **FDR**, false discovery rate." Reprinted from Wünsche et al. (2018).

3.4.9 The time point of ISs detection indicates activity of enhancer modules in HSCs and progenitor populations

So far we have established that ISs from WAS GT patients detected after the switch from short to long-term hematopoiesis originate from bona fide long-term HSCs. Moreover, we showed that ISs locate in chromatin regions most accessible in HSCs compared to other hematopoietic cell populations and that genes nearby clusters of WAS ISs have been previously linked to HSC functions. However, as observed for the eR1 RUNX1 enhancer sub-module (Figure 37B), the time point of sequencing of WAS ISs harbors additional information, as WAS ISs sequenced early are more likely to be derived from progenitors, while WAS ISs sequenced later are increasingly more HSC-specific. Along with the 10 key TFs described above, MYC also displays another key hematopoietic TF with well-established roles in HSC biology. Although the gene promoter itself harbors only very few IS, "[…] the recently reported blood enhancer cluster (BENC) that has been shown to physically interact in CD34$^+$ cells with the 1.7 megabases downstream located MYC gene shows well defined clusters of ISs (Figure 39A) (Bahr et al., 2018). The human BENC consists of at least 8 enhancer modules with selective activity in different blood cell populations, this way regulating MYC expression throughout the hematopoietic hierarchy. By assessing early and late occurring ISs as well as the last time point of clone detection within BENC modules, we were able to recapitulate the reported results (Figure 4B). The median time point correlated well with their suggested activity, with persistent ISs in the HSC specific modules C/D and more transient or short-lived clones detected in the progenitor like modules G/I. In line, module A/B, which is equally active in HSCs and progenitors showed both, transient and long-lived clones (Figure 39C). Strikingly, we also identified three additional modules (X1, X2 and X3) that are located outside of the reported module boundaries and therefore were not experimentally addressed before, all of which showed equal or even higher HSC specificity than modules C/D (Figure 39C). Collectively, these data demonstrate the power of γRV ISs mapping in long-term engrafted CD34$^+$ cells to identify HSC regulatory regions.”

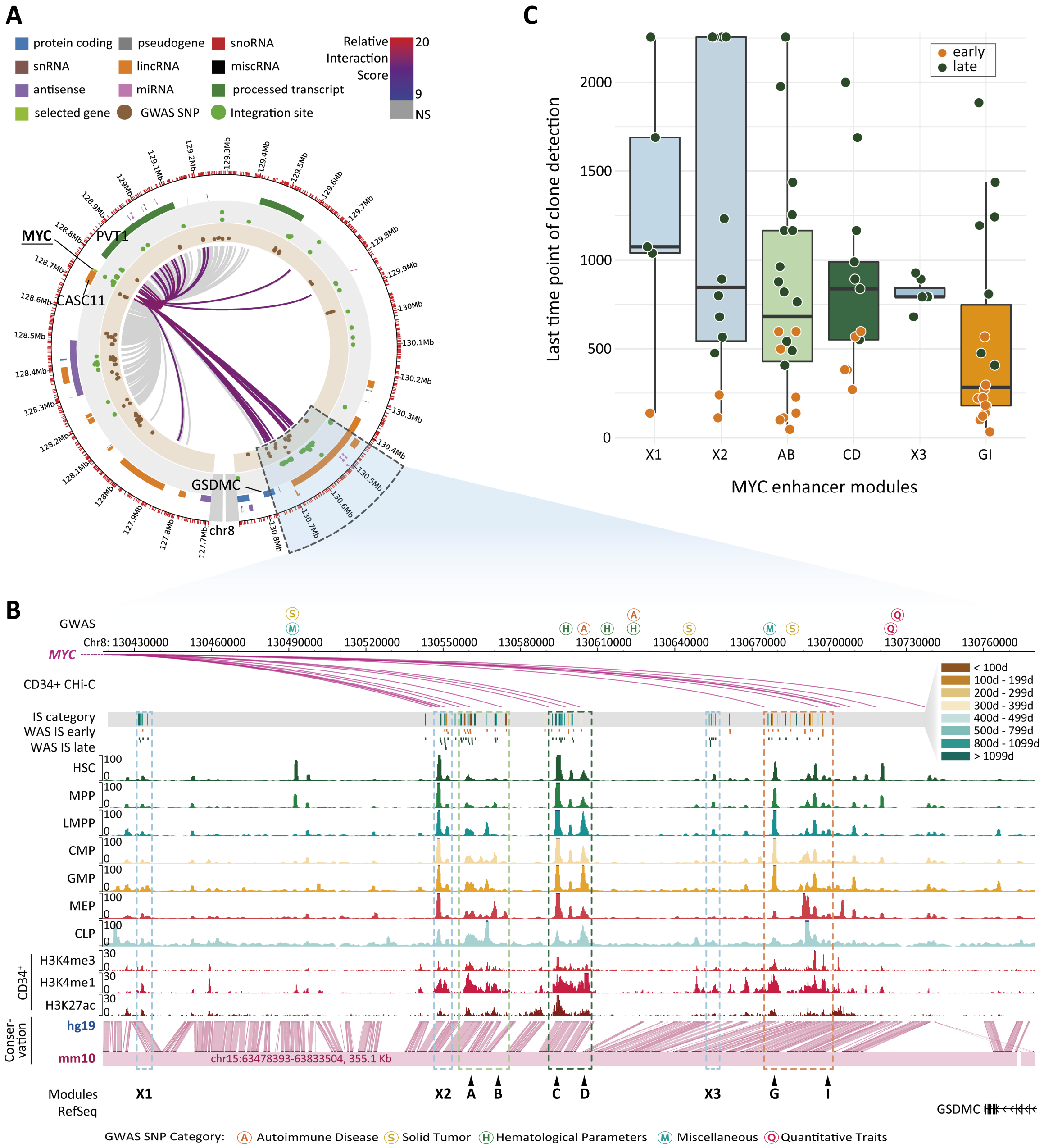


Figure 39 | “**Differential IS patterning at MYC enhancer correlates with selective activity of enhancer modules in HSCs and progenitor populations.”**

“A | Circular plot of the MYC locus including the BENC region. Shown are significant interactions (color of arcs represents interaction score) with the MYC gene (light green). Patient ISs are depicted as green dots. The BENC module region is highlighted by a dashed box. **B** | WashU Epigenome Browser view of the BENC region with ATAC-seq signal intensities, GWAS SNPs as well as early and late patient IS clusters. BENC enhancer sub module are indicated by the colored dashed boxes. **C** | Representation of the sequencing time points of ISs at known (A, B, C, D, G, I) and novel sub modules (X1, X2 and X3). **BENC**, blood enhancer cluster.” Reprinted from Wünsche et al. (2018).

3.4.10 Tagged enhancers and disease variants can be linked to their putative target gene by integration with long-range interaction data

“A major challenge for the investigation of regulatory non-coding elements remains the prediction of their target promoters as not all genes are controlled by regulatory elements located in close proximity to their TSS. As described above, we used promoter capture Hi-C (CHi-C) data from primary CD34^{+} cells to assign non-coding regions of interest to their prospective gene or promoter (Mifsud et al., 2015). One of the most heavily IS-tagged genes in our dataset was Nuclear receptor interacting protein 1 (NRIP1), which shows interaction with an array of IS-tagged enhancers spanning over more than 500kb (Figure 40A). Interestingly, despite the global lack of correlation of gene expression and number of IS, NRIP1 is expressed highest in HSCs with a gradual decline towards more committed progenitors (Figure 34B and C, Figure 40B), indicating that for some genes a high expression is indeed associated with a high number of IS. As many IS clusters are generally associated with interactions (Figure 37A and Figure 40A), we next examined if this enrichment was significantly higher than expected by chance. Interestingly, patient ISs showed a striking enrichment, which was highest for promoter interactions (approx. 6.4-fold, $p \approx 0$), followed by all interactions (approx. 2-fold, $p \approx 0$) and non-promoter interactions (approx. 1.6-fold, $p \approx 0$) (Figure 40C). ISs from CD34^{+} cells were slightly more enriched, which was also reflected by the higher Chi-Square value (39,084 vs. 17,340). However, as the enrichment was not overly different between WAS ISs and ISs from CD34^{+} cells, it seems conceivable that HSCs and CD34^{+} cells share the majority of interactions, increasing the likelihood that interactions at clusters of WAS ISs indeed point towards their putative target gene.”

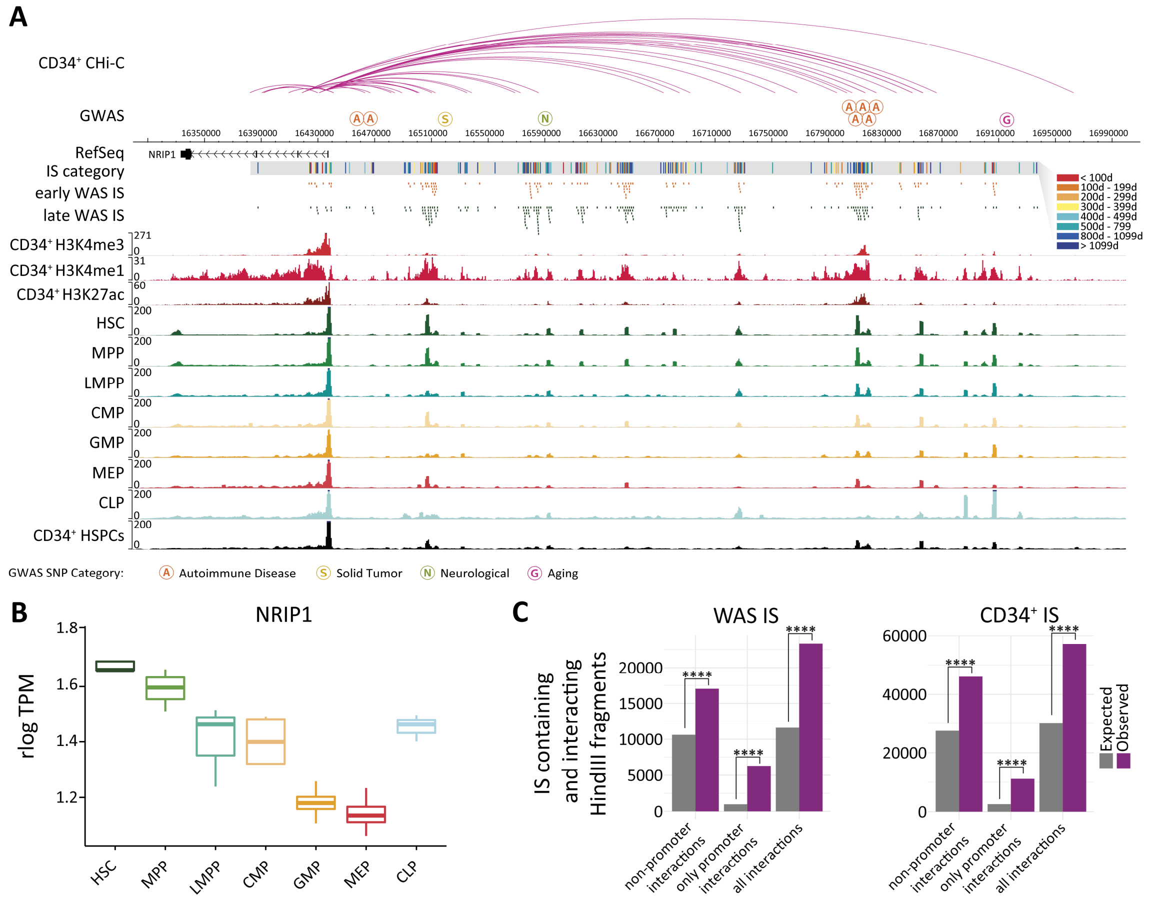


Figure 40 | **"Integration of long-range interaction data links ISs to enhancers and disease variants of their putative target gene."**

"A | ATAC-seq and ChIP-seq signal intensities, early and late WAS ISs and capture Hi-C (Chi-C) interactions from CD34+ cells at the NRIP1 locus. GWAS SNPs are indicated according to position and subclass. **B** | RNA-seq TPM expression data of NRIP1 in HSCs and progenitor populations. **C** | Chi-square test with Yates's correction for continuity to compare interaction fragments either containing or not containing ISs (WAS or CD34^{+}). Genomic HindIII restriction sites coordinates are equivalent to those of interaction fragments. **TPM**, transcripts per million." Reprinted from Wünsche et al. (2018).

3.4.11 Clusters of ISs co-localize with hematological GWAS SNPs

"We next asked whether IS-tagged regions possess functional properties across tissues and if these regions are particularly important in the human blood system. To address this question, we utilized a collection of 24,495 genome wide association study single nucleotide polymorphisms (GWAS SNPs) along with number-matched random samplings of approx. 3.8 x 10^{9} common SNPs with no known medical impact. As enhancers are generally more prone to carry disease variants compared to other non-coding regions (Corradin and Scacheri, 2014; Ernst et al., 2011), we expected γRV ISs to be enriched in the vicinity of GWAS SNPs. Indeed, WAS ISs were highly enriched near GWAS SNPs compared to common SNPs (p = 8.35 x 10^{-64}), which was also observed to a similar degree for all other in vitro IS datasets (Figure 41A). Next, we classified all traits and diseases into 17 categories (adapted from Mifsud et al., 2015; Maurano et al., 2012, Figure 41B) and calculated their relative enrichment or depletion. GWAS SNPs categorized into hematological parameters were the most significant and highest enriched sub-class (Figure 41C). A similar pattern was observed for $CD34^+$ and K562 derived IS, while the hepatocellular carcinoma cell line HepG2 showed strongest enrichment for serum metabolites SNPs (Figure 41D-F). Collectively, these findings suggest that clusters of γRV ISs do not only mark regulatory regions but also seem to favor elements with functional roles in a cell type-specific manner."

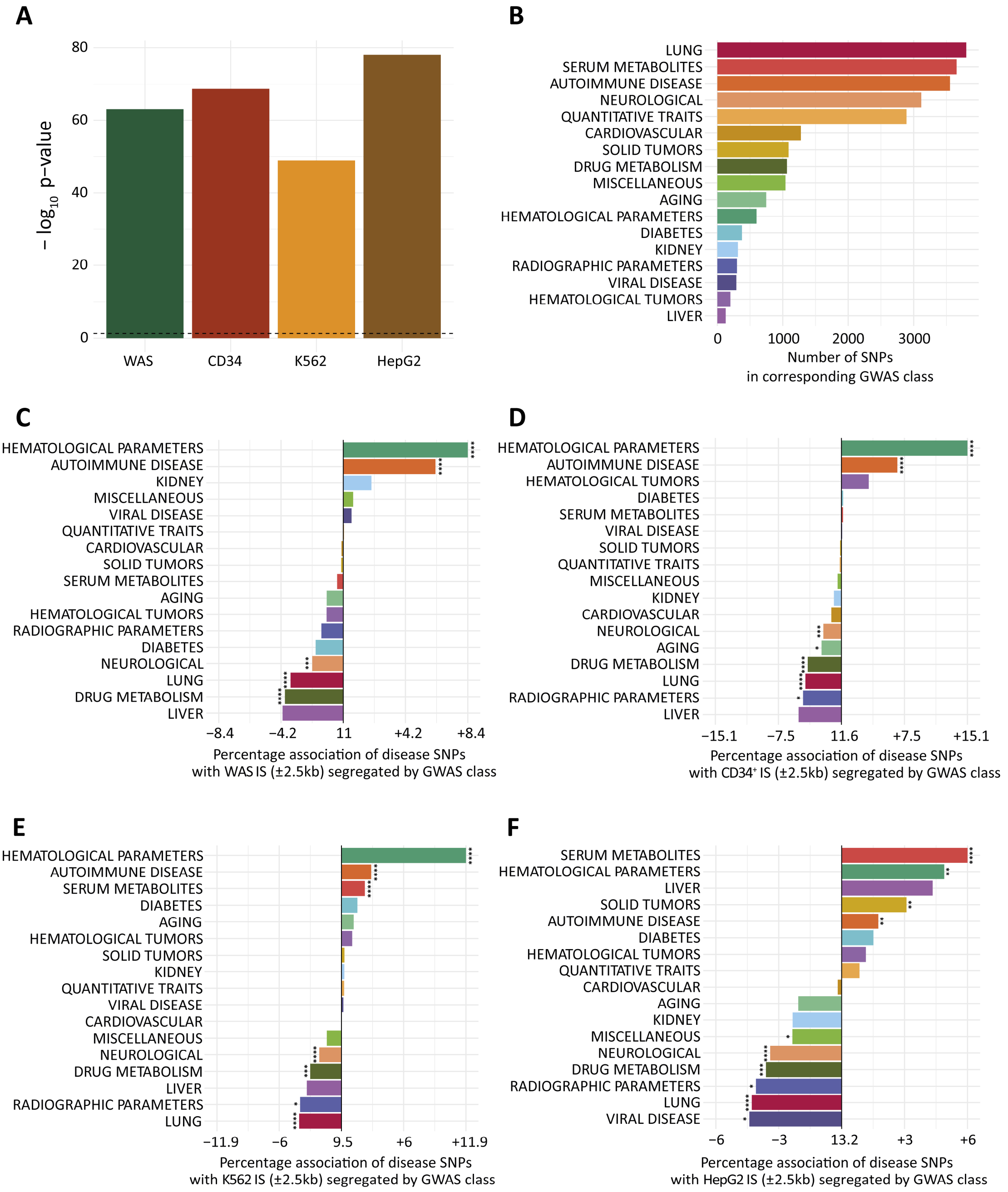


Figure 41 | "**p-values of GWAS SNP enrichment at sites of ISs along with categorical enrichment across IS datasets.**"

"**A** | Significance of GWAS SNPs co-occurring with ISs compared to common SNPs. Black dashed line indicates –log10 p value of 0.05. **B** | Number of GWAS SNPs for each category. **C-F** | Percentage of GWAS SNPs that overlap with a window of 5 kb around IS-derived from indicated cells segregated by subclasses. Bar plot shows relative enrichment or depletion over mean association. Bars with no asterisks were not significantly altered from mean. **No asterisk**, not significance; *p < 0.05; **p < 0.01; ***p < 0.001; ****p < 0.0001." Adapted from Wünsche et al. (2018).

3.4.12 Integration sites show elevated sequence conservation

Since we detected a significant enrichment of GWAS SNPs at sites of viral integration, we investigated the sequence conservation around ISs across 46 primates. First, we checked whether cluster positions coincided with particularly conserved regions by visual inspection of some of the most prominent clusters. Interestingly, conservation appeared to peak to a greater extent in close vicinity to clusters than at clusters itself (Figure 42A and B). To expand on this, we measured the phyloP and phastCons conservation scores in a 500bp window around ISs as well as around GWAS SNPs, and common SNPs, and used the genomic mean for comparison (Figure 42C and D). PhastCons score describe the likelihood for each nucleotide to belong to a conserved element based on the multiple alignment of *n* given species, such as the 46 primates used here. Moreover, phastCons score also considers the flanking elements, making the score more sensitive to consecutive stretches of conserved elements. By contrast, phyloP scores ignore the context of neighboring elements, rendering it more appropriate for estimating the evolutionary selection at particular nucleotides or classes of nucleotides. While phyloP score can measure evolutionary acceleration, indicated by a negative score (-log p-values under a null hypothesis of neutral evolution), phastCons score only represent probabilities of negative selection and range between 0 and 1 (Hubisz et al., 2011; Pollard et al., 2010; Siepel et al., 2005). As described before, both phastCons and phyloP scores sharply decline at sites of common SNPs (Castle, 2011). Likewise, GWAS SNPs showed a similar pattern, although the overall scores are higher compared to common SNPs and averaged above the genomic mean (Figure 42C and D). This is in line with the results from Ma et al. (2015), who showed that GWAS SNPs categorized into complex disease variants (the majority of GWAS catalog SNPs are complex disease variants) show no conservation, indicated by another conservation score – the GERP score. Interestingly, phastCons scores also decline at sites of IS, yet to a much lesser extent than GWAS SNPs and common SNPs ($p = 5 \times 10^{-7}$ and 2.9×10^{-9}, respectively). In contrast, phyloP scores for ISs did not show any decline, but instead appeared relatively stable across the entire genomic window ($p = 4.9 \times 10^{-12}$ - 2.5×10^{-81}). Collectively, the conservation at sites of γRV insertions was significantly higher as compared to the genomic mean or SNPs and GWAS SNPs, pointing towards a functional role of the IS-tagged regulatory regions.

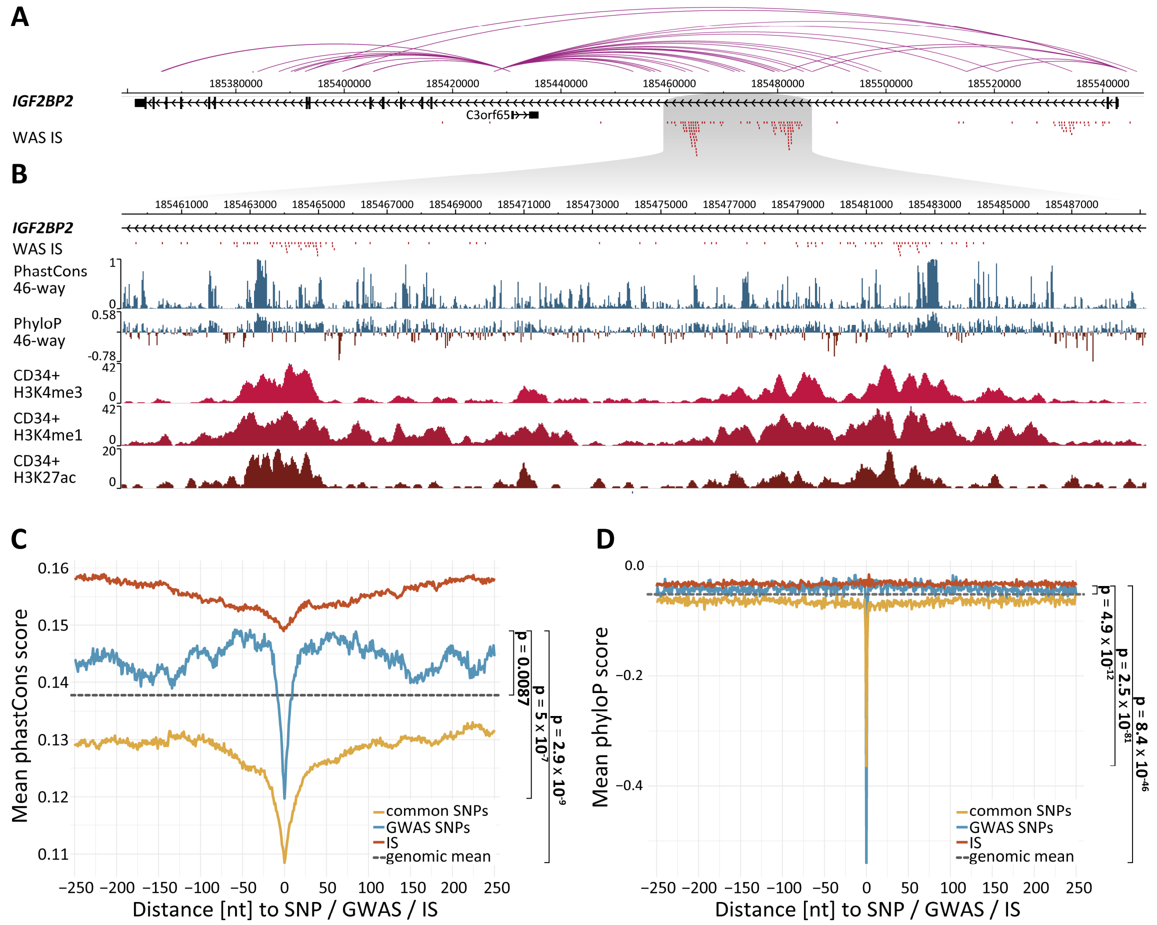


Figure 42 | “**Global phastCons and phyloP conservation scores at ISs […]” point towards functional roles of IS-tagged regulatory regions.**

“**A** | Complete IGF2BP2 locus with CHi-C interactions and WAS IS. **B** | Zoom of above panel as indicated by grey box with tracks for phyloP and phastCons scores as well as ChIP-seq signal for active promoter and enhancer marks in CD34$^+$ cells. **C** | Mean phastCons scores from 46 primates in a 500 nucleotide window centered on 130,000 random SNPs, 24,434GWAS SNPs, or 130,637 IS, respectively. **D** | Mean phyloP scores from 46 primates displayed as described in C.” Reprinted from Wünsche et al. (2018).

3.4.13 Boundaries of topological associated domains but not CTCF sites show significantly more ISs than expected by chance

“As boundaries of topologically associated domains (TAD) also show higher conservation (Harmston et al., 2017), we checked for enrichment of ISs at these sites. Moreover, we also assessed the relative enrichment of ISs at CTCF sites, as both, TAD boundaries as well as CTCF sites are relevant for chromatin integrity that could potentially be disrupted by IS. As TAD boundaries are known to be transcriptionally active (Dixon et al., 2012), we expectedly observed significantly more ISs within TAD boundaries mapped in K562 cells (Rao et al., 2014) than expected by chance (expected: 2.21%-2.26%; observed: 2.60%-3.19%; Percent ISs within TAD boundaries). However, ISs from CD34$^+$ cells before and after transplantation showed a very comparable percentages of IS, […]" thus were neither further enriched nor depleted, "[…] indicating that ISs at boundaries do not grossly disrupt or enhance cell function (Figure 43A). In contrast, we neither observed significant enrichment nor depletion of ISs in CTCF sites mapped in CD34$^+$ cells (expected: 0.12%-0.13%; observed: 0.08%-0.15%; Percent ISs within CTCF sites) (Jeong et al., 2017) (Figure 43B)." Collectively, these results further promote the notion of directed integration at active sites, however also suggest that the integration *per se* did not grossly disrupt the function of such loci.

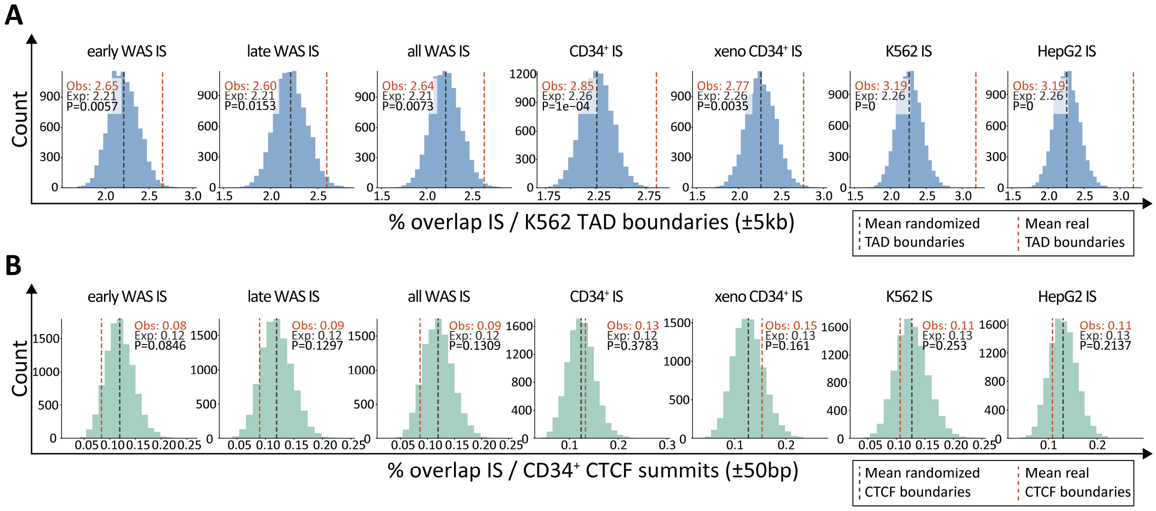


Figure 43 | **TAD boundaries but not CTCF sites show significantly more ISs than expected by chance**
"A | Comparison of the percentage of overlap between indicated IS datasets, with observed and randomized K562 TAD boundaries (±2.5 kb) (Rao et al., 2014). For statistical testing, TAD domains were shuffled 10,000 times while maintaining the original characteristics (size, distance, genome gap exclusion). IS data sets were down-sampled 10,000 times to match the smallest data set (xeno CD34$^+$ IS). Expected (grey) and observed (red) observations are indicated by dashed lines. **B** | Comparison of the percentage of overlap indicated IS datasets, with observed and randomized CTCF sites from CD34+ cells (±50 bp) (Jeong et al., 2017). For statistical testing, CTCF domains were shuffled 10,000 times. IS data sets were down-sampled 10,000 times to match the smallest data set (xeno CD34$^+$ IS). Expected (grey) and observed (red) observations are indicated by dashed lines. **Obs,** observed; **Exp,** expected; **P**, p-value." Adapted from Wünsche et al. (2018).

3.4.14 γRV provide a catalogue of >3,000 regulatory regions in functionally define human long-term HSCs

"In summary, we show that γRV ISs can be used as molecular tags not only for clonal tracking, but also to mark regulatory regions in functionally defined cell populations. Unlike sequencing approaches of phenotypically defined cell populations, our method exploits the natural selection process enriching for long-term repopulating HSCs after transplantation." Through extensive analysis of our own data in conjunction with publicly available datasets, "[…] we were able to detect >79,000 genomic tags from 10 patients," creating a rich resource of >3,000 active gene-regulatory regions in human repopulating long-term HSCs (Figure 44). "These data provide new insights into active regions and regulatory mechanisms of repopulating HSCs and represent a solid basis and comprehensive resource for functional studies investigating stem cell self-renewal and differentiation."

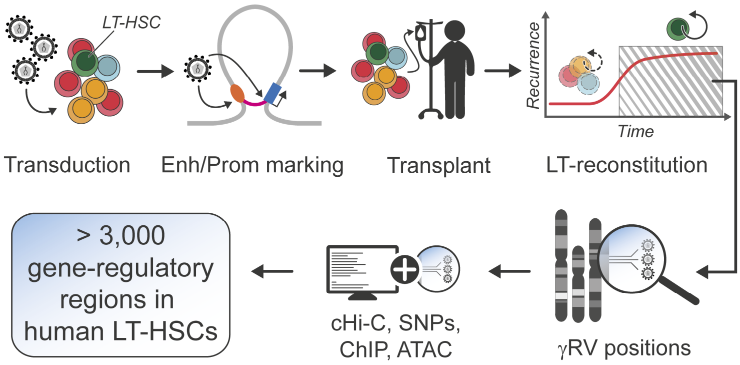


Figure 44 | **Schematic representation showing the concept and strategy of mapping the gene-regulatory regions in human repopulating long-term HSCs.**
After transduction of heterogeneous CD34+ cells from WAS patients, γRV pre-integration complexes mark active enhancers and promoters. Next, cells are re-infused into the patients and engraft which naturally selects LT-HSCs over time. Through LAM-PCR and HT-seq methods, γRV positions are mapped to the genome. Integration of additional data-sets such as CHI-C, GWAS SNPS, ChIP-seq and ATAC-seq data provide further information on IS-enriched regions and validate our approach, eventually leading to the resource of more than 3,000 active gene-regulatory region in human repopulating LT-HSCs. Adapted from Wünsche et al. (2018).

4 Discussion

In the present study, we used a large collection of γ-retroviral integration sites (γRV ISs) that were collected prior to this thesis over a period of 6 years from 10 Wiskott-Aldrich-Syndrome (WAS) patients. In total, we used 181,055 ISs, which map to 130,637 unique sites for the purpose of identifying, selecting and finally validating novel hematopoietic regulatory genes as well as creating a genome-wide resource of active gene-regulatory regions in human repopulation long-term HSCs.

4.1 The lentiviral overexpression-library approach is largely limited by transduction efficiencies and cell numbers

After the selection of candidate genes based on IS cluster size and proximity to the TSS, we first aimed to collect *in vitro* and *in vivo* data for all genes to re-evaluate the list of candidates and to eventually focus on one or two candidates for functional and mechanistic experiments. We hypothesized that transduction of LSK or LSK-SLAM cells with a pool containing all candidate gene constructs would allow to study their influence on hematopoietic dynamics simultaneously. Through unique barcodes (BCs) for every candidate within a small cassette that is suitable for genomic amplification and HT-seq, we aimed to track changes in relative proportion throughout the experimental timeline, caused by alterations in proliferation or differentiation upon candidate over expression.

4.1.1 Tracking clonal dynamics *in vitro* and *in vivo* through genetic barcodes and multiplexed HT-seq is technically feasible

In order to trace and distinguish transduced cells throughout the experiment, we first established the lentiviral overexpression pool. During the first step of the nested PCR, the BC-containing cassette is amplified from genomic DNA. During the second step, different samples can be indexed, using one of 96 different multiplex reverse primers (Figure 17). Finally, up to 96 indexed samples can be mixed to be sequenced on a single Illumina HiSeq 2000, which can eventually be de-multiplexed using the index-read. To ensure that BCs (18nt) and indices (8nt) were sufficiently different in their sequence to prevent cross-contamination ("bleeding") of BCs or indices caused by sequencing errors, we ran a defined series of PCR reactions. Here, each of the 96 BCs was amplified with one of the 96 indices and subsequently pooled and sequenced (3.2.3, Figure 18). The results demonstrated that BC or index-bleeding, respectively, occurred to a negligible degree (0.059% off-target reads per index or 0.0006% off-target reads per index and BC; Figure 19B). This indicated that 1) it is feasible to distinguish all BCs in our small library and 2) multiplex (index) up to 96 samples without considerable BC or index-bleeding. These results were expected due to the relatively high BC/index diversity (minimum distance of two nucleotides between BCs and indices) and low error rate of Illumina HiSeq platforms (~0.1% or 1 in 1,000) (Manley et al., 2016). Thus, the chance that the two nucleotides that distinguish indices/BCs from one another are both sequenced wrongly is only 1,000 x 1,000 hence 1 in 1 million (0.0001%). In case that any base substitution at two positions would generate another BC/index sequence also present in the library (worst case scenario), the chance would increase to 95 in 1 million, thus approx. 1 in 10,000 or 0.01%. However, as many BCs/indices are different in more than two positions, this chance is likely too high. Note, that a difference in three positions between BCs/indices would decrease the chance to 1 in 1 billion. Taken together, the observed rate of wrongly annotated read per single index was slightly higher than expected by chance (exp.: 0.0001%; obs.: 0.0006%). This could be explained by minor spillover or aerosol contaminations between wells during sample preparation, which is also possible during sample preparation of actual experiments. However, overall these miss-allocations of BCs/indices occurred at a very low rate, thus should not have grossly effected the results of the following experiments.

4.1.2 Titer variations in cDNA overexpression library screens

Most *in vitro* screens are performed using knock-down of cellular transcripts or knock-out of genes using RNA interference or more recently CRISPR/Cas9 libraries. Such libraries have the great advantage of uniformly sized shRNAs, siRNAs or sgRNA inserts, resulting in comparable virus titers for all constructs within the library (Miles et al., 2016; Mohr et al., 2010). Unlike gene knock-down or knock-out screens, overexpression screens usually work through continuous high expression levels of wild-type genes. Although the first overexpression screen was already performed in 1982 in yeast cells using a library of random genomic fragments, they are more complex than RNAi or CRISPR screens, due to varying fragment sizes (Carlson and Botstein, 1982). Today, also more sophisticated libraries have been developed for additional organisms including

Homo sapiens covering ~15,000 full-length human cDNAs driven by a cytomegalovirus (CMV) promoter (Liu et al., 2007). While relatively easy in yeast due to transient transfection with 2µ vector-based plasmid libraries, stable overexpression of cDNAs in mammals usually requires γRV or LV packaging and stable transduction (Ludwig and Bruschi, 1991; Prelich, 2012). However, systematic investigation of the packaging limit of HIV-based vectors (LV) has shown that virus titers decreased semi-logarithmical with increasing construct size. Although we did not observe such a striking correlation between size and titer, we still observed some significant variations (3.2.4, Figure 20A). As a consequence, the initial representation of cDNAs or BCs, respectively, might be imbalanced, possibly leading to the over or underrepresentation. To compensate for this, we used an inducible vector system in order to measure the relative proportion of all BCs before cDNA expression for normalization. However, gross variation in titers could lead to biases regarding transduced cell types or bottleneck effects e.g. during passaging or engraftment. In fact, even for uniformly-sized pooled libraries, usually a 500x coverage is aimed for, meaning that the number of infected cells should be at least 500 times higher than the number of genes or BCs present in the library (Doench, 2017). Naturally, with increasing titer variations, this number has to be even higher. Additionally, one has to consider that transduction efficiencies should be kept in the 5-20% range, as higher rates will massively increase the likelihood of double or even multiple integration sites per cell (Doench, 2017). Thus, at a transduction efficiency of 20% the number of starting cells has to be at least 2500 times the number of BCs, making genetic screens with limited primary material extremely challenging.

4.1.3 Transduction with pooled overexpression library in CFU assays is limited by colony number and transduction efficiency

After having established the technical requirement for the overexpression pool, we aimed to assess the influence of our candidate genes on proliferation, differentiation and self-renewal. First, we tested our library in a colony-forming unit (CFU) assay using LSK-SLAM$^{\text{Rosa26 rtTA}}$ cells. As discussed in 4.1.2, it is very important to establish and maintain a sufficient library representation throughout the experiment. However, this can be very challenging when working with limited primary material. Due to the experimental setup, we did not achieve a higher BC representation than 120-fold for the

GFP pool or 30-fold for the gene of interest (GOI) pool, respectively. The fold coverage for individual BCs in the GOI pool might be even higher or lower, depending on the individual titers. This makes interpretation of the results very challenging, as it increases the chance of observing significant chances in BC representation by chance (Figure 45).

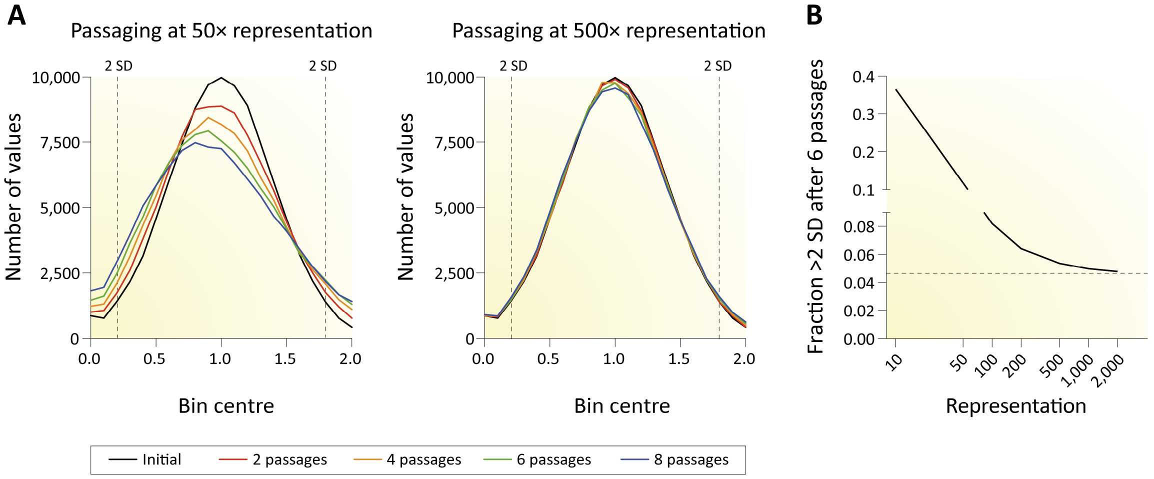


Figure 45 | **Effects of insufficient library representation during several rounds of passaging.**
A | Simulation of the library distribution across passages. At a low 50-fold (50x) representation, the library spreads out, leading to a relatively high amount of perturbations falling outside of the 2 standard deviations (SD). This effect is observed without any selective pressure but is only due to random chance. In contrast, a simulation of a 500-fold (500x) library representation does not show spreading. **B** | Fractions of perturbations greater than 2 SD after 6 passages for different library representations. Adapted by permission from Springer Customer Service Centre GmbH: Springer Nature, Nature Reviews Genetics, Doench (2017), copyright 2017.

One way to control for random enrichment or depletion can be partially achieved through redundancies. This can either be a large number of repetitions of the same experiment or through the usage of different BCs for the same gene. For the latter, we hypothesized that a similar behavior of 2 independent BCs that depict the same gene indicate a biological effect, while opposing behavior of 2 independent BCs that depict the same gene indicate a non-biological effect. However, the relative change of BC representation can generally only attain three different states: Increasing, static or declining. Accordingly, the assessment whether a change in representation occurs as a consequence of a biological effect should also take the magnitude of change into consideration. Applying the above-mentioned criteria, only a few genes showed a consistent phenotype through all passages in the CFU assay. For example, both BCs for *Plcb4* showed a very consistent decline over time, similar to *Irf2bp2*, *Xbp1S* and *Znf217*. None of the genes showed a comparably consistent increase over time.

Conclusively, due to the poor library representation in the CFU assay we did not derive interpretable results. Although we tried to increase the coverage in follow-up experiments, the limited number of cells, the relatively low virus titers as well as the limited number of colonies that can be grown on a plate and the costly reagents did not allow for significant improvements.

4.1.4 Cell trace experiments showed improved library representation but insufficient timeframe

In order to improve upon the low number of cells possible in CFU assays, we aimed to address changes in proliferation using the CellTrace^{TM} dye. To this end, we transduced approx. 1×10^5 LSK$^{Rosa26\ rtTA}$ cells with the GFP pool (10 BCs) and approx. 3×10^5 LSK$^{Rosa26\ rtTA}$ cells with the GOI pool (40 BCs). At a transduction efficiency of 20-25%, this equals a ~2,000-fold BC representation in the GFP pool transduced cells and a ~1500-fold representation in the GOI pool. In fact, the results of the individual experiments showed substantially less deviation between BCs of the same genes compared to the CFU assay. Moreover, most genes behaved similarly in both independent experiments, indicating that the approach was technically more feasible compared to the CFU assay. However, the effect on proliferation was very moderate for most genes. This could be either explained by little or no effect of our candidate genes on proliferation of LSK cells *in vitro* or by the relatively short cultivation time of 3 or 5 days, respectively. The cultivation time however is largely limited by the detection of the dye, which loses approx. half its signal intensity with every cell division (Filby et al., 2015). Usually, the maximum number of divisions that can be traced with this dye is around 7-8, which roughly equals 5-8 days of fast cycling LSK cell *in vitro*. Collectively, this approach seems to overcome the problems with the library representation, however, is not suitable to address subtle changes in proliferation.

4.1.5 Pooled overexpression screen *in vivo* mostly suffers from low transduction efficiencies.

As discussed above, a critical step for genetic screens in general is to maintain an unbiased representation of the library in order to observe real biological effects rather than stochastic noise. Most *in vivo* screens are performed in tumor models using cell lines as starting material. Due to the high number of available cells, these approaches even allow for genome-wide screens with tens of thousands of different BCs or shRNAs/sgRNAs (Chen et al., 2015; Crotty and Pipkin, 2015). In contrast, genetic screens that aim to elucidate the impact of e.g. gene-knockdown on engraftment or differentiation and self-renewal of stem and progenitor (HSPCs) cells are largely restricted by limited input material and poor engraftment. For example, a study that used molecular barcoding to track HSCs (LSK, CD34$^-$ CD150$^+$ CD135$^-$) after transplantation in mice showed that out of 9,000 only 50-80 HSCs stably engraft and propagate per mouse (Lu et al., 2011). Correspondingly, a study by Hope et al. (2010) focused on only 20 genes, each targeted by 2-3 shRNAs to identify new fate determinants of HSCs *in vivo*. Despite the small scale and high gene transfer efficiency of 59% on average, the results still showed substantial variation. In another study from Holmfeldt et al. (2016), a total of 41 genes were screened, each represented by ≥2 shRNAs. However, due to the large number of constructs, the authors had to transplant and analyze a very high number of >1,300 mice. In contrast to the RNAi screens, another group performed a gain-of-function cDNA overexpression screen on 104 candidates and 3 mice for every cDNA (Deneault et al., 2009). However, it has to be noted that all of the above-mentioned screens did not pool the candidates but instead transplanted single-gene transduced cells into individual mice. Thus, engraftment, differentiation as well as self-renewal could be traced by FACS and fluorescent markers and did not require HT-seq detection of BCs. Thus, the tracking is more cost efficient and immediate. Moreover, the transduction efficiency can be drastically increased as double or triple integrations would refer to the same gene and not lead to the overexpression or knock-down of different genes in the same cell. However, these approaches also have downsides: Differences while handling or transplanting the cells or even differences between mice and irradiation can induce effects that are not attributable to the overexpressed or knocked-out gene (Deneault et al., 2009).

Considering the insights from the above mentioned studies, the results from the *in vivo* screen performed in this study have to be treated with caution. Assuming a maximum of 100 engrafted HSCs per mouse would naturally not allow for a proper BC representation, regardless of the library size. One has also to take into consideration, that only ~20% of the transplanted cells are transduced, limiting the approach even further. Some of these limitations could probably be bypassed through larger cohorts of mice or fewer genes. On the other hand, not only HSCs contribute to blood production in the first weeks after transplantation, but also progenitor cells (Lu et al., 2011). Depending on the focus of the study, changes in the progenitor compartment can also be of interest. Hence, the approach itself has the capacity to generate usable results, however, has to be further improved with respect to the above mentioned limitations. In fact, using the same approach applied in the present study but focusing on overexpression of miRNA instead, did indeed generate robust results (data from Elias Eckert, published in Wünsche et al. (2018)). A uniformly sized miRNA overexpression library with fewer constructs and higher engraftment and transduction efficiency led to the identification of miR-10a and miR-335 as regulators of early hematopoiesis. Both miRNAs were validated individually in transplantation experiments, demonstrating that the approach can be sensitive and robust enough to pick up subtle changes in BC representation (Wünsche et al., 2018).

4.1.6 γRV genotoxicity is probably less universal than expected.

Albeit the fact that the data discussed above did not allow for detailed insights on how the candidate genes affect hematopoiesis, some conclusions can still be drawn from these experiments. Throughout all experiments, we never observed clonal expansion or neoplastic growth of transduced cells. This was partially unexpected, as the field of gene therapy postulates that large CIS occur due to clonal expansion of cells (Biasco et al., 2012). Undeniably, this effect can certainly be observed for some well-characterized proto-oncogenes such as *MECOM* (*EVI1*/*MDS1*), *LMO2*, *MN1*, *SETBP1*, *PRDM16* and *CCND2* (Boztug et al., 2010; Braun et al., 2014). However, at this point, our candidate genes do not appear to fall into this category. Especially NRIP1 showed some of the largest clusters of ISs among all genes, yet its overexpression never caused an overproportional expansion of cells. Instead, the opposite was observed, ultimately leading to the question, if there are other reasons why CIS are differently sized and what the location in the genome dictates. In fact, investigations on γRV integration preference have answered many questions and propose a directed integration mechanisms, which leads to a heterogeneous distribution of ISs in the genome, even without clonal selection (Kvaratskhelia et al., 2014).

The following paragraphs of the discussion contain text sections that have been taken from Wünsche et al., (2018) and have been originally written by myself. All literal quotes are indicated by quotation marks (" … "), following the guidelines of good scientific practice of the Ruperto-Carola University of Heidelberg.
Reprinted figures from Wünsche et al., (2018) are indicated as such in the figure legend.

4.2 Oncogenic γ-retroviral integration events are less common than expected

The fact that not a single gene included in our screen led to a measurable clonal expansion neither *in vitro* nor *in vivo* experiments was rather unexpected. For many years, the general perception was that large CIS are most certainly a consequence of overproportionally expanding cells or clonal selection, respectively. Thus, genes that carry a lot of ISs were classically suspected to be causal for this overproportional expansion (Biasco et al., 2012). Yet, the complete picture appears more complex. In 1993, Stocking and colleagues tried to estimate the frequency at which transforming events occur. Using *in vitro* growth-factor dependency assays, the proposed risk was only ~2 x 10^{-7} (Stocking et al., 1993). Other studies suggested comparable rates, ranging from 10^{-8} to 10^{-6} (King et al., 1985; Moolten and Cupples, 1992). With increasing knowledge from insertional mutagenesis screens, the frequency at which ISs were thought to hit proto-oncogenes increased significantly to 10^{-5} to 10^{-2}. However, the way these numbers were generated was still very theoretical and subject to many assumptions (Baum et al., 2003). Moreover, the rate at which proto-oncogenes are hit and the rate at which clonal outgrowth or malignant transformation occurs should be very different. First of all, gene activation or disruption due to viral integrants arise monoallelic, restricting a profound effect to dominant proto-oncogenes only. Recessive oncogenes are likely to require an additional mutation in the second allele to cause a phenotype. Moreover, human carcinogenesis is usually a multistep process and requires for example additional genetic lesions, maintenance of *Telomerase Reverse Transcriptase* (*TERT)* expression to overcome oncogene-induced senescence, or immune escape (Baum et al., 2004). This is in line with "[…] a study from Howe et al. (2008), showed that clonal outgrowth often depends on combinatorial processes between somatic mutations, deletions, or translocations and retroviral overexpression, restricting the pool of transformed clones dramatically.”

The theory of multiple cooperating hits required for malignant transformation also supports our notion of rarely transforming IS. Using our IS data in comparison to other IS data, “[…] we neither observed a striking difference in the global number of clusters, cluster sizes or dimensions, nor a significant enrichment of genes contained within leukemia related gene sets.” Also, the analysis of cumulative ISs per gene only revealed 13 genes that were statistically noticeable. Moreover, “[…] despite the high prevalence of leukemia observed in the WAS GT trial, the global pattern of ISs remained stable, again indicating that γRV-mediated transformation is restricted to only few loci. In fact, across various γRV gene therapy trials all patients presented with only a small number of dominant clones, driven by the same limited set of recurrent genes (Braun et al., 2014; Hacein-Bey-Abina et al., 2003b; Ott et al., 2006).” In summary, “[…] the genotoxic potential of γRV insertions appears to take effect only under certain circumstances, whereas the vast majority of ISs leave the cells unchanged. Of course, one must keep in mind that the IS pattern might be indeed skewed for known leukemogenic drivers such as *MECOM* and *LMO2*. However, regardless of the skewing, these genes may still play an important role during normal hematopoiesis, as inactive genes are unlikely to be targeted by γRV insertions and consequently would not carry any IS. In fact, many of the known leukemogenic drivers identified in gene therapy studies and γRV screenings in mice have also been implicated in normal HSC regulation or were shown to be specifically active in HSCs (Aguilo et al., 2011; Yamada et al., 1998; Zhang et al., 2011).” Given the non-detectable incidence of clonal outgrowth upon transduction with our candidate genes and the low probability of a genome-wide skewed γRV IS pattern, we hypothesized that γRV IS from patients can be harnessed for other purposes.

4.3 Using γ-retroviral integration sites from long-term repopulating clones to map active gene-regulatory regions in HSCs

An essential assumption made in the present study is the post-transplant enrichment of HSCs. Historically, HSCs were defined as cells that have the capacity to repopulate the entire blood system of an organisms through self-renewal and differentiation into all blood-cell lineages (multipotency, see 1.1.1). Later, the discovery of surface marker combinations enabled the prospective enrichment of phenotypic HSC (see 1.1.2). Both names are often used interchangeably, although functional and phenotypic HSCs are not exactly the same. While phenotypic markers only enrich for HSCs, functional experiments can usually prove the aforementioned characteristics. Although we were able to address the self-renewal capacity of IS-tagged cells through positive association matrices and a mathematical model, our data does not provide the necessary lineage information required to analyze the multipotency. To infer that IS-tagged cells originated from HSCs using only information on self-renewal capacity (time after transplantation and recurrence of clones), it is important to discuss the current state of research regarding the definition of phenotypic and functional HSCs.

4.3.1 HSCs differ in their capacity to differentiate and self-renew

To date, countless studies have conducted functional experiments to investigate HSC biology. While HSCs were historically regarded as a homogenous population of cells that can be separated from committed downstream progenitors, results from recent experiments suggest that the transition from HSCs to MPPs to progenitors is not as sharp as previously imagined. In fact, the contrary was observed – a multitude of stages between multipotent HSCs and oligopotent or lineage restricted progenitors, respectively, ultimately questioning our current definition of HSCs (Haas et al., 2018; Laurenti and Göttgens, 2018). However, given the focus on HSCs in this study, the following paragraphs discuss the key findings regarding HSC biology and nomenclature to elaborate more on our concept of post-transplant HSC enrichment.

Many studies, both in humans and in mice could show that a substantial proportion of phenotypic and also functional HSCs are lineage biased (Haas et al., 2018; Laurenti and Göttgens, 2018). In functional experiments, this manifests through a shifted output of long-term clones towards either lymphoid or myeloid cells. Importantly, lineage-biased cells are different from lineage-restricted cells. Lineage-biased cells can still be regarded as multipotent, as they contribute to both lineages although to different quantities. Lineage-restricted cells in contrast are oligopotent and only contribute to either lymphoid or myeloid cells. One of the pioneering studies that experimentally addressed HSC heterogeneity in more detail was conducted by Dykstra et al. (2007). Here, flow cytometry sorted murine phenotypic HSCs were transplanted into sublethally irradiated mice as single cells or single cell derived clonal cultures. The output of individual cells was followed over time focusing on lineage contribution and self-renewal capacity. Interestingly, four HSC-subtypes with differing self-renewal and differentiation capacity were identified, termed α-, β-, γ-, and δ-cells. Out of these four, only β-cells showed a balanced output, while α-cells showed a strong myeloid bias and γ-cells a corresponding lymphoid bias. In contrast, δ-cells were almost completely lineage-restricted to only lymphoid cells. Importantly, the majority of HSCs were found to be β-cells (39%) and α-cells (27%), which also exhibited the highest self-renewal capacity. In contrast, γ- and δ-cells were not capable of reconstituting secondary or tertiary recipients. In conclusion, this study demonstrated the presence of lineage-restricted cells (δ-cells) within the HSC pool, however, these cells showed the lowest percent of donor blood contribution (<10%) and were not serially transplantable (Dykstra et al., 2007). In contrast, ~60% of the recipient blood was reconstituted by balanced β-cells. Similar results were generated without single-cell transplantation experiments but instead using viral genetic barcoding. Here, two different HSC sub-populations were identified, one biased towards B and T cells, and one biased towards B cells and granulocytes. Importantly, both sub-populations were biased but not restricted, indicating multipotency (Glimm et al., 2011; Lu et al., 2011).

So far, only lineage-balanced or lineage-biased cells were shown to possess long-term repopulating capacity. Lineage-restricted cells, on the other hand, which would not fulfill the HSC-criterion of multipotency only showed short-term but not long-term self-renewal capacity. Another study that demonstrated short-term repopulation capacity of lineage-restricted progenitors came from Yamamoto et al. (2013). Here, the authors describe the presence of myeloid-restricted progenitors

(MyRPs) within the pool of phenotypic HSCs. MyRPs contain common myeloid repopulating progenitor (CMRP), megakaryocyte-erythroid repopulating progenitor (MERP) and megakaryocyte repopulating progenitors (MkRP) that are derived from HSCs by asymmetric cell division. Tracking of MyRPs showed that MERPs and MkRPs only self-renew for about 20 weeks and were not serially transplantable, similar to CMRPs, which vanished latest 4 weeks after re-transplantation. Interestingly, the repopulation kinetics were very comparable to those of short-term HSCs, the equivalent to human MPPs. To date, MkRPs were also identified by many others, that show that these highly lineage-restricted stem-like cells indeed exhibit short-term self-renewal capacity and reside within the pool of phenotypic LT-HSCs (Carrelha et al., 2018; Grinenko et al., 2018; Haas et al., 2015; Sanjuan-Pla et al., 2013; Shin et al., 2014). A comprehensive study from Carrelha et al. (2018) investigated systematically cells of the megakaryocyte/platelet, erythroid, myeloid and B and T cell lineages after single cell transplantation at an unpreceded resolution. Interestingly, Carrelha and colleagues also identified a distinct class of megakaryocyte/platelet-restricted HSCs, however claim that these cells maintain their multipotency albeit lineage-restriction. Importantly, no other HSC sub-class was observed that contributed to one lineage only. Moreover, lineage-biased HSCs retained their multipotency.

In summary, these insights into the murine hematopoiesis clearly indicate that long-term self-renewal capacity almost always coincides with multipotency. Because of the tight connection between these two properties, it is safe to assume that long-term repopulating cells are equivalent to HSCs.

The tight connection between long-term self-renewal capacity and multipotency in mice is in fact also in line with insights from studies in non-human primates. Clonal tracking studies in rhesus macaque revealed that already after one month uni-lineage progenitors were replaced by myeloid, then by myeloid-B and later by stable myeloid-B-T multipotent HSCs (Wu et al., 2014). This is partially in line with another study in primates, which shows that after 7-13 months uni-lineage progenitors were replaced by long-term multipotent clones, contributing to >80% of the total blood cell population (Kim et al., 2014). Interestingly, similar observations were made in a human lentiviral gene therapy trial, which showed the greatest overall multilineage output over time by HSCs followed a lesser extent by MPPs (Biasco et al., 2016). These studies again suggest that it is safe to assume that long-term repopulating cells are equivalent to HSCs and to a lesser extent by MPPs.

While all of the aforementioned studies addressed blood reconstitution and maintenance after transplantation, new *in situ* genetic barcoding techniques also enabled researchers to study clonal dynamics in an unperturbed or naïve state (Busch et al., 2015; Rodriguez-Fraticelli et al., 2018; Sun et al., 2014; Yu et al., 2016). Intriguingly, these studies observed a much higher contribution of phenotypic MPPs to steady-state hematopoiesis than HSCs, indicating that post-transplant and naïve or unperturbed hematopoiesis differ at least in some aspects. Importantly however, HSC heterogeneity regarding lineage-bias or restriction were shown to be both features of reconstitution after transplantation as well as naïve hematopoiesis (Carrelha et al., 2018; Rodriguez-Fraticelli et al., 2018).

Taken together, the new insights into HSC heterogeneity and lineage biases are challenging the classical linear tree-model of hematopoiesis. Moreover, the presence of self-renewing lineage-restricted MkPRs within the pool of phenotypic HSCs is questioning our current definition of HSCs even more. Nevertheless, the current literature clearly shows that the capacity of clones to reconstitute an entire organism over a period of up to 6 years, at least in the context of post-transplant hematopoiesis, can be regarded as a feature that is unique to HSCs only. In other word, multipotency hence HSC properties can be inferred from the time information alone. This is of great importance, as we lack the lineage information of ISs to show multipotency of clones during steady long-term hematopoiesis by demonstrating both myeloid and lymphoid output.

4.3.2 The post-transplant enrichment of HSCs can be also be visualized using ATAC-seq signal intensity and gene set enrichment analysis

Apart from inferring HSC-specificity of ISs using only the time information, we also analyzed the ATAC-seq signal intensity at ISs from three independent studies across 13 primary human blood cell types (ISs from the WAS GT study, from Aiuti et al. (2007) and from De Ravin et al. (2014)). "We detected a striking difference between CD34$^+$ ISs and WAS ISs or pre- and post-transplantation IS pattern, respectively. This indicates that cells with long-term engraftment capabilities, thus with HSC-like chromatin structure, are selected after transplantation, while displacing short-lived progenitors. The fact that this phenomenon was observed across three separate studies clearly suggests that this selection occurs independently of disease background or transduction or cultivation protocols. Intriguingly, the enrichment of HSC-specific ATAC-seq peaks is very comparable between early and late IS, thus occurs much faster than the switch observed using pairwise positive association. However, the CD34$^+$ pool contains a high number of cells that do not engraft upon transplantation, hence ISs in these cells are lost from the pool very fast and consequently could also cause a change in the ATAC-seq signal intensity at ISs after transplantation. Secondly, the ATAC-seq data is derived from immunophenotypically defined HSCs, which are still relatively heterogeneous and moreover were almost indistinguishable from MPPs (Corces et al., 2016). In contrast, murine HSCs can be sorted to much higher purities. Recently, high resolution gene expression data of murine LT-HSCs and their immediate progenitor populations (MPP1-4) has been published (Cabezas-Wallscheid et al., 2014). Strikingly, genes that are significantly up-regulated in murine LT-HSCs compared to any of the downstream MPP populations were also significantly enriched in our human dataset, not only implying conserved functions of these genes across species, but also further supporting our hypothesis of tagging long-term specific regulatory regions."

In summary, the enrichment of HSC-specific ATAC-seq peaks and LT-HSC specific genes after transplantation strongly supports the notion that long-term engraftment naturally selects for HSCs and thus endorses our approach of using γRV ISs to map active regulatory regions in human repopulating HSCs.

4.3.3 γ-retroviral integration might also point towards super-enhancers

The puzzle of how and where γRV ISs locate in the genome has slowly been deciphered in the last decades. Researchers came a long way from the assumption that ISs spread out completely random in the genome to the concept of directed integration. Just recently, "[…] γRVs have been shown to preferentially integrate in active regulatory elements such as enhancers through tethering of the viral intasome to chromatin through the interaction with BET proteins" (Cattoglio et al., 2010; De Ravin et al., 2014; De Rijck et al., 2013; Gupta et al., 2013; Kvaratskhelia et al., 2014; LaFave et al., 2014; Larue et al., 2014; Sharma et al., 2013). Due to the stable integration, sequencing of γRV

ISs reveals the location of strong enhancers and active promoters. In fact, γRV ISs have already been used to successfully map regulatory regions in HSPCs, MPPs and EPPs to address epigenetic changes associated to HSPC lineage commitment, prior to our study (Romano et al., 2016). Interestingly, in that study Romano and colleagues found a highly significant enrichment of γRV clusters at super-enhancers (SE) compared to normal enhancers (53% vs. 12%), raising the question whether our IS clusters point more specifically to SEs rather than normal enhancers in human HSCs. However, as only one study reported this coherence so far, it is probably not yet save to generalize this observation for other cells types as well (enrichment of γRV ISs at strong enhancers and active promoters has been shown by multiple independent studies). One way to address this question could be to re-analyze existing data sets with matching samples of ChIP-seq against at least H3K4me3, H3K4me1 and H3K27ac and γRV ISs with a focus on SEs. Moreover, the ever-improving understanding of histone modifications and their associated chromatin states as well as newly available technologies might even refine the current concept of γRV integration preference, this way also improving the interpretation of the regulatory regions mapped in this study.

4.3.4 Prediction of enhancer activity during early hematopoiesis using the time point of IS detection

“In addition to the investigation of IS-tagged genes we also demonstrated the HSC-specificity of ISs on cluster level, by analyzing the recently reported blood enhancer cluster (BENC), which has been functionally dissected with enhancer sub-module resolution. Strikingly, we were not only able to recapitulate the results from Bahr et al. (2018), but also identified three additional sub-modules with equal or even higher HSC specificity. These results indicate that our IS data has sufficient resolution to identify enhancer modules and that the information about the time point of detection allows to predict the activity of these enhancers in the early hematopoietic hierarchy.”

4.4 Conclusion and perspectives

"In summary, we show that γRV ISs can be used as molecular tags not only for clonal tracking, but also to mark regulatory regions in functionally defined cell populations. Unlike sequencing approaches of phenotypically defined cell populations, our method exploits the natural selection process enriching for long-term repopulating HSCs after transplantation. This approach may be even extended to other vector types such as lentiviruses (active gene bodies) or adeno-associated viruses (AAV5: transcriptional activity) (Janovitz et al., 2014). Likewise, γRVs may also be used to tag regulatory regions in other rare cell types. Using our strategy, we were able to detect >79,000 genomic tags from 10 patients which point towards >3,000 regulatory regions in human long-term repopulating HSCs. These data provide new insights into active regions and regulatory mechanisms of repopulating HSCs and represent a solid basis and comprehensive resource for functional studies investigating stem cell self-renewal and differentiation." Although we did not observe a striking phenotype of the IS-tagged protein-coding genes analyzed with the pooled over-expression approach, future experiments with improved robustness or sensitivity might provide further insights into if and how these potential HSC regulators could affect hematopoiesis.

5 MATERIAL AND METHODS

5.1 Material

5.1.1 Technical equipment

Table 6 | **Overview of technical equipment and devices**

Instrument	Manufacturer
Agarose gel electrophoresis chambers	VWR Peqlab
Agarose gel electrophoresis power supply	Elchrom Scientific
Avanti J-30I Ultracentrifuge	Beckman Coulter
Bacteria incubator	Sanyo
Bacteria shaker	Infors
Benchtop centrifuges	Eppendorf
Benchtop centrifuges, cooling	Heraeus
Cell culture centrifuge	Heraeus
Cell culture hood	Thermo Scientific
Cell culture incubator	Thermo Scientific
Cell sorter FACS Aria II	BD Biosciences
ChemiDoc XRS Imaging System	Bio-Rad
Cobas z 480	Roche
cOmplete EDTA-free Protease Inhibitor Cocktail	Roche
Flow cytometer LSRII	BD Biosciences
Freezer -20 °C	Liebherr
Freezer -80 °C	Sanyo
Fridge 4 °C	Liebherr
Fume hood	WALDNER
Gel documentation station	VWR Peqlab
Hotplate stirrer	VWR
Ice machine	Hoshizaki
L8-55M Ultracentrifuge	Beckman Coulter
Liquid nitrogen tank	German-Cryo
Microscope Axiovert 40C	Zeiss
Microwave	Bartscher
Mr. Frosty freezing containers	Thermo Scientific
NanoDrop Spectrophotometer ND-1000	Thermo Scientific
PAGE running chambers	Bio-Rad
pH meter	Mettler Toledo

Pipetboy	Integra Biosciences
Pipettes	Eppendorf
Qubit 2.0 Fluorometer	Invitrogen
SDS-PAGE power supply	Bio-Rad
Thermocycler peqSTAR	VWR Peqlab
Thermomixer	Eppendorf
Trans-Blot Turbo Transfer System	Bio-Rad
Vacuum pump	VACUUBRAND
Vortex	IKA
Water purification system (for ddH$_2$O)	Thermo Scientific

5.1.2 Commercial kits

Table 7 | **Commercial kits**

Kit	**Manufacturer**
DNeasy Blood & Tissue Kit	Qiagen
EasySep™ Mouse Hematopoietic Progenitor Cell Isolation Kit	Stem Cell Technologies
GeneJET Plasmid Maxiprep Kit	Thermo Scientific
GeneMATRIX Plasmid Miniprep DNA Purification Kit	EURx
QIAquick Gel Purification Kit	Qiagen
QIAquick PCR Purification Kit	Qiagen
Qubit dsDNA HS Assay Kit	Thermo Scientific
RNase-Free DNase Set	Qiagen
RNeasy Micro or Mini Kit	Qiagen
SuperScript® III First-Strand Synthesis SuperMix	Invitrogen

5.1.3 Reagents

Table 8 | **Reagents**

Reagent	**Manufacturer**
10x Tris/Glycine/SDS Running Buffer	Bio-Rad
2-Mercaptoethanol	Sigma-Aldrich
ACK Lysing Buffer	Thermo Scientific
Agar	Sigma-Aldrich
Agarose for DNA Electrophoresis	Serva
Ammonium chloride	Sigma
Ammonium persulfate (APS)	Biorad

Ampicillin sodium salt	Sigma-Aldrich
Bovine Serum Albumin (BSA)	Linaris
cOmplete EDTA-free Protease Inhibitor Cocktail	Roche
Cytokines: mIl3, mFlt3-L, mTPO, mSCF	R&D systems
Dimethyl sulfoxide (DMSO)	Sigma-Aldrich
DNA Gel Loading Dye	Thermo Scientific
DNase I, RNase-frei	Epicentre
dNTPs	Genaxxon
Dulbecco's Phosphate-Buffered Saline (PBS)	Gibco
Ethanol	Sigma-Aldrich
Ethidium Bromide solution 0.07 %	AppliChem
FACS Flow Sheath Fluid	BD
Fetal bovine serum (FBS)	GE Healthcare
Fluoro-Gold™ (Hydroxystilbamidine bis(methanesulfonate))	Sigma-Aldrich
GeneRuler DNA Ladders 100 bp, 1 kb	Thermo Scientific
IGEPAL CA-630 (NP40 substitute)	Sigma-Aldrich
Iscove's Modified Dulbecco's Medium (IMDM)	Invitrogen
Laemmli Sample Buffer, 4x Concentrate	Bio-Rad
Luria Broth Base powder	Invitrogen
Magnesiumchloride, $MgCl_2$	Sigma-Aldrich
Methanol	Sigma-Aldrich
MethoCult™ M3434	Stemcell Technologies
Penicillin-Streptomycin (P/S)	Gibco
Phusion High-Fidelity DNA Polymerase	NEB
Polybrene	Chemicon
Polyethylenimine, branched	Sigma-Aldrich
Precision Plus Protein Standards (Dual Color)	BioRad
Protamine Sulfate	Sigma-Aldrich
Proteinase K	Qiagen
REDTaq® ReadyMix™ PCR Reaction Mix	Sigma Aldrich
Restriction enzymes (BamHI, SbfI, ClaI + Cut Smart buffer)	New England Biolabs
RNase A	Qiagen
RNaseOUT™ Recombinant Ribonuclease Inhibitor	Thermo Scientific
RoboSep Buffer	Stem Cell Technologies
Roswell Park Memorial Institute (RPMI)-1640 Medium	Gibco
S.O.C. Medium	Invitrogen
Sodium chloride	Sigma-Aldrich
StemSpan SFEM	Stem Cell Technologies

T4 DNA Ligase + buffer	New England Biolabs
T4 Polynucleotide Kinase + buffer	New England Biolabs
Taq DNA polymerase + PCR reaction buffer	Qiagen
TBE buffer (10X)	Genaxxon
Titanium Taq DNA polymerase + PCR reaction buffer	Takara
Trans-Blot Turbo Midi PVDF Transfer Packs	Bio-Rad
Tris powder	Bio-Rad
Trypan blue	Gibco
Trypsin-EDTA	Gibco
TWEEN 20	Sigma-Aldrich
Western Lightning Plus-ECL	PerkinElmer

5.1.4 Consumables

Table 9 | **Consumables**

Consumable	**Manufacturer**
Cell Counting Chambers	Neubauer
Cell culture dishes 10 cm, 15 cm	Corning
Cell culture flasks T25, T75, T225	Fisher Scientific
Cell culture plates 6-well, 12-well, 24-well, 48-well, 96-well	Greiner Bio-One
Cell strainer 40 µm, 70 µm, 100 µm	Corning
Conical tubes 15 mL, 50 mL	BD
Cryo tubes 2.0 mL (sterile)	Genaxxon
FACS tubes, 4.5 mL conical bottom polystyrene test tube	Greiner Bio-One
FACS tubes, 5 mL round bottom polystyrene test tube	BD
Filter foil, 85 µm, SEFAR NITEX 03-85/35	Sefar
Filter pipette tips 10 µL, 20 µL, 200 µL, 1,000 µL	Greiner Bio-One
Microvette CB 300 K2E	Sarstedt
PCR plate sealing foil	Steinbrenner
PCR reaction plate, 96-well	Greiner Bio-One
PCR strips	Biozym
Petri dishes	Corning
Pipettes 2 mL, 5 mL, 10 mL, 25 mL, 50 mL	Corning
Plastic flasks 125 mL, 250 mL	Nunc
Polyallomer Centrifuge Tubes (Ultracentrifuge)	Beckman Coulter
qPCR plate sealing foil	Biozym
qPCR reaction plate, 96-well	Biozym

Qubit Assay Tubes	Invitrogen
SafeSeal reaction tubes 0.5 mL, 1.5 mL, 2 mL	Sarstedt

5.1.5 Plasmids

Table 10 | **Plasmids**

Handling name	**Full name/Elements**	**Reference**
LV101	3rd generation lentiviral plasmid (*gag-pol*)	In house plasmid stocks
LV102	3rd generation lentiviral plasmid (*rev*)	In house plasmid stocks
LV103	3rd generation lentiviral plasmid (*vsv-g*)	In house plasmid stocks
p602	pCCL.SIN.cPPT.PGK.IRES.eGFP.wPRE	Luigi Naldini, (Herbst et al., 2012)
p612	pCCL.SIN.cPPT.pTight.IRES.eGFP.wPRE	Cloned using p602 backbone with pTight (pLVX) promoter
P902	pCCL.SIN.cPPT.BC.pTight.IRES.eGFP.wPRE	Cloned from p612 backbone with Barcode cassette (BC)
pMA-RQ	pMA-RQ transfer vector with Ampicillin resistance	Invitrogen (GeneArt)
pMA-T	pMA-T transfer vector with Ampicillin resistance	Invitrogen (GeneArt)
pMK-RQ	pMA-RQ transfer vector with Kanamycin resistance	Invitrogen (GeneArt)
pRSI9	pRSI9-U6-(sh)-HTS3-UbiC-TagRFP-2A-Puro	Cellecta, Inc.

P612 plasmid was originally cloned by Shayda Hemmati.

5.1.6 Western blot Antibodies

Table 11 | **Westernblot Antibodies**

Antibody	**Host**	**Supplier**	**Cat. no.**	**Dilution**
Anti Igf2bp2	Rabbit	Antibodies-Online	ABIN502002	1:1,000
Anti-rabbit-HRP	Goat	Abcam	ab6721	1:10,000

5.1.7 Flow cytometry antibodies and staining panels

Table 12 | **Antibodies used for sorting of LSK and LSK-SLAM cells**

Antibody	Format	Clone	Host	Supplier	Cat. no.	Dilution
CD117	PE	2B8	Rat IgG2b	BD	553355	1:200
CD150	PE-Cy5	17A2	Rat IgG2b	BD	555276	1:500
CD48	AlexaFlour700	HM48-1	Armenian Hamster	Biozol	B188338	1:200
Lineage Cocktail	APC	mix	Isotype Cocktail	BD	558074	1:100
Ly-6A/E (Sca-1)	PE-Cy7	D7	Rat IgG2a	BD	558162	1:200

Establishment of color combinations and antibody dilutions for the detection of indicated cell populations were jointly established with Elias Eckert.

Table 13 | **Antibodies used for sorting of HSCs**

Antibody	Format	Clone	Host	Supplier	Cat. no.	Dilution
CD117	APC	2B8	Rat IgG2b	BD	553991	1:200
CD135	PE	A2F10.1	Rat IgG2a	BD	553930	1:100
CD150	PE-Cy5	17A2	Rat IgG2b	BD	555276	1:500
CD34	FITC	RAM34	Rat IgG2b	BD	560238	1:30
CD48	AlexaFlour700	HM48-1	Armenian Hamster	Biozol	B188338	1:400
Lineage Cocktail	PE-Cy7	mix	Isotype Cocktail	See below	See below	1:00
Ly-6A/E (Sca-1)	APC-Cy7	D7	Rat IgG2a	BD	552770	1:200

Establishment of color combinations and antibody dilutions for the detection of indicated cell populations were jointly established with Elias Eckert.

Table 14 | **Lineage cocktail used for the HSC sort staining**

Antibody	Format	Clone	Host	Supplier	Cat. no.	Dilution
CD11b	PE-Cy7	M1/70	Rat IgG2b	BD	552850	-
CD3	PE-Cy7	17A2	Rat IgG2b	BD	552849	-
CD45R	PE-Cy7	RA3-6B2	Rat IgG2a	BD	552772	-
Ly6G/C (Gr-1)	PE-Cy7	RB6-8C5	Rat IgG2b	BD	552894	-
Ter119	PE-Cy7	Ter-119	Rat IgG2b	BD	553673	-

Antibodies in Table 14 are mixed to equal proportions and used in at a final dilution of 1:400. Establishment of color combinations and antibody dilutions for the detection of indicated cell populations were jointly established with Elias Eckert.

Table 15 | **Antibodies used for detection or sort of progenitor cells**

Antibody	Format	Clone	Host	Supplier	Cat. no.	Dilution
CD117	APC	2B8	Rat IgG2b	BD	553991	1:200
CD127 (IL7R)	PE-Cy5	A7R34	Rat IgG2a	eBioscience	15-1271-82	1:200
CD16/32 (FcγR)	PE	93	Rat IgG2a	BioLegend	101308	1:200
CD34	FITC	RAM34	Rat IgG2b	BD	560238	1:30
Lineage Cocktail	PE-Cy7	mix	Isotype Cocktail	See below	See below	1:300
Ly-6A/E (Sca-1)	APC-Cy7	D7	Rat IgG2a	BD	552770	1:200

Establishment of color combinations and antibody dilutions for the detection of indicated cell populations were jointly established with Elias Eckert.

Table 16 | **Antibodies used for analysis of HSCs with presents of GFP$^+$ cells**

Antibody	Format	Clone	Host	Supplier	Cat. no.	Dilution
CD117	PE	2B8	Rat IgG2b	BD	553355	1:200
CD150	PE-Cy5	17A2	Rat IgG2b	BD	555276	1:500
CD34	AlexaFlour700	RAM34	Rat IgG2a	BD	560518	1:30
CD45.2	PacificBlue	104	Mouse IgG2a	BioLegend	109819	1:100
CD48	PE-Cy7	HM48-1	Armenian Hamster	BD	560731	1:200
Lineage Cocktail	APC	mix	Isotype Cocktail	BD	558074	1:100
Ly-6A/E (Sca-1)	APC-Cy7	D7	Rat IgG2a	BD	552770	1:200

Establishment of color combinations and antibody dilutions for the detection of indicated cell populations were jointly established with Elias Eckert.

Table 17 | **Antibodies used for analysis and sort of differentiated lymphoid and myeloid cells**

Antibody	Format	Clone	Host	Supplier	Cat. no.	Dilution
CD11b	PerCP-Cy 5.5	M1/70	Rat IgG2a	BD	550764	1:200
CD3	PE-Cy7	17A2	Rat IgG2b	BD	560591	1:200
CD45.1	PE	A20	Rat IgG2a	BD	553930	1:200
CD45.2	APC-Cy7	04	Rat IgG2a	BD	550882	1:200
CD45R	AlexaFluor700	RA3-6B2	Rat IgG2a	BD	557957	1:200
Ly6G	APC	1A8	Rat IgG2a	BD	560599	1:200

Establishment of color combinations and antibody dilutions for the detection of indicated cell populations were jointly established with Elias Eckert.

5.1.8 Oligonucleotides

Table 18 | **Genotyping primers for B6.Cg-Gt(ROSA)26Sor tm1(rtTA*M2)Jae/J mice**

Name	Sequence (5’-3’)
Rosa A	aaagtcgctctgagttgttat
Rosa B	gcgaagagtttgtcctcaacc
Rosa C	ggagcgggagaaatggatatg

Amplicon lengths: Wildtype ~330bp; Mutant ~550bp.

Table 19 | **Primers for amplifying the barcode cassette from the pRSI9 cellecta library and introducing the ClaI restriction enzyme cutsite**

Name	Sequence (5’-3’)
PW_C_Bar_Cla1_f	ttacagatcgatttttttggcaagcaaaagacg
PW_C_Bar_Cla1_r	atctatatcgattgccatttgtctcgaggtcg

Table 20 | **qPCR primers for endogenous expression of candidate genes**

Gene	Primer name	Forward primer (5'-3')	Reverse primer (5'-3')	Amplicon
Amica1	mAmica1_a	atgaaaaagcccgtggaact	gttgtatcacctactcggactctg	74
Amica1	mAmica1_b	agcctggagaacaaagagaagat	ctctgtcgtctcccacgtagt	75
Ccnd3	mCcnd3_a	ggcatactggatgctggag	ccaggtagttcatagccagagg	77
Ccnd3	mCcnd3_b	attgagaagctttgcatctatacg	gaccagcacctcccactc	72
Evl*	Evl	atgagtgaacagagtatctgcc	tctttgccacagacggggtt	-
Fbxl18	mFbxl18_a	tgtacatgcctgctcttgct	aagtagggctgctccaacc	76
Fbxl18	mFbxl18_b	ggctagctccggagagga	tcatcggagaagccaagc	87
Igf2bp2	mIgf2bp2_a	gctggtgcctccatcaag	tgaccatcctctcactgacatc	61
Igf2bp2	mIgf2bp2_b	tgacaagagaagaggcaaagc	catcggggatgtaggaaatc	90
Irf2bpl*	Irf2bpl	agatgctagctgtcccatgc	tgttcctcaccgagcttcag	-
Lair1	mLair1c	aatctagctactaatggcctggag	ttgaaggtctcctgcaactg	108
Lair1L	mLair1l_a	ggtgatcaaagaaaatgtcatcc	gctgtatgtctttagccaagatgtat	76
Lair1L	mLair1l_b	gtgcctgggatggaaaatta	tcataagacttgaattagggaagatg	77
Lair1S	mLair1s	tcatccagttatcctgctggt	gccaagatgtatcctcctgtg	74
Mbnl1	mMbnl1_a	aacatctgccacaagtgttcc	tgttcggcagatattatgggta	72
Mbnl1	mMbnl1_b	ttgattcagcagaagaacatgg	ggtgcaactgaaaacattgg	107
Ninj2	mNinj2_a	caggacctccagcaatccta	acaaaggctgaagtggctcta	74
Ninj2	mNinj2_b	ccctagtcaccctcatcattg	tggcagcattgttgagcttat	132
Nrip1	mNrip1_a	gcttttcaacagccttctcag	tcatctttcgttgctcacca	97
Nrip1	mNrip1_b	cctttaacattcgggaggaa	ggctgttgaaaagcaactctg	103
Plcb4*	Plcb4	atgcgggtaccttctcaagc	tttccgtatggtgtcggtgg	-
Prkcb	mPrkcb_a	gggatgaaatgcgacacct	cgttccgtgtggtcagtg	89
Prkcb	mPrkcb_b	gaaactcgaacgcaaggaga	accggtcgaagttttcagc	77
Slx4ip	mSlx4ipc	gaggaacgctctgaaggaaa	cactagatcttcccacgaggtc	98
Slx4ipSL	mSlx4ipsl	attgccacaaggttcaaaca	tgtgatctgaaagccataacctc	75
Slx4ipSL	mSlx4ipsl	attgccacaaggttcaaaca	tgtgatctgaaagccataacctc	113
Swap70	mSwap70_a	acctttgaaatcagtgcctca	tgcccagcttcaacagatg	88
Swap70	mSwap70_b	cggcaggatgaagagactg	ccagctctgccctcttagaa	76
Xbp1	mXbp1_a	ctgacgaggttccagaggtg	gcagaggtgcacatagtctgag	96
Xbp1	mXbp1_b	agcaagtggtggatttggaa	ccgtgagttttctcccgtaa	76
Znf217*	Znf217	tgaggatggactccctgacg	gctgcggcatactcacagaa	-

Primers/Genes marked with an asterisk were originally designed by Shayda Hemmati. **m**, amplifies murine gene; **L**, long isoform; **S**, short isoform; **l**, primer that amplifies long isoform; **s**, primer that amplifies short isoform; **c**, primer that amplifies both isoforms ("common").

Table 21 | **qPCR primers for expression of codon optimized candidate genes**

Gene	Primer name	Forward primer (5'-3')	Reverse primer (5'-3')	Amplicon
Amica1	mAmica1co	cgtgaccaaagtgaactgga	gtcgtagctcagcacggttt	70
Ccnd3	mCcnd3co	gaaagctgaagtgggacctg	cacgagggcctgtctgtc	98
Evl*	Evl_opt	atctaccacaacaccgccag	aggtggtggcttcctctttg	-
Fbxl18	mFbxlco_a	cctagctacggcgtggtg	tggttctgtccaggatctca	76
Fbxl18	mFbxlco_b	ccctagagccgatagagcac	gcactttcttgccgaagc	71
Igf2bp2	mIgf2co_a	gccctcctcacagagctagag	agtgggaagtcgatctgtctg	76
Igf2bp2	mIgf2co_b	ctgtacccccaccaccact	ttgggatgaacagagacacg	82
Irf2bpl*	Irf2bpl_opt	aaacagagccgaggaatggg	gccggtgggatactcgatg	-
Lair1L	mLair1lco	gcccgacatcaccatctt	cgctgtagctgcacacga	80
Lair1S	mLair1sco	tcaacacccaggaagatacca	cctctgctgctgtctcttgtt	164
Lrrc33*	Lrrc33_opt	ccgacaacagactgagcgag	tcgaagatgctgtcgtccag	-
Mbnl1	mMbnl1co	ctgccttcaacccttacctg	gattgcctgtcacgagcat	90
Ninj2	mNinj2co	actacaccaccctcgtgacc	caggttcaggatggcgataa	92
Nrip1_a	mNrip1co_a	aggaaaacggccagaaagac	tagcctgtccgttcaggtg	75
Nrip1_b	mNrip1co_b	tgaacagccaccagaaagtg		
Nrip1_c	mNrip1co_c	tcaggacttcagcttcagca	tgtgggacttgtcctgctc	95
Plcb4*	Plcb4_opt	gaagtccgagggcaaagagg	caccatgtaggtgaagccga	-
Prkcb_a	mPrkcbco_a	agggcgagtacttcaacgtg	ccgatcttggctctctcg	88
Prkcb_b	mPrkcbco_b	gcaagtgggcagattcaaag	ccctgtagatgatgcccttg	98
Slx4ipL	mSlx4co_l	ccgcgtgaaagaatacgtg	gctgcttctggtgaactcg	70
Slx4ipS	mSlx4co_s	ggttcagcgagcagaaaaa	gttctggacacggtgaaggt	88
Swap70	mSwap70co	aggcggaaagagctgagaa	ctgctgtttgttctcgttgg	96
Xbp1	mXbp1co	cagaacatcttcccttggaca	gtgtccagctggtccagaa	88
Xbp1S	mXbp1Sco	tcatcgtgtccgtgaagaaa	ctcaggcagtgggagctg	91
Znf217*	Znf217_opt	accccgaagtgctgatgatg	acttgctgtgagggctgaaa	-

Primers/Genes marked with an asterisk were originally designed by Shayda Hemmati. **m**, amplifies murine gene; **L**, long isoform; **S**, short isoform; **l**, primer that amplifies long isoform; **s**, primer that amplifies short isoform; **c**, primer that amplifies both isoforms ("common").

Table 22 | **Barcoding sequencing primers**

Name	**Sequence (5’-3’)**
deltaP5_1stPCR_R	gagaggttcagagttctacagtccgaaac
Multiplex Primer	aatgatacggcgaccaccgaggatcggaagagcacacgtctgaactccagtcac $(N)_8$ gagaggttcagagttctacagtccg
pRSI9_FwdGex	caagcagaagacggcatacgaga
pRSI9_FwdHTS	ttctctggcaagcaaaagacggcata
pRSI9_GexSeqN	acagtccgaaaccccaaacgcacgaa

The multiplex primer contains an eight nucleotide index sequence, indicated by $(N)_8$. The index sequence is listed below in Table 23. Primers were jointly designed with Elias Eckert.

Table 23 | **High throughput multiplexing primers**

Name	Index	Name	Index	Name	Index
NuGene_Ind_A01	TAGACGTG	NuGene_Ind_A05	ATCGCCAT	NuGene_Ind_A09	GCTTCTTG
NuGene_Ind_B01	CACTAGCT	NuGene_Ind_B05	AAGGCGTT	NuGene_Ind_B09	CTCATCAG
NuGene_Ind_C01	GCGATAGT	NuGene_Ind_C05	CACCTTAC	NuGene_Ind_C09	TGTTCGAG
NuGene_Ind_D01	TGATACGC	NuGene_Ind_D05	AGTCGACA	NuGene_Ind_D09	CTTGTCGA
NuGene_Ind_E01	TGGAGAGT	NuGene_Ind_E05	CTCAGAGT	NuGene_Ind_E09	GATGCACT
NuGene_Ind_F01	AATGGACG	NuGene_Ind_F05	ACTCCATC	NuGene_Ind_F09	TGTAGCCA
NuGene_Ind_G01	TTACGGCT	NuGene_Ind_G05	TGAGCTAG	NuGene_Ind_G09	TTGTGTGC
NuGene_Ind_H01	CTCTACTC	NuGene_Ind_H05	TGGTACAG	NuGene_Ind_H09	GACTATGC
NuGene_Ind_A02	AACGACGT	NuGene_Ind_A06	TTGACAGG	NuGene_Ind_A10	AGCGTGTT
NuGene_Ind_B02	AACAGGAC	NuGene_Ind_B06	ATACGACC	NuGene_Ind_B10	TCCGTGAA
NuGene_Ind_C02	AGGCTTCT	NuGene_Ind_C06	TATCAGCG	NuGene_Ind_C10	TCACAGCA
NuGene_Ind_D02	GGATCTTC	NuGene_Ind_D06	GGAAGCTA	NuGene_Ind_D10	ATTCGAGG
NuGene_Ind_E02	CTCAGCTA	NuGene_Ind_E06	ACGACTTG	NuGene_Ind_E10	AAGCCACA
NuGene_Ind_F02	TTGGACGT	NuGene_Ind_F06	GATGAGAC	NuGene_Ind_F10	TACCACAG
NuGene_Ind_G02	GATGTGTG	NuGene_Ind_G06	TGCTTGGT	NuGene_Ind_G10	TCGAGTGA
NuGene_Ind_H02	TTGATCCG	NuGene_Ind_H06	ACCTGACT	NuGene_Ind_H10	GTAGGAGT
NuGene_Ind_A03	AAGGCTGA	NuGene_Ind_A07	TCGCGATA	NuGene_Ind_A11	TGTTGTGG
NuGene_Ind_B03	AGAGCCTT	NuGene_Ind_B07	TCCTGCTA	NuGene_Ind_B11	TTAAGCGG
NuGene_Ind_C03	ACGGAACA	NuGene_Ind_C07	GTCCTTCT	NuGene_Ind_C11	CATACCAC
NuGene_Ind_D03	GACATTCC	NuGene_Ind_D07	ACAGCTCA	NuGene_Ind_D11	TGTACACC
NuGene_Ind_E03	CTTGGATG	NuGene_Ind_E07	TCACTCTG	NuGene_Ind_E11	CTCCTAGA
NuGene_Ind_F03	CTGTTGAC	NuGene_Ind_F07	AGGAACCT	NuGene_Ind_F11	TGCTTCCA
NuGene_Ind_G03	GATAGCGA	NuGene_Ind_G07	CAAGGTCT	NuGene_Ind_G11	GTGGTGTT
NuGene_Ind_H03	GATAGGCT	NuGene_Ind_H07	GAAGGAAG	NuGene_Ind_H11	TCCGTATG
NuGene_Ind_A04	GAATCCGA	NuGene_Ind_A08	TGAACCTG	NuGene_Ind_A12	TCTGAGAG
NuGene_Ind_B04	ATACTCCG	NuGene_Ind_B08	ACACCAGT	NuGene_Ind_B12	ACCAGCTT
NuGene_Ind_C04	GAAGGTTC	NuGene_Ind_C08	GTGAATCC	NuGene_Ind_C12	AAGGACAC
NuGene_Ind_D04	CATCCTCT	NuGene_Ind_D08	GAATCGTG	NuGene_Ind_D12	ACAGACCT
NuGene_Ind_E04	AGAAGCGT	NuGene_Ind_E08	GCATGTCT	NuGene_Ind_E12	GACGAATG
NuGene_Ind_F04	TTCCTGTG	NuGene_Ind_F08	ACTGTGTC	NuGene_Ind_F12	TTGGTGAG
NuGene_Ind_G04	GGATTCGT	NuGene_Ind_G08	TCAACTGG	NuGene_Ind_G12	TCGTAGTC
NuGene_Ind_H04	ATGGAAGG	NuGene_Ind_H08	GATCCATG	NuGene_Ind_H12	CTGCGTAT

Multiplexing strategy and primer design was jointly established with Elias Eckert.

Table 24 | **Cellecta barcode sequences**

#	Gene	18nt-BC Sequence	#	Gene	18nt-BC Sequence
1	miR_10a_oe_1	ACACACACTGGTCATGCA	49	Mbnl1_1	TGACGTTGGTACCAGTCA
2	miR_10a_oe_2	GTGTCAACGTGTTGGTCA	50	Mbnl1_2	GTCATGACCAACGTGTGT
3	miR_26a-1_oe_1	ACTGACGTACGTTGCATG	51	Ninj2_1	CAACACGTCAGTACCATG
4	miR_26a-1_oe_2	GTTGCACACAGTACCATG	52	Ninj2_2	GTACACACTCACCAGTTG
5	miR_101_oe_1	ACGTACTGTGACACCAAC	53	Nrip1_1	TGTGGTGTTGCAACGTGT
6	miR_101_oe_2	TGACTGTGCAGTACCATG	54	Nrip1_2	TGCACATGGTCAACCAAC
7	miR_146a_oe_1	ACTGACTGGTACACGTAC	55	Prkcb_1	ACTGACCATGCAGTGTGT
8	miR_146a_oe_2	ACACTGACTGGTACGTCA	56	Prkcb_2	TGTGGTCATGTGACCAGT
9	miR_148b_oe_1	TGGTGTTGGTTGACGTTG	57	Slx4ipL_1	ACACGTGTGTGTTGCATG
10	miR_148b_oe_2	GTCAGTGTCAACGTCACA	58	Slx4ipL_2	GTACTGACGTACCAGTCA
11	miR_326_oe_1	ACACGTCAACGTGTCAGT	59	Slx4ipS_1	GTACACTGTGACACCATG
12	miR_326_oe_2	ACGTTGCAGTACGTGTTG	60	Slx4ipS_2	GTCATGCACATGACGTCA
13	miR_335_oe_1	CACATGGTACACCAGTCA	61	Swap70_1	GTCACATGTGTGTGCATG
14	miR_335_oe_2	TGTGTGCAGTGTACCAGT	62	Swap70_2	GTGTGTTGTGTGCATGAC
15	miR_342_oe_1	ACCATGCAGTACACGTAC	63	Xbp1_1	GTCATGGTCAGTCAGTAC
16	miR_342_oe_2	ACGTGTACTGCAACGATG	64	Xbp1_2	TGGTGTTGTGACACCACA
17	miR_eGFP_1	GTCACATGGTTGTGCACA	65	Xbp1S_1	GTACTGACACACTGGTGT
18	miR_eGFP_2	CACACAACCAACACGTTT	66	Xbp1S_2	GTGTGTACACCAACCAAC
19	miR_eGFP_3	ACGTCATGCACATGGTCA	67	Evl_1	CATGTGGTACCATGCACA
20	miR_eGFP_4	GTGTCATGGTTGACCACA	68	Evl_2	GTGTCAGTACTGACCAGT
21	miR_eGFP_5	TGGTTGACTGCATGCACA	69	Irf2bp2_1	ACGTCACATGTGCAGTTG
22	miR_eGFP_6	TGCAACACCATGCATGCA	70	Irf2bp2_2	GTGTCACATGACTGCAGT
23	miR_eGFP_7	ACTGACTGGTGTCAGTCA	71	Lrrc33_1	ACTGCAACTGGTTGCAGT
24	miR_eGFP_8	ACACTGTGTGACCAGTTG	72	Lrrc33_2	ACCAGTGTTGCACATGGT
25	miR_eGFP_9	CATGACTGTGTGCATGAC	73	Plcb4_1	TGGTTGTGCACACATGGT
26	miR_eGFP_10	TGACGTCATGGTACCAGT	74	Plcb4_2	ACCATGTGGTTGCATGTG
27	miR_eGFP_11	ACCAGTTGTGTGTGGTGT	75	Znf217_1	ACACCATGACTGCAGTCA
28	miR_eGFP_12	ACACTGCAGTTGCATGTG	76	Znf217_2	GTGTCACATGACGTCATG
29	miR_eGFP_13	TGTGCACAGTGTTGGTGT	77	GFPctrl_1	GTCACATGACGTACCAGT
30	miR_eGFP_14	ACACCAGTGTGTCATGCA	78	GFPctrl_2	TGCAACACTGACTGGTTG
31	miR_eGFP_15	TGACACGTGTTGTGGTGT	79	GFPctrl_3	GTGTTGACGTGTGTGTAC
32	miR_eGFP_16	CAGTTTTGCATGACGTGT	80	GFPctrl_4	TGTGGTGTTGTGCATGCA
33	miR_eGFP_17	ACCAACCAGTCAGTGTGT	81	GFPctrl_5	TGACCACATGCAACCAGT
34	miR_eGFP_18	GTCATGGTGTCAACCAGT	82	GFPctrl_6	ACGTCATGCAACACGTTG
35	add_BC_1	GTACGTGTACGTTGGTAC	83	GFPctrl_7	GTACTGGTGTACTGCACA

36	add_BC_2	TGCATGGTCAGTACCAGT	84	GFPctrl_8	ACTGCATGTGACCAGTAC
37	Amica1_1	TGTGCACAGTTGACCACA	85	GFPctrl_9	CAACTGACACGTTGCAAC
38	Amica1_2	CATGACGTTGCAACCAGT	86	GFPctrl_10	ACTGACACTGACGTGTTG
39	Ccnd3_1	GTCATGACGTGTTGGTGT	87	add _BC_3	GTACTGTGGTTGACGTGT
40	Ccnd3_2	TGCAGTCAGTCAGTGTTG	88	add _BC_4	TGTGTGCATGTGTGCAGT
41	Fbxl18_1	TGGTTGGTGTGTTGCATG	89	add _BC_5	TGTGGTTGACCAACCAAC
42	Fbxl18_2	TGTGTGTGCATGTCCATG	90	add _BC_6	TGGTGTGTTGTGACCAGT
43	Igf2bp2_1	ACGTACACGTTGACGTCA	91	add _BC_7	ACCATGTGACTGACCAAC
44	Igf2bp2_2	TGACACACTGGTTGCATG	92	add _BC_8	ACCATGGTGTACACCAAC
45	Lair1L_1	TGCAGTTGACGTACCATG	93	add _BC_9	GTGTACACACACGTCATG
46	Lair1L_2	ACACTGACTGACTGCATG	94	add _BC_10	CATGTGCATGGTTGGTCA
47	Lair1S_1	TGGTTGTGTGACCAGTAC	95	add _BC_11	CAGTACTGGTACACGTCA
48	Lair1S_2	GTCAACACTGCACAGTAC	96	add _BC_12	GTACGTACCATGACCAGT

Barcode sequences were randomly picked from the Cellecta libray through bacterial transformation followed by colony PCR. Barcodes starting with "miR_" were cloned by Elias Eckert. Barcodes starting with "add_" were jointly cloned with Elias Eckert.

5.1.9 Mouse Strains

Table 25 | **Mouse strains**

Abbreviation	Full name	Allogenic marker	Supplier
BoyJ	B6.SJL-Ptprca-Pep3b-/BoyJ	CD45.1	Charles River Italia
Bl6$^{Rosa\ rtTA}$	B6.Cg-Gt(ROSA)26Sor tm1(rtTA*M2)Jae/J	CD45.2	In house breeding

5.1.10 Software

Table 26 | **Software**

Software	Company	Application
FACS Diva V8	BD	FACS analysis
FlowJo V10	FlowJo	FACS analysis
Lasergene V12.2	DNASTAR	Cloning and sequencing analysis
LightCycler 480 Software V1.5	Roche	qRT-PCR analysis
R version 3.4.2	R Foundation	Statistical analysis
RStudio Desktop 1.0.143	RStudio	Statistical analysis

Table 27 | **R packages and Unix programs**

Name	**Version**	**OS**	**Link**
bigWigToBedGraph	-	U	http://hgdownload.soe.ucsc.edu/admin/exe/linux.x86_64/bigWigToBedGraph
ClusteredMutations	1.0.1	W	https://cran.r-project.org/web/packages/ClusteredMutations/index.html
ComplexHeatmap	1.14.0	W	https://bioconductor.org/packages/release/bioc/html/ComplexHeatmap.html
data.table	1.10.4	W	https://cran.r-project.org/web/packages/data.table/index.html
dplyr	0.7.1	W	https://cran.r-project.org/web/packages/dplyr/index.html
GenomicRanges	1.28.3	W	http://bioconductor.org/packages/release/bioc/html/GenomicRanges.html
ggplot2	2.2.1	W	https://cran.r-project.org/web/packages/ggplot2/index.html
ggrepel	0.6.5	W	https://cran.r-project.org/web/packages/ggrepel/index.html
ggsignif	0.3.0	W	https://cran.r-project.org/web/packages/ggsignif/index.html
gplots	3.0.1	W	https://cran.r-project.org/web/packages/gplots/index.html
HOMER	4.9	U/O	http://homer.ucsd.edu/homer/index.html
MACS2	2.2.1	U	https://github.com/taoliu/MACS
piano	1.16.1	W	https://bioconductor.org/packages/release/bioc/html/piano.html
scales	0.4.1	W	https://cran.r-project.org/web/packages/scales/index.html
seriation	1.2-2	W	https://cran.r-project.org/web/packages/seriation/index.html
wigToBigWig	-	U	http://hgdownload.soe.ucsc.edu/admin/exe/linux.x86_64/wigToBigWig

OS, Operating System; **W**, Windows 7; **U**, Unix/Linux; **O**, OS X Yosemity; Note: MACS2 was used python version 2.7

5.2 Methods

5.2.1 Molecular cell biology

5.2.1.1 RNA isolation

Isolation of total cellular RNA was performed using either the RNeasy Mini or Micro Kit (Qiagen) according to the manufacturer protocol for RLT-Buffer lysis and homogenation with the Qiagen Shredder Columns. RNA was either directly reversely transcribed (5.2.1.4) or stored at -80°C until further processing.

5.2.1.2 DNA isolation for mouse genotyping

Genomic DNA was isolated from mouse tails using the DNeasy Blood & Tissue Kit (Qiagen) following the manufacturers protocol for rodent tails. DNA was either directly used for the genotyping PCR (5.2.1.3) or stored at -20°C until further usage.

5.2.1.3 Genotyping PCR

For genotyping of B6.Cg-Gt(ROSA)26Sor tm1(rtTA*M2)Jae/J mice, DNA was isolated as described in 5.2.1.2. For the PCR reaction, 5uL of DNA with a concentration of 5-100ng/µL was used. The Primer-Mix was produced using 50 µL of Rosa A, 35µL of Rosa B and 50µL of Rosa C primers (Table 18) and mixed with 1215µL of distilled water. PCR mix and amplification conditions are listed below.

Table 28 | **Master Mix composition and PCR conditions for genotyping PCR of B6.Cg-Gt(ROSA)26Sor tm1(rtTA*M2)Jae/J mice**

Reagent	Volume [µL]
Template (DNA)	5
H2O	3.95
Red-Tag-Mix	11
$MgCl_2$	0,7
Primer-Mix (see above)	1.35

Step	Temp [°C]	time [s]	#
Initial denaturing	95	45	1
Denaturing	95	45	2
Annealing	55	45	3
Elongation	72	60	4
---------- Repeat #2-4 30x ----------			
Final Elongation	72	420	5
Storage	4	Infinite	6

5.2.1.4 Reverse transcription and Quantitative real-time PCR (qRT-PCR)

mRNAs were reverse transcribed into cDNA using the SuperScript® III First-Strand Synthesis SuperMix (Invitrogen) with the addition of RNaseOUT™ Recombinant Ribonuclease Inhibitor following the manufacturers protocol. Next, cDNA was diluted 1:10 in case 1µg of total RNA was used and 1:20 when 2µg of RNA was used during reverse transcription. Finally, a master mix was prepared and aliquoted into wells of a 96-well plate followed by dispensing cDNA into the well. All primers were design to anneal at 60°C, in order to facilitate to grouping of different amplifications on the same 96-well plate.

Table 29 | **Master Mix composition and PCR conditions for qPCR reactions**

Reagent	Volume [µL]	Step	Temp [°C]	time [s]	#
Template (cDNA)	2	Initial denaturing	95	600	1
H2O	7	Denaturing	95	15	2
SYBR green	10	Annealing	60	30	3
Forward Primer	0.5	Elongation	72	30	4
Reverse Primer	0.5	---------- Repeat #2-4 40x -----------			
		Storage	RT	Infinite	6

5.2.1.5 Protein isolation and semi dry Western blot for the detection of Igf2bp2

1 x 10^7 cells were lyzed in 200µL RIPA buffer (conditioned with 1 cOmplete™ ULTRA Tablets, Mini, EASYpack Protease Inhibitor Cocktail tablet per 10mL of RIPA buffer) for 30 minutes at 4°C while shaking. Next, suspension is centrifuged at 13,000rpm and 4°C in a benchtop centrifuge for 15min and supernatant is transferred to a fresh tube. After protein isolation, concentration is measured using the Pierce™ BCA Protein Assay Kit according to the manufacturers' protocol. For Igf2bp2, a 5% resolving gel and a 10% stacking gel was casted according to Table 30 and Table 31. 10µg of total protein was mixed with 4x Laemmli-buffer and incubated at 95°C for 5min, followed by centrifugation at 4°C and maximum speed and incubation on ice before loading. Denatured protein was loaded and separated at 120V for 1.5h in running buffer. Gel was blotted onto a PVDF, which was active for 1 minute in 100% methanol, for 1.5h at 25V. Finally, membrane was blocked in 5% low fat milk powder dissolved in TBS-T for 60min

at RT before incubation with primary antibody against mouse Igf2bp2 over night at 4°C. To this end, the antibody was diluted 1:1,000 in 5% low fat milk powder dissolved in TBS-T and 0.01% NaN_3.

After incubation with the primary antibody, membrane was washed thoroughly in TBS-T, followed by incubation with secondary antibody coupled to horse-reddish peroxidase, raised against host species of first antibody at a dilution of 1:10,000 for 1 hour at room temperatur. Finally, membrane was washed and secondary antibody was detected by incubation with Western Lightning Plus-ECL for 1min using the Gel documentation station.

Table 30 | **Reagents and concentrations for the production blotting buffer as well as the resolving and stacking gel for Western blot**

Puffer	Reagent	Final concentration
RIPA buffer	NaCl	150mM
	Tris	50mM
	Nonident P-40	1%
	Na-Deoxycholate	0.5%
	SDS	1%
Blotting buffer	Tris	25mM
	Glycin	150mM
	(methanol)	10%
	(20% SDS)	1%
10x PBS-T-buffer	KCl	27mM
	NaCl	1,37M
	Tween 20	0,5%
	Na_2HPO_4	100mM
	KH_2PO_4	20mM
2x Loading Dye	20% SDS	4%
	Glycerol	20%
	Beta-Mercaptoethanol	10%
	Bromphenolblau	0,004%
	Tris	125mM
10x TBS-T-buffer	1M Tris/HCL pH 7,5	100mM
	NaCl	1,5M
	Tween 20	0,5%
LGP (lower gel buffer):	Tris	1,5M
	20% SDS	0,4%
UGP (gel buffer):upper	Tris	0,5M
	20% SDS	0,4%
10xSDS Running buffer	Tris	25mM
	Glycin	190mM
	20% SDS	1%
10% APS	1g per 10mL	10%

Table 31 | **Volumes used for resolving and stacking gel**

1x resolving gel	**5%**	**7,5%**	**10%**	**13%**	**15%**	**18%**	**1x stacking gel**	**5%**
LGB (Lower Gel Buffer) in mL	2	2	**2**	2	2	2	UGB (Upper Gel Buffer) in mL	1,2
H2O in mL	4,6	4	**3,3**	2,5	2	1,2	H2O in mL	3
30% Acrylamid in mL	1,3	2	**2,6**	3,4	4	4,8	30% Acrylamid in mL	0,8
10% APS in µL	50	50	**50**	50	50	50	10% APS in µL	50
TEMED in µL	5	5	**5**	5	5	5	TEMED in µL	5

For the western blot of Igf2bp2, a 10% resolving gel was used.

5.2.1.6 Cloning of the barcode cassette containing p902 target vector

The p902 target vector was cloned from the p612 vector by inserting the barcode cassettes upstream of the pLVX promoter. Barcode cassettes were amplified from the Cellecta pRSI6 shRNA library using the PW_C_BAR_Cla1_f/r primers, followed by purification with the PCR purification kit (Qiagen). Next, PCR products were digested with ClaI and ligated into ClaI-linearized p612 vectors. Finally, ligation product was transformed and spread on LB-Agar plate to pick single colonies. (Note that every colony represents a unique barcode). Successful insertion as well as barcode sequence was monitored by sanger sequencing.

5.2.1.7 Synthesis of candidate gene cDNAs and cloning into p902 target vector

The cDNA sequences were accessed for the GRCm38 mouse genome from Ensemble (https://www.ensembl.org/index.html), choosing those isoforms with available protein (CCDS) sequence. All cDNA sequences were equipped with a BamHI recognition sequence followed by a Kozak sequence ultimately before the start codon as well as a Sbf1 recognition site ultimately after the stop codon. (cDNAs with colour coded recognition sites as well as start and stop codons are supplied in the Appendix. All cDNAs were synthesized by GeneArt (ThermoFisher) and codon optimized for enhanced transcription and translation using a software implemented algorithm. Next, cDNAs were amplified from transfer vector (Table 10) using standard protocols for bacterial transformation, cultivation and plasmid isolation. cDNAs were extracted from plasmids through digestion with BamH1 and Sbf1, followed by gel extraction using the QIAquick Gel extraction Kit (Qiagen). Finally, p902 plasmids were linearized using BamHI and SbfI restriction enzymes and cDNA fragments were ligated into the vector. Correct insertion and sequence integrity was monitored through sanger sequencing.

5.2.1.8 Barcode amplification followed by high-throughput sequencing

In order to assess barcodes in transduced cells, cells were isolated and washed in PBS followed by centrifugation at 10,000 rpm for 10 minutes. After quantitative removal of supernatant, cells were directly lysed in a PCR compatible buffer (10mM Tris-HCl at pH7.5-pH7.5, 50mM NaCl, 6.25mM $MgCl_2$, 0.045 % IGEPAL CA-630, 0.45% Tween-20, freshly added proteinase K to 1µg/µL) at 56°C for 60 minutes followed by a proteinase K inactivation step at 98°C for 10 minutes. Samples were either used directly or stored at -80°C until further usage. To amplify the barcode sequence, two nested PCRs were conducted as displayed in Figure 17, using the volumes and condition depicted in Table 32 and Table 33. After both nested PCRs, multiplexed samples were pooled and purified using the QIAquick Gel extraction Kit (Qiagen). Finally, samples adjusted to a molar concentration of 10nM (~1ng/µL) using the Qubit (Thermo Fisher) along with the DNA high sensitivity detection kit and subsequently sequenced using the Illumina HiSeq 2000 platform with V4 reagents at the DKFZ Genomics and Proteomic Core Facility using the pRSI9_GexSeqN primer and Illumina standard sequencing primers. FASTA files were de-multiplexed by the DKFZ Genomics and Proteomic Core Facility and barcode sequences were retrieved and counted using the edgeR package (Dai et al., 2014).

Table 32 | **1st PCR during nested PCR for barcode amplification**

Reagent	**Volume [µL]**	**Step**	**Temp [°C]**	**time [s]**	**#**
H_2O	30.5	Initial denaturing	94	180	1
10x Ti-Taq Buffer	5	Denaturing	94	30	2
FwdHTS (10µM) primer	1.5	Annealing	60	10	3
deltaP5_1stPCR_R primer	1.5	Elongation	72	20	4
dNTPmix (10µM each)	1	----------- Repeat #2-4 18x -----------			
Titanium Taq polymerase	0.5	Elongation	68	120	5
Sample (direct lysis)	10	Storage	8	Infinite	6

Table 33 | **2nd PCR during nested PCR for barcode amplification**

Reagent	Volume [µL]	Step	Temp [°C]	time [s]	#
H_2O	26	Initial denaturing	94	180	1
10x Ti-Taq Buffer	5	Denaturing	94	30	2
FwdGex (10µM) primer	2.5	Annealing	60	10	3
Illu_pRSI9_X (2.5µM) primers	10	Elongation	72	20	4
dNTPmix (10µM each)	1	----------- Repeat #2-4 22x -----------			
Titanium Taq polymerase	0.5	Elongation	68	120	5
Sample (direct lysis)	5	Storage	8	Infinite	6

Illu_pRSI9_X is one of 96 multiplexing primers from Table 23.

5.2.1.9 Cell culture

Culturing and passaging

$HL60^{rtTA}$ cells were cultured in RPMI supplemented with 10% FBS and 1% glutamine and passaged through centrifugation at 150g for 5min at room temperature (RT). HEK293T cells were cultured in IMDM supplemented with 10% FBS and 1% glutamine (complete medium) and passaged by gently washing the plates with PBS, followed by incubation with 0.025% trypsin at 37°C until cells were fully detached. Next, trypsin was block using complete medium and centrifuged at 150g for 5min at room temperature (RT). $LSK^{Rosa26\ rtTA}$ or $LSK\text{-}SLAM^{Rosa26\ rtTA}$ cells were cultured in StemSpan medium supplemented with 1% P/S, 100ng/mL rmSCF, 100ng/mL Flt3 Ligand, 100ng/mL rmTPO and 20ng/mL rmIL3 and passaged through centrifugation at 150g for 5min at room temperature (RT). All cells were cultured in humidified incubators at 37°C and 5% CO_2.

Freezing and storage

Vital freezing of HEK293T cells was performed after washing, detaching and centrifugation as described above, followed by re-suspension in complete medium with the addition of 10% DMSO. Subsequently, cell suspension is cooled down gradually in a -80°C freezer inside a MrFrosty container. $HL60^{rtTA}$ cells were frozen as described above, without the detaching step. Fully frozen samples were transferred to liquid nitrogen tanks for long-term storage.

5.2.1.10 Virus production

Production of lentiviral particles for candidate gene overexpression was usually performed in 15 x 15cm plates seeded with 1 x 10^7 low-passage HEK293T cells at day 1. After 24h, medium was replaced with fresh medium, followed by transient co-transfection of cells. To this end, plasmids were mixed according to Table 34 and topped off with IMDM without supplements to a total volume of 500µL. Plasmid/IMDM solution was filtered sterile using a 0.22µm filter. Separately, 500 µL IMDM was mixed with polyethylenimine (PEI) at a concentration that the final ratio between µg DNA and µg PEI is 1:3. Plasmid/IMDM and PEI/IMDM solutions were mixed and thoroughly vortexed before incubation for 15-30 minutes. Subsequently, DNA/PEI/IMDM solution was applied dropwise to 15cm plates with as little disturbance of PEI:DNA complexes as possible through i.e. additional mixing etc. After 12h, medium was replaced. Virus containing supernatant was harvested 24h, 48h and 72h after changing the medium and filtered through a 0.22µm filter. Up to 35mL of supernatant were transferred to an ultra-centrifuge tube and centrifuged for 2h at 20,000 rpm at RT. After centrifugation, supernatant was discarded and viral pellets were carefully resuspended in 50uL PBS or StemSpan without supplements and pooled into a 1.5mL Eppendorf tube. Finally, the virus concentrate was mixed on a rotary stand for 20 minutes at RT, aliquoted into 0.5 mL Eppendorf tubes and stored at -80°C until use.

Table 34 | **µg plasmids used per 15cm dish.**

Plasmid	Genes	µg
LV101	*gag-pol*	12.5
LV102	*rev*	6.25
LV103	*vsv-g*	9
p902	GOI	32

5.2.1.11 Lentiviral titer calculation

In order to calculate the number of transducing units (TU) particles per volume, 5 x 10^4 HL60^{rtTA} cells were re-suspended in 500µL RPMI with supplements and plated into 6-well plates. 500µL virus dilution were made in 1:10 steps using RPMI with supplements plus 16µg/mL protamine sulfate and 2 µg/mL doxycycline (DOX), resulting in dilutions of 1:200, 1:2,000, 1:20,000 and 1:200,000. Next, 500µL virus dilutions were placed into each well resulting in a 1:2 dilution. After 12-16h, 1mL of fresh medium was added to each well and GFP was measured 48 after transduction. Finally, the number of infectious viral particles per mL is calculated using Equation 1, considering only GFP$^+$ percentages below 20%.

Equation 1 | **Calculation of virus titer**

$$\text{Titer}(TU/mL) = \frac{50{,}000 \times \%(GFP^{+})}{100 \times dilution\ factor}$$

5.2.1.12 Lentiviral transduction of LSK$^{\text{Rosa26 rtTA}}$ or LSK-SLAM$^{\text{Rosa26 rtTA}}$ cells

Before transduction, LSK$^{\text{Rosa26 rtTA}}$ or LSK-SLAM$^{\text{Rosa26 rtTA}}$ cells were cultured as described above, or re-suspended in transduction media (see below) after sorting. The multiplicity of infection (MOI) ranged between 20-80, however was kept below 25% GFP$^+$ cells to prevent multiple integrations per cell. Cells were transduced in U-bottom 96-well plates with approx. 100,000 cells per well and re-suspended in 200µL IMDM plus supplements and 8µg/mL protamine sulfate. Culture medium was changed by centrifugation of the 96-well plate at 150g and RT for 5 minutes, followed by careful aspiration of the supernatant. Cells were cultured at least for another 24h before cell counting or flow cytometry analysis and proceeding with subsequent experiments (5.2.1.13, 5.2.1.14, 0).

5.2.1.13 Colony forming unit assay

For colony forming unit (CFU) assays promoting erythro-myeloid differentiation (MethoCult™ GF M3434, Stem Cell Technologies) LSK-SLAM$^{\text{Rosa26 rtTA}}$ cells were transduced as described in (0). Subsequently, cells were washed with StemSpan medium including supplements and cultured for additional 72h in the presence of 1µg/mL DOX. Next, cells were flow cytometry sorted sorted for GFP directly into StemSpan medium without supplements, pelleted at 150g for 5min at RT and re-suspended in MethoCult™ GF M3434 supplemented with 1% P/S and 1µg/mL DOX at a concentration of 1,200 cells per 3mL. Cells were cultured for seven days. Subsequently, cells were washed off the plates by diluting the semisolid medium with PBS and FACS sorted again for GFP$^+$ cells. For the second culture period, 10,000 GFP+ cells were plated per 3mL dish and culture as described above and cultured for another 8 days. For the third culture period, cells were again re-suspended in PBS, not sorted for GFP because of the near-100% GFP positivity, re-plated again at a concentration of 100,000 cells per 3mL of MethoCult™ GF M3434 and cultured for another 13 days. Colonies were counted before every replating under a light microscope at 40x magnification. Cell leftovers were kept before 1st plating and for every re-plating step for BC amplification and sequencing.

5.2.1.14 Cell trace assay

For the cell trace assay, 4 x 10^5 LSK$^{Rosa26\ rtTA}$ cells from donor Rosa26 rtTA mice were transduced as described above (0) after splitting the cells into $^1/_3$ GFP pool and $^2/_3$ GOI pool cells. Next, transgene expression was initiated by adding 1µg/mL DOX to StemSpan with supplements and cells were stained with CellTrace™ violet according to the manufacturers protocol. Stained cells were cultured for 3 or 5 days, respectively, and finally sorted into fast (weakest signal intensity) intermediate fast, intermediate slow and slow (highest signal intensity) cycling cells. Lastly, cell were lyzed and BCs were amplified and sequenced.

5.2.1.15 Fluorescence activated cell analyzing and sorting

Cell populations were sorted and analyzed using either the LSR II or Aria II from Becton Dickinson (BD). To this end, cells were labelled using FACS antibodies and fluorochrome combinations as described in 5.1.7 for 30min on ice with the addition of 1µg/mL FluoroGold^{TM} (Hydroxystilbamidine bis(methanesulfonate). Before sorting, cells were washed and re-suspended in Hanks balanced salt solution containing 2% FBS and filtered through a 85µm strainer. LSK$^{Rosa26\ rtT}$ or LSK-SLAM$^{Rosa26\ rtT}$ cells were either sorted directly into StemSpan or Hanks balanced salt solution containing 2% FBS. Cell populations were distinguished using the following surface marker combinations listed in Table 35.

Table 35 | **Surface marker combinations for the identification of immunophenotypic cell populations**

Abbreviation	**Cell type**	**Source**	**Cell surface combinations**
LSK	Hematopoietic stem and progenitor cells	BM	Lineage$^+$, CD117$^+$, Ly6A/E$^+$
LSK-SLAM	Hematopoietic stem cells	BM	Lineage$^+$, CD117$^+$, Ly6A/E$^+$, CD150$^+$, CD48$^-$
Granu	Granulocytes	PB	CD11b$^+$, Ly6G$^+$
Macro/Mono/DC	Macrophages, Monocytes, Dendritic cells	PB	CD11b$^+$, Ly6G$^-$
Ery	Erythroid progenitor cells	PB	Ter119$^+$
T cells	T cells, peripheral blood	PB	CD3$^+$
CD4	CD4 positive T cells (mature T helper cells)	PB	CD3$^+$, CD4$^+$
CD8	CD8 positive T cells (cytotoxic T cells)	PB	CD3$^+$, CD8$^+$
B cells	B cells, peripheral blood	PB	CD45R$^+$

5.2.2 Mouse experiments

5.2.2.1 Harvesting of bone marrow samples for the isolation of hematopoietic stem and progenitor cells

After mice were euthanizing the mice by cervical dislocation, hind legs, hip and spine bones (femur, tibia; ilium; spina) were harvested and cleaned from muscle tissue. Next, bones were crushed in Hanks balanced salt solution containing 2% FBS using pistil and mortar. Bones were rinsed until white and cell suspension was filtered through a 70µm strainer. Cells were depleted from differentiated lineage cells using the EasySep™ Mouse Hematopoietic Progenitor Cell Isolation Kit (StemCellTechnologies) according to the manufacturers' protocol and finally stained with antibodies according to Table 15 flow cytometry sorted.

5.2.2.2 Acquiring of blood samples from transduced mice

Approx. 200µL of peripheral blood samples were taken every four weeks from transduced mice by punctuation of the vena saphena and collected in EDTA tubes (Microvette CB 300 K2E). Before staining for FACS or aliquotation for direct lysis and BC amplification, erythrocytes were lysed twice by incubation with red blood cell lysis buffer for 5minutes each at RT. Cells were washed with Hanks balanced salt solution containing

2% FBS and ¼ was pelleted at 10,000 rpm for 5min and 4°C and subjected to direct lysis. The remainder was stained for flow cytometry sorting according to Table 17

5.2.2.3 Bone marrow transplantations

Approx. 24h before bone marrow transplantations, mice were lethally irradiated with 950cGy separated into two irradiation sessions. Transduced LSK cells were injected into the tail vein, which was dilated beforehand using infra-red light. In order to prevent immunogenic reactions against the transplant, cells were thoroughly washed with pure PBS to eliminate potential traces of FBS or cytokines.

5.2.3 Patient integration site data sample collection from Wiskott-Aldrich syndrome gene therapy study

For biosafety reasons, integration site (IS) data for all patients in the Wiskott-Aldrich syndrome (WAS) study was collected over a period of 6 years (Sample collection, ISs retrieval and mapping performed in the group of Manfred Schmidt, DKFZ and Genewerk, Heidelberg, Germany) (Boztug et al., 2010; Braun et al., 2014). In brief, 10 male patients between 2 to 14 years old received autologous bone marrow transplants using rhG-CSF and plerixafor (*n* = 8) or G-CSF (*n* = 2) mobilized CD34$^+$ cells. The number of infused cells ranged from 9.7 x 10^6 to 24.9 x 10^6 cells/kg for 9 out of 10 patients. Only patient 3 received significantly less cells (2.9 x 10^6 cells/kg) due to inefficient mobilization and leukopheresis. Depending on the patients weight, the total amount of infused CD34$^+$ cells varied between 2.05 x 10^9 to 6.99 x 10^9 cells (patient 3: 3.3 x 10^6 cells, median all patients: 3.5 x 10^9). This equals to approx. 20,565 to 69,877 LT-HSCs per patient (median: 35,633; sum: 369,686), assuming a LT-HSC frequency of 0.01% within CD34$^+$ cells (Biasco et al., 2016; Kim et al., 2014). Amplification and detection of γRV ISs from genomic DNA was done by linear-amplification-mediated PCR (LAM-PCR) as described before (Boztug et al., 2010; Schmidt et al., 2007). For more detailed information about patients see Braun et al. (2014) and Boztug et al. (2010).

The following paragraphs on computational analysis of the material and methods part contain text sections that have been taken from Wünsche et al. (2018) and have been originally written by myself. All literal quotes are indicated by quotation marks (” … “), following the guidelines of good scientific practice of the Ruperto-Carola University of Heidelberg.

5.2.4 Computational analysis

5.2.4.1 Acquisition of datasets used in this study

“Pre-transplant ISs positions from CD34$^+$ cells as well as post-transplant positions from CD34$^+$ cells transplanted into NSG mice and analyzed 2 month later were downloaded as supplemental data from the original publication (De Ravin et al., 2014). Pre and post-transplantation ISs positions from the ADA-SCID gene therapy trail on five patients followed up to 47 months after transplantation were acquired from the supplement from the original publication (Aiuti et al., 2007). IS positions from K562 and HepG2 cells were taken from (LaFave et al., 2014) and downloaded from https://research.nhgri.nih.gov/software/GeIST/download.shtml. Fast ATAC-seq as well as RNA-seq data of 13 primary cell types were downloaded as supplementary tables from (Corces et al., 2016). RNA-seq raw counts were converted to transcripts per million (TPM) using a custom python script and normalized using the R package DEseq2 (Love et al., 2014). Significant Capture Hi-C interactions were downloaded from ArrayExpress database under accession E-MTAB-2323. ChIP-seq data from CD34$^+$ cells were downloaded from NCBI GEO under the accession GSM706845, GSM772865, GSM772870, GSM772938, GSM772951, GSM773041. GWAS SNPs were downloaded from the NHGRI-EBI Catalog of published genome-wide association studies (https://www.ebi.ac.uk/gwas/) as *“All associations v1.0”*. Common SNPs with no reported phenotype were downloaded from NCBI Variation resource as SNPs in VCF format. Topologically associated domains were downloaded from NCBI GEO under the accession number GSE63525 as “Arrowhead_domainlist” and CTCF sites were downloaded from NCBI GEO under the accession number GSM2861703 as a BigWig file and further processed as described below. All datasets were downloaded for the hg19

(GRCh37) build or converted to hg19 using the liftOver utility with default settings from UCSC (https://genome.ucsc.edu/cgi-bin/hgLiftOver)."

5.2.4.2 Cluster analysis for the selection of candidate genes

Before the use of the ClusterCount function as described below, the local accumulation of patient ISs (common integration site; CIS) and their closest TSS was assessed using the Cyctoscape software (version 2.8.3; www.cytoscape.com) with the addition of a plugin which was kindly provided by Raffaele Fronza (NCT/DKFZ Heidelberg). Here, the threshold for the minimal distance between two ISs within the same CIS was set to 10kb and the maximum distance from the boundary of the CIS to the nearest TSS was set to 50kb.

5.2.4.3 Centered distance of ISs to TSSs and percentage of genes tagged by IS

Transcription start site (TSS) positions for protein-coding, lincRNA and miRNA genes for hg19 / GRCh37 were downloaded from Ensembl Biomart and distance for each IS to the nearest TSS was calculated. To calculate the percentage of genes tagged by IS, genes were segregated by gene class and genes with at least one IS in a 10 kb window around the TSS were considered tagged. Data for both analyses were visualized using ggplot2 for R.

5.2.4.4 Rainfall plots

Distance between patient ISs and number matched ISs from CD34$^+$ cells was calculated using the *imd* function from the ClusteredMutations package for R. ISs contained in the 100 biggest clusters were labeled and results were plotted using ggplot2 for R (see Table 27).

5.2.4.5 Cluster prediction

For all bioinformatical analysis except the candidate gene selection – the clusters were predicted using the *ClusterCount* function (provided with the Appendix) with the maximum distance between two consecutive ISs set to 2500 bp and the minimum amount of ISs per cluster to 10 IS, unless stated otherwise.

5.2.4.6 Overlap of ChIP-seq signal and IS positions

"The enrichment of histone marks was calculated as the percentage of all ChIP-seq reads which contain at least one IS. Plots were drawn using ggplot2 for R."

5.2.4.7 Pearson correlation and PCA

The similarity of all IS datasets was assessed using Pearson correlation and PCA. Therefore, the genome was divided into 10kb bins and ISs were counted for each bin. Because of the differently sized datasets CD34$^+$, K562, and HepG2 cells, datasets were randomly sampled to match the number of WAS ISs (130,673) 1,000 times and ISs per bin were counted for each sampling. Subsequently, the mean for each bin was calculated and Pearson correlation using the *cor* function for R with missing values handled by casewise deletion (use="complete"). Datasets were ordered according to the principle component 1 from the PCA analysis. For the PCA, data was scaled and centered and the calculation was performed using the *prcomp* function in R. Data was visualized with ggplot2 for R (see Table 27).

5.2.4.8 Estimation of hematopoietic switch after transplantation using pairwise positive association matrix

"The calculation of the positive association matrices for each patient was adapted from (Biasco et al., 2016). In brief, odds ratios (OR) were calculated from binarized values for ISs (1 detected or 0 not detected) for each combination of two time points (i, j) per patient […]" as described below in Equation 2.

Equation 2 | **Calculation of odds ratios between patient samples**

$$OR_{ij} = \frac{N(IS_i = 1 \wedge IS_j = 1)/N(IS_i = 1 \wedge IS_j = 0)}{N\left(IS_i = 0 \wedge IS_j = 1\right)/N(IS_i = 0 \wedge IS_j = 0)}$$

"Time points with positive correlation take *OR* indices from (1;∞) while negative associations range from (0;1). For heatmaps, only *OR*≥1 were used, infinite values (diagonal) and values below 1 were set 1. The heatmap was drawn with the R package gplots with color intensities being proportional to $\log_2$(*OR*). Note that data was not hierarchically clustered but ranked for sequencing time point, to allow for a visual identification of a change in association."

5.2.4.9 Comparison of cluster dimension and sizes between IS datasets

“Median cluster size (number of ISs per cluster) and cluster dimension (cluster span in bp) was determined for all *in vitro* data sets and WAS ISs using the *ClusterCount* function in R. For cluster prediction, the maximum distance (*d*) between the lagging and the leading IS in two consecutive cluster was set to 2.5kb and the minimum cluster size to 10 ISs per cluster. Datasets were randomly sampled 1,000 times to match the numbers of ISs between *in vitro* data sets or early and late WAS ISs. A two-sample Wilcoxon signed-rank test was performed to test for significant differences between WAS patients and each matched sampling of the corresponding *in vitro* dataset. Figure 31 shows cluster size and dimension for one representative sampling and the median p-value of 100 samplings.”

5.2.4.10 Assessment of growth kinetics to address the frequency of overproportional clonal expansion

“The closest gene for every unique IS at a given sequencing time point was determined using 159,884 TSS positions (corresponding to 30,381 genes) that were downloaded from BioMart for hg19/GRCh37 for protein-coding genes, miRNAs and lincRNAs. Next, the area under the curve (AUC) was calculated for all genes with more than 50 ISs per gene as described […]” in Equation 3 “[…] below, where $x_i, \dots, x_n$ describes the sequencing time point and $f(x_i)$ the corresponding cumulative number of IS. To compensate for non-Gaussian distribution, AUCs were logarithmized and mean and standard deviation were calculated. Next, the 95% confidence interval was calculated and subsequently back transformed. Finally, genes with an AUC outside the 95% confidence interval were marked and denoted with a red line. Plots were drawn with base R and ggplot2.”

Equation 3 | **Calculation of area under the curve**

$$\mathrm{AUC} = \sum_{i=1}^{n-1} (x_{i+1} - x_i) f(x_i)$$

5.2.4.11 Fitting the mean of the $\log_2$ odds ratio using a 4-parametric log-logistic model

"First, the mean of all $\log_2$ odds ratios (OR) between a given sample and all corresponding samples and/or time points was calculated for all patients (mean $\log_2$ OR for every column of the positive association heatmap). Next, curve fitting was performed using the "drm" function from the DRC package for R (Ritz et al., 2015). Next, the onset of stable hematopoiesis was estimated empirically for the most robustly sequenced patients 1 and 2 so that the turning point matches the onset observed in the heatmap and set to 30% of the functions maximum. Finally, the same 30% were applied to all remaining patients that had sufficient data for the curve fitting. The exact computed switch time points for patients with sufficient data are listed in Table 4. For patients 3, 6, 7 and 10, the median switch date of 404 days was used, which was calculated from Patients 1, 2, 4, 5, 8, and 9."

5.2.4.12 Assessment of ATAC-seq signal and gene expression at sites of integration

"Clusters of ISs were predicted as described above and assigned to its closest gene using 159,884 TSS positions (corresponding to 30,381 genes) that were downloaded from BioMart for hg19/GRCh37 for protein-coding genes, miRNAs and lincRNAs. Next all clusters belonging to the same gene were aggregated and the sum of ISs was plotted against TPM gene expression with ggplot2. For the median expression of all IS-tagged genes, TPMs were extracted for all 13 primary cell types and plotted as violin plot using ggplot2. Pairwise statistical comparison was performed using a two-sample Wilcoxon signed-rank test. For the median signal intensity at sites of integration, either all ATAC-seq signals or signature peak ATAC-seq signals were normalized using the "normalize.quantiles" function from the preprocessCore package for R, and extracted in a genomic window of 1kb around all IS. CD34$^+$ ISs were randomly sampled 1,000 times to correct for the lower number of WAS IS. Signal intensities were plotted for all IS-containing ATAC-seq peaks as boxplots using ggplot2 and pairwise statistical comparison was performed using a two-sample Wilcoxon signed-rank test."

5.2.4.13 Differentially tagged Ggnes between CD34$^+$ cells and patients and gene set enrichment analysis

"Differential number of ISs per gene was calculated by first determining cluster positions and number of ISs per cluster as described above for late occurring ISs and CD34$^+$ cells (maximum distance: 2,500bp; minimum number of ISs per cluster: 10). ISs from CD34$^+$ cells were randomly sampled 1,000 times to match the lower number of ISs from patients, and clusters were allocated to its closest gene for each iteration as described above, followed by aggregation of clusters marking the same gene. Next, genes were dismissed which were tagged by clusters at a lower frequency than the mean number of genes tagged for each iteration. Finally, pairwise comparison was performed by subtracting (difference) or dividing (fold change) the mean sum of CD34$^+$ ISs from the sum of late occurring WAS ISs for each gene. In order to calculate fold changes for genes that have 0 ISs in either dataset, one pseudo IS was added to all genes. For gene set enrichment analysis (GSEA), only genes showing a difference > 10 or < -10 ISs were used to filter out genes that only show a minor difference between ISs from CD34$^+$ cells and WAS IS. […] Next, the $\log_2$ fold changes were used and matched to the C2 curated gene set v6.0 from the Molecular Signatures Database (MSigDB) alongside with 20 custom gene sets generated from Cabezas-Wallscheid et al. (2014). The GSEA analysis was conducted using the piano package in R (Varemo et al., 2013) with arguments set to signifMethod = "nullDist", geneSetStat = "page", and adjMethod = "fdr". The network was generated using the piano package on HSC relevant gene sets using the standard parameters. Genes contained in network gene sets were extracted, sorted according to p-value of the gene set and $\log_2$ fold change and plotted as heatmap using ggplot2."

5.2.4.14 Assessment of activity of MYC enhancer modules using WAS ISs

"Genomic positions of BENC modules were taken from Bahr et al. (2018) and slightly expanded to fit the ATAC-seq pattern of HSCs: A/B = chr8:130555999-130575896; C/D = chr8:130592385-130606829; G/I = chr8:130675980-130700504. Genomic positions of modules X1, X2 and X3 were set to match the ATAC-seq signal in HSC and the IS pattern observed. X1 = chr8:130429984-130436953; X2 = chr8:130546989-130552736; X3 = chr8:130652912-130657000. ISs were extracted given genomic ranges and duplicates were discarded by retaining the IS (clone) that was sequenced at the last time point. ISs were categorized into early and late with the patient specific cut off plotted according to their sequencing time point using ggplot2."

5.2.4.15 Enrichment of ISs in CHi-C interaction fragments

"As the CHi-C fragments from (Mifsud et al., 2015) were generated using the HindIII restriction enzyme, we first digested the hg19 genome *in silico* using the HindIII restriction enzyme, through a custom Java script and hg19 fasta files from http://hgdownload.cse.ucsc.edu/goldenPath/hg19/chromosomes/). Next, contingency tables were produced by determining the number of HindIII fragments that either interact but do not harbor any IS, harbor ISs but do not interact, neither interact nor harbor any IS, or harbor both, interaction and IS. Expected values and significance were calculated using a Chi-square test with Yates's correction for continuity for R and plotted using ggplot2."

5.2.4.16 GWAS and common SNP enrichment analysis

"In total 33,044 GWAS SNPs and 38,138,476 common SNP were downloaded as described above. GWAS SNPs from non-European studies as well as chromosomal translocations and abnormalities were filtered out in order to match the genetic background of the patients with the data, yielding 24,434 remaining GWAS SNPs. Next, the overlap of GWAS SNPs and IS positions (±2.5kb) or the same quantity of randomly sampled common SNPs and the same IS positions, respectively. The significance of difference was estimated by Chi-squared test, using the mean overlap of 1,000 random samplings from common SNPs vs. the overlap of GWAS SNPs. Next, GWAS SNPs were classified into 17 categories, adapted from Mifsud et al. (2015) and Maurano et al. (2012) […]. The categorical enrichment was calculated as the percent overlap of the categorical GWAS SNPs subtracted by the mean percent of overlapping GWAS SNPs. Statistical significance was calculated using the Fisher's exact test by comparing the overlapping and the non-overlapping GWAS SNPs of each category with all un-categorized overlapping and non-overlapping GWAS SNPs."

5.2.4.17 Circularized view of CD34$^+$ capture Hi-C long-range interactions and WAS ISs and WashU epigenome browser custom track

For the publication (Wünsche et al., 2018), "circular plots including CHi-C interactions and patient ISs […]" were prepared and "[…] can be accessed using the Capture HiC Plotter (Schofield et al., 2016) (https://www.chicp.org/). Additionally, a Washington University EpiGenome Browser session (ID wgTns1P1rr) is available for displaying fully analyzed data for number matched patient and CD34$^+$ IS, CD34$^+$ CHi-C interactions, ATAC-seq peaks of 13 primary cell types and ChIP-seq data for selected histone modifications in CD34$^+$ cells."

5.2.4.18 Sequence conservation at sites of integration

"PhastCons and PhyloP scores for multiple alignments of 45 primates to the human genome (46way) were downloaded as *fixed step wiggle* files from UCSC for the hg19 genome build. First, files were converted to bigwig format using the wigToBigWig and subsequently converted to bedGraph file using the bigWigToBedGraph utilities (http://hgdownload.cse.ucsc.edu/admin/exe/). After converting all data into 1-based positions, both phastCons and phyloP scores were extracted chromosome wise for IS, GWAS SNP, as well as common SNP positions. Statistical differences were determined for each chromosome using a two-sample Wilcoxon signed-rank test comparing WAS ISs and GWAS SNPs, common SNPs as well as 1 x 10^6 randomly extracted scores. Likewise, scores for all positions 250 bp upstream as well as 250 bp downstream were extracted. Finally, the mean score for all IS, GWAS SNPs or common SNPs at every basepair coordinate was calculated and plotted. Statistical comparison was only performed at the actual position of ISs or SNPs and is presented as the median p-value for all chromosomes."

5.2.4.19 Permutation test to assess the enrichment of ISs at TAD boundaries

"As a control 10,000 random TADs were created from the original TADs from Rao et al. (2014) using "shuffle" from bedtools 2.26.0 with arguments set to –excl –chrom –noOverlapping –allowBeyondChromEnd, where –excl depicts all hg19 genome gaps (http://hgdownload.cse.*ucsc*.edu/goldenPath/hg19/database/*gap.txt.gz*). Next, TAD boundaries were created by extending both, the start and the end coordinates of a TAD, by ±2.5kb. Finally, the percentage of overlap for all 10,000 random TAD boundaries was assessed and plotted as a histogram using ggplot2. The p-value is calculated by dividing the number of events where the overlap is either greater (depletion) or smaller (enrichment) than the actual overlap, by the number of permutations.”

5.2.4.20 Permutation test to assess the enrichment of ISs at CTCF sites from CD34$^+$ cells

“First, the bigwig file was converted to BedGraph using the bigWigToBedGraph binary from http://hgdownload.cse.ucsc.edu/admin/exe/linux.x86_64/bigWigToBedGraph and peak calling with MACS2.1.1 (Liu, 2014) was performed using the bdgpeakcall function with minimum gap (-l) set to 50bp and default max gap (-g; 30bp). Next, peaks were filtered for *in silico* predicted CTCF binding sites generated for hg19 using the probability matrix from JASPAR (http://jaspar.genereg.net/matrix/MA0139.1/) and FIMO (http://meme-suite.org/tools/fimo) with default settings, resulting in approx. 31,000 high confidence CTCF peaks in CD34$^+$ cells.”

6 References

Adolfsson, J., Borge, O.J., Bryder, D., Theilgaard-Monch, K., Astrand-Grundstrom, I., Sitnicka, E., Sasaki, Y., and Jacobsen, S.E. (2001). Upregulation of Flt3 expression within the bone marrow Lin(-)Sca1(+)c-kit(+) stem cell compartment is accompanied by loss of self-renewal capacity. Immunity *15*, 659-669.

Aguilo, F., Avagyan, S., Labar, A., Sevilla, A., Lee, D.F., Kumar, P., Lemischka, I.R., Zhou, B.Y., and Snoeck, H.W. (2011). Prdm16 is a physiologic regulator of hematopoietic stem cells. Blood *117*, 5057-5066.

Aiuti, A., Cassani, B., Andolfi, G., Mirolo, M., Biasco, L., Recchia, A., Urbinati, F., Valacca, C., Scaramuzza, S., Aker, M.*, et al.* (2007). Multilineage hematopoietic reconstitution without clonal selection in ADA-SCID patients treated with stem cell gene therapy. J Clin Invest *117*, 2233-2240.

Aiuti, A., Slavin, S., Aker, M., Ficara, F., Deola, S., Mortellaro, A., Morecki, S., Andolfi, G., Tabucchi, A., Carlucci, F.*, et al.* (2002). Correction of ADA-SCID by stem cell gene therapy combined with nonmyeloablative conditioning. Science *296*, 2410-2413.

Aker, M., Tubb, J., Miller, D.G., Stamatoyannopoulos, G., and Emery, D.W. (2006). Integration Bias of Gammaretrovirus Vectors following Transduction and Growth of Primary Mouse Hematopoietic Progenitor Cells with and without Selection. Molecular Therapy *14*, 226-235.

Albert, M.H., Notarangelo, L.D., and Ochs, H.D. (2011). Clinical spectrum, pathophysiology and treatment of the Wiskott-Aldrich syndrome. Current opinion in hematology *18*, 42-48.

Alberts, B., Johnson, A., Lewsi, J., Raff, M., Roberts, K., and Walter, P. (2008). Molecular Biology of the Cell. Garland Science *5th Edition*.

Aldrich, R.A., Steinberg, A.G., and Campbell, D.C. (1954). Pedigree demonstrating a sex-linked recessive condition characterized by draining ears, eczematoid dermatitis and bloody diarrhea. Pediatrics *13*, 133-139.

Ali, T., Renkawitz, R., and Bartkuhn, M. (2016). Insulators and domains of gene expression. Current Opinion in Genetics & Development *37*, 17-26.

Allen, B.L., and Taatjes, D.J. (2015). The Mediator complex: a central integrator of transcription. Nature reviews Molecular cell biology *16*, 155-166.

Amano, T., Sagai, T., Tanabe, H., Mizushina, Y., Nakazawa, H., and Shiroishi, T. (2009). Chromosomal dynamics at the Shh locus: limb bud-specific differential regulation of competence and active transcription. Dev Cell *16*, 47-57.

Arnold, C.D., Gerlach, D., Stelzer, C., Boryn, L.M., Rath, M., and Stark, A. (2013). Genome-wide quantitative enhancer activity maps identified by STARR-seq. Science *339*, 1074-1077.

Bahr, C., von Paleske, L., Uslu, V.V., Remeseiro, S., Takayama, N., Ng, S.W., Murison, A., Langenfeld, K., Petretich, M., Scognamiglio, R., *et al.* (2018). A Myc enhancer cluster regulates normal and leukaemic haematopoietic stem cell hierarchies. Nature *553*, 515-520.

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

Balvay, L., Lastra, M.L., Sargueil, B., Darlix, J.-L., and Ohlmann, T. (2007). Translational control of retroviruses. Nature Reviews Microbiology *5*, 128.

Banerji, J., Rusconi, S., and Schaffner, W. (1981). Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. Cell *27*, 299-308.

Barnes, D.W.H., Corp, M.J., Loutit, J.F., and Neal, F.E. (1956). Treatment of Murine Leukaemia with X Rays and Homologous Bone Marrow. British Medical Journal *2*, 626-627.

Baum, C., Dullmann, 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. Molecular Therapy *9*, 5-13.

Baum, C.M., Weissman, I.L., Tsukamoto, A.S., Buckle, A.M., and Peault, B. (1992). Isolation of a candidate human hematopoietic stem-cell population. Proc Natl Acad Sci U S A *89*, 2804-2808.

Bhatia, M., Wang, J.C., Kapp, U., Bonnet, D., and Dick, J.E. (1997). Purification of primitive human hematopoietic cells capable of repopulating immune-deficient mice. Proc Natl Acad Sci U S A *94*, 5320-5325.

Biasco, L., Baricordi, C., and Aiuti, A. (2012). Retroviral integrations in gene therapy trials. Mol Ther *20*, 709-716.

Biasco, L., Pellin, D., Scala, S., Dionisio, F., Basso-Ricci, L., Leonardelli, L., Scaramuzza, S., Baricordi, C., Ferrua, F., Cicalese, M.P., *et al.* (2016). In Vivo Tracking of Human Hematopoiesis Reveals Patterns of Clonal Dynamics during Early and Steady-State Reconstitution Phases. Cell Stem Cell *19*, 107-119.

Bishop, J.M. (1983). Cellular oncogenes and retroviruses. Annual review of biochemistry *52*, 301-354.

Bortin, M.M. (1970). A COMPENDIUM OF REPORTED HUMAN BONE MARROW TRANSPLANTS. Transplantation *9*, 571-587.

Boyle, A.P., Davis, S., Shulha, H.P., Meltzer, P., Margulies, E.H., Weng, Z., Furey, T.S., and Crawford, G.E. (2008). High-Resolution Mapping and Characterization of Open Chromatin across the Genome. Cell *132*, 311-322.

Boztug, K., Schmidt, M., Schwarzer, A., Banerjee, P.P., Diez, I.A., Dewey, R.A., Bohm, M., Nowrouzi, A., Ball, C.R., Glimm, H*., et al.* (2010). Stem-cell gene therapy for the Wiskott-Aldrich syndrome. N Engl J Med *363*, 1918-1927.

Braun, C.J., Boztug, K., Paruzynski, A., Witzel, M., Schwarzer, A., Rothe, M., Modlich, U., Beier, R., Gohring, G., Steinemann, D*., et al.* (2014). Gene therapy for Wiskott-Aldrich syndrome--long-term efficacy and genotoxicity. Sci Transl Med *6*, 227ra233.

Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y., and Greenleaf, W.J. (2013). Transposition of native chromatin for multimodal regulatory analysis and personal epigenomics. Nature methods *10*, 1213-1218.

Buenrostro, J.D., Wu, B., Chang, H.Y., and Greenleaf, W.J. (2015). ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. Curr Protoc Mol Biol *109*, 21 29 21-29.

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

Busch, K., and Rodewald, H.R. (2016). Unperturbed vs. post-transplantation hematopoiesis: both in vivo but different. Current opinion in hematology *23*, 295-303.

Cabezas-Wallscheid, N., Klimmeck, D., Hansson, J., Lipka, D.B., Reyes, A., Wang, Q., Weichenhan, D., Lier, A., von Paleske, L., Renders, S*., et al.* (2014). Identification of regulatory networks in HSCs and their immediate progeny via integrated proteome, transcriptome, and DNA methylome analysis. Cell Stem Cell *15*, 507-522.

Carlson, M., and Botstein, D. (1982). Two differentially regulated mRNAs with different 5' ends encode secreted with intracellular forms of yeast invertase. Cell *28*, 145-154.

Carrelha, J., Meng, Y., Kettyle, L.M., Luis, T.C., Norfo, R., Alcolea, V., Boukarabila, H., Grasso, F., Gambardella, A., Grover, A*., et al.* (2018). Hierarchically related lineage-restricted fates of multipotent haematopoietic stem cells. Nature *554*, 106.

Castle, J.C. (2011). SNPs occur in regions with less genomic sequence conservation. PLoS One *6*, e20660.

Cattoglio, C., Facchini, G., Sartori, D., Antonelli, A., Miccio, A., Cassani, B., Schmidt, M., von Kalle, C., Howe, S., Thrasher, A.J.*, et al.* (2007). Hot spots of retroviral integration in human CD34+ hematopoietic cells. Blood *110*, 1770-1778.

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.

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.

Chen, S., Sanjana, Neville E., Zheng, K., Shalem, O., Lee, K., Shi, X., Scott, David A., Song, J., Pan, Jen Q., Weissleder, R.*, et al.* (2015). Genome-wide CRISPR Screen in a Mouse Model of Tumor Growth and Metastasis. Cell *160*, 1246-1260.

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 U S A *98*, 14541-14546.

Civin, C.I., Strauss, L.C., Brovall, C., Fackler, M.J., Schwartz, J.F., and Shaper, J.H. (1984). Antigenic analysis of hematopoiesis. III. A hematopoietic progenitor cell surface antigen defined by a monoclonal antibody raised against KG-1a cells. Journal of immunology (Baltimore, Md : 1950) *133*, 157-165.

Copley, M.R., Babovic, S., Benz, C., Knapp, D.J.H.F., Beer, P.A., Kent, D.G., Wohrer, S., Treloar, D.Q., Day, C., Rowe, K.*, et al.* (2013). The Lin28b–let-7–Hmga2 axis determines the higher self-renewal potential of fetal haematopoietic stem cells. Nature Cell Biology *15*, 916.

Coppola, C.J., C. Ramaker, R., and Mendenhall, E.M. (2016). Identification and function of enhancers in the human genome. Human Molecular Genetics *25*, R190-R197.

Corces, M.R., Buenrostro, J.D., Wu, B., Greenside, P.G., Chan, S.M., Koenig, J.L., Snyder, M.P., Pritchard, J.K., Kundaje, A., Greenleaf, W.J.*, et al.* (2016). Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. Nat Genet *48*, 1193-1203.

Corradin, O., and Scacheri, P.C. (2014). Enhancer variants: evaluating functions in common disease. Genome medicine *6*, 85.

Crotty, S., and Pipkin, M.E. (2015). In vivo RNAi screens: concepts and applications. Trends in immunology *36*, 315-322.

Dai, Z., Sheridan, J.M., Gearing, L.J., Moore, D.L., Su, S., Wormald, S., Wilcox, S., O'Connor, L., Dickins, R.A., Blewitt, M.E.*, et al.* (2014). edgeR: a versatile tool for the analysis of shRNA-seq and CRISPR-Cas9 genetic screens. F1000Res *3*, 95.

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

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

Deichmann, A., Brugman, M.H., Bartholomae, C.C., Schwarzwaelder, K., Verstegen, M.M., Howe, S.J., Arens, A., Ott, M.G., Hoelzer, D., Seger, R., *et al.* (2011). Insertion sites in engrafted cells cluster within a limited repertoire of genomic areas after gammaretroviral vector gene therapy. Mol Ther *19*, 2031-2039.

Dekker, J. (2006). The three 'C' s of chromosome conformation capture: controls, controls, controls. Nat Methods *3*, 17-21.

Dekker, J., Rippe, K., Dekker, M., and Kleckner, N. (2002). Capturing chromosome conformation. Science *295*, 1306-1311.

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

Deneault, E., Cellot, S., Faubert, A., Laverdure, J.P., Frechette, M., Chagraoui, J., Mayotte, N., Sauvageau, M., Ting, S.B., and Sauvageau, G. (2009). A functional screen to identify novel effectors of hematopoietic stem cell activity. Cell *137*, 369-379.

Derry, J.M., Ochs, H.D., and Francke, U. (1994). Isolation of a novel gene mutated in Wiskott-Aldrich syndrome. Cell *78*, 635-644.

Dixon, Jesse R., Gorkin, David U., and Ren, B. (2016). Chromatin Domains: The Unit of Chromosome Organization. Molecular cell *62*, 668-680.

Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature *485*, 376-380.

Doench, J.G. (2017). Am I ready for CRISPR? A user&#39;s guide to genetic screens. Nature Reviews Genetics *19*, 67.

Dostie, J., Richmond, T.A., Arnaout, R.A., Selzer, R.R., Lee, W.L., Honan, T.A., Rubio, E.D., Krumm, A., Lamb, J., Nusbaum, C., *et al.* (2006). Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. Genome Res *16*, 1299-1309.

Doulatov, S., Notta, F., Laurenti, E., and Dick, J.E. (2012). Hematopoiesis: a human perspective. Cell Stem Cell *10*, 120-136.

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.J., Brinkman, R., and Eaves, C. (2007). Long-term propagation of distinct hematopoietic differentiation programs in vivo. Cell Stem Cell *1*, 218-229.

Ellermann, V., and Bang, O. (1908). Experimentelle Leukämie bei Hühnern. Centralblatt für Bakteriologie *86*, 595-609.

Ernst, J., and Kellis, M. (2017). Chromatin-state discovery and genome annotation with ChromHMM. Nature Protocols *12*, 2478.

Ernst, J., Kheradpour, P., Mikkelsen, T.S., Shoresh, N., Ward, L.D., Epstein, C.B., Zhang, X., Wang, L., Issner, R., Coyne, M*., et al.* (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. Nature *473*, 43-49.

Filby, A., Begum, J., Jalal, M., and Day, W. (2015). Appraising the suitability of succinimidyl and lipophilic fluorescent dyes to track proliferation in non-quiescent cells by dye dilution. Methods (San Diego, Calif) *82*, 29-37.

Ford, C.E., Hamerton, J.L., Barnes, D.W., and Loutit, J.F. (1956). Cytological identification of radiation-chimaeras. Nature *177*, 452-454.

Forman, S.J., Negrin, R.S., Antin, J.H., and Appelbaum, F.R. (2015). Thomas' hematopoietic cell transplantation : stem cell transplantation, Fifth edition edn (England: Chichester, West Sussex, United Kingdom ; Hoboken, NJ : John Wiley & Sons Inc., 2015.).

Geyer, P.K. (1997). The role of insulator elements in defining domains of gene expression. Curr Opin Genet Dev *7*, 242-248.

Glimm, H., Ball, C.R., and von Kalle, C. (2011). You can count on this: barcoded hematopoietic stem cells. Cell Stem Cell *9*, 390-392.

Gorkin, D.U., Leung, D., and Ren, B. (2014). The 3D Genome in Transcriptional Regulation and Pluripotency. Cell stem cell *14*, 762-775.

Grinenko, T., Eugster, A., Thielecke, L., Ramasz, B., Krüger, A., Dietz, S., Glauche, I., Gerbaulet, A., von Bonin, M., Basak, O*., et al.* (2018). Hematopoietic stem cells can differentiate into restricted myeloid progenitors before cell division in mice. Nature Communications *9*, 1898.

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

Haas, S., Hansson, J., Klimmeck, D., Loeffler, D., Velten, L., Uckelmann, H., Wurzer, S., Prendergast, A.M., Schnell, A., Hexel, K.*, et al.* (2015). Inflammation-Induced Emergency Megakaryopoiesis Driven by Hematopoietic Stem Cell-like Megakaryocyte Progenitors. Cell Stem Cell *17*, 422-434.

Haas, S., Trumpp, A., and Milsom, M.D. (2018). Causes and Consequences of Hematopoietic Stem Cell Heterogeneity. Cell Stem Cell *22*, 627-638.

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. The Journal of Clinical Investigation *118*, 3132-3142.

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

Hajdu, S.I. (2003). A note from history: The discovery of blood cells. Annals of clinical and laboratory science *33*, 237-238.

Harmston, N., Ing-Simmons, E., Tan, G., Perry, M., Merkenschlager, M., and Lenhard, B. (2017). Topologically associating domains are ancient features that coincide with Metazoan clusters of extreme noncoding conservation. Nat Commun *8*, 441.

Heinz, S., Romanoski, C.E., Benner, C., and Glass, C.K. (2015). The selection and function of cell type-specific enhancers. Nat Rev Mol Cell Biol *16*, 144-154.

Holmfeldt, P., Ganuza, M., Marathe, H., He, B., Hall, T., Kang, G., Moen, J., Pardieck, J., Saulsberry, A.C., Cico, A.*, et al.* (2016). Functional screen identifies regulators of murine hematopoietic stem cell repopulation. The Journal of Experimental Medicine *213*, 433-449.

Hong, S., and Kim, D. (2017). Computational characterization of chromatin domain boundary-associated genomic elements. Nucleic Acids Research *45*, 10403-10414.

Hope, K.J., Cellot, S., Ting, S.B., MacRae, T., Mayotte, N., Iscove, N.N., and Sauvageau, G. (2010). An RNAi screen identifies Msi2 and Prox1 as having opposite roles in the regulation of hematopoietic stem cell activity. Cell Stem Cell *7*, 101-113.

Hou, C., Zhao, H., Tanimoto, K., and Dean, A. (2008). CTCF-dependent enhancer-blocking by alternative chromatin loop formation. Proceedings of the National Academy of Sciences of the United States of America *105*, 20398-20403.

Howe, S.J., Mansour, M.R., Schwarzwaelder, K., Bartholomae, C., Hubank, M., Kempski, H., Brugman, M.H., Pike-Overzet, K., Chatters, S.J., de Ridder, D*., et al.* (2008). Insertional mutagenesis combined with acquired somatic mutations causes leukemogenesis following gene therapy of SCID-X1 patients. J Clin Invest *118*, 3143-3150.

Hubisz, M.J., Pollard, K.S., and Siepel, A. (2011). PHAST and RPHAST: phylogenetic analysis with space/time models. Briefings in bioinformatics *12*, 41-51.

Huntsman, H.D., Bat, T., Cheng, H., Cash, A., Cheruku, P.S., Fu, J.F., Keyvanfar, K., Childs, R.W., Dunbar, C.E., and Larochelle, A. (2015). Human hematopoietic stem cells from mobilized peripheral blood can be purified based on CD49f integrin expression. Blood *126*, 1631-1633.

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

Jeong, M., Huang, X., Zhang, X., Su, J., Shamim, M., Bochkov, I., Reyes, J., Jung, H., Heikamp, E., Presser Aiden, A*., et al.* (2017). A Cell Type-Specific Class of Chromatin Loops Anchored at Large DNA Methylation Nadirs. bioRxiv.

Karamitros, D., Stoilova, B., Aboukhalil, Z., Hamey, F., Reinisch, A., Samitsch, M., Quek, L., Otto, G., Repapi, E., Doondeea, J*., et al.* (2018). Single-cell analysis reveals the continuum of human lympho-myeloid progenitor cells. Nature immunology *19*, 85-97.

Karki, R., Pandya, D., Elston, R.C., and Ferlini, C. (2015). Defining “mutation” and “polymorphism” in the era of personal genomics. BMC Medical Genomics *8*, 37.

Kataoka, K., Sato, T., Yoshimi, A., Goyama, S., Tsuruta, T., Kobayashi, H., Shimabe, M., Arai, S., Nakagawa, M., Imai, Y*., et al.* (2011). Evi1 is essential for hematopoietic stem cell self-renewal, and its expression marks hematopoietic cells with long-term multilineage repopulating activity. J Exp Med *208*, 2403-2416.

Kiel, M.J., Yilmaz, O.H., Iwashita, T., Yilmaz, O.H., Terhorst, C., and Morrison, S.J. (2005). SLAM family receptors distinguish hematopoietic stem and progenitor cells and reveal endothelial niches for stem cells. Cell *121*, 1109-1121.

Kim, S., Kim, N., Presson, A.P., Metzger, M.E., Bonifacino, A.C., Sehl, M., Chow, S.A., Crooks, G.M., Dunbar, C.E., An, D.S*., et al.* (2014). Dynamics of HSPC repopulation in nonhuman primates revealed by a decade-long clonal-tracking study. Cell Stem Cell *14*, 473-485.

King, W., Patel, M.D., Lobel, L.I., Goff, S.P., and Nguyen-Huu, M.C. (1985). Insertion mutagenesis of embryonal carcinoma cells by retroviruses. Science *228*, 554-558.

Knight, S., Collins, M., and Takeuchi, Y. (2013). Insertional mutagenesis by retroviral vectors: current concepts and methods of analysis. Curr Gene Ther *13*, 211-227.

Kool, J., and Berns, A. (2009). High-throughput insertional mutagenesis screens in mice to identify oncogenic networks. Nature reviews Cancer *9*, 389-399.

Kumar, M., Keller, B., Makalou, N., and Sutton, R.E. (2001). Systematic determination of the packaging limit of lentiviral vectors. Hum Gene Ther *12*, 1893-1905.

Kvaratskhelia, M., Sharma, A., Larue, R.C., Serrao, E., and Engelman, A. (2014). Molecular mechanisms of retroviral integration site selection. Nucleic Acids Res *42*, 10209-10225.

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

Lansdorp, P.M., Sutherland, H.J., and Eaves, C.J. (1990). Selective expression of CD45 isoforms on functional subpopulations of CD34+ hemopoietic cells from human bone marrow. J Exp Med *172*, 363-366.

Larue, R.C., Plumb, M.R., Crowe, B.L., Shkriabai, N., Sharma, A., DiFiore, J., Malani, N., Aiyer, S.S., Roth, M.J., Bushman, F.D.*, et al.* (2014). Bimodal high-affinity association of Brd4 with murine leukemia virus integrase and mononucleosomes. Nucleic Acids Res *42*, 4868-4881.

Laurenti, E., and Göttgens, B. (2018). From haematopoietic stem cells to complex differentiation landscapes. Nature *553*, 418.

Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O.*, et al.* (2009). Comprehensive mapping of long range interactions reveals folding principles of the human genome. Science (New York, NY) *326*, 289-293.

Liu, G., Mattick, J.S., and Taft, R.J. (2013). A meta-analysis of the genomic and transcriptomic composition of complex life. Cell Cycle *12*, 2061-2072.

Liu, T. (2014). Use model-based Analysis of ChIP-Seq (MACS) to analyze short reads generated by sequencing protein-DNA interactions in embryonic stem cells. Methods Mol Biol *1150*, 81-95.

Liu, Y., Kern, J.T., Walker, J.R., Johnson, J.A., Schultz, P.G., and Luesch, H. (2007). A genomic screen for activators of the antioxidant response element. Proc Natl Acad Sci U S A *104*, 5205-5210.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol *15*, 550.

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.

Ludwig, D.L., and Bruschi, C.V. (1991). The 2-micron plasmid as a nonselectable, stable, high copy number yeast vector. Plasmid *25*, 81-95.

Ma, M., Ru, Y., Chuang, L.S., Hsu, N.Y., Shi, L.S., Hakenberg, J., Cheng, W.Y., Uzilov, A., Ding, W., Glicksberg, B.S.*, et al.* (2015). Disease-associated variants in different categories of disease located in distinct regulatory elements. BMC Genomics *16 Suppl 8*, S3.

Malinova, D., Fritzsche, M., Nowosad, C.R., Armer, H., Munro, P.M., Blundell, M.P., Charras, G., Tolar, P., Bouma, G., and Thrasher, A.J. (2016). WASp-dependent actin cytoskeleton stability at the dendritic cell immunological synapse is required for extensive, functional T cell contacts. Journal of leukocyte biology *99*, 699-710.

Manley, L.J., Ma, D., and Levine, S.S. (2016). Monitoring Error Rates In Illumina Sequencing. Journal of biomolecular techniques : JBT *27*, 125-128.

Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H., Brody, J.*, et al.* (2012). Systematic localization of common disease-associated variation in regulatory DNA. Science *337*, 1190-1195.

McKenzie, J.L., Gan, O.I., Doedens, M., Wang, J.C., and Dick, J.E. (2006). Individual stem cells with highly variable proliferation and self-renewal properties comprise the human hematopoietic stem cell compartment. Nature immunology *7*, 1225-1233.

McKenzie, J.L., Takenaka, K., Gan, O.I., Doedens, M., and Dick, J.E. (2007). Low rhodamine 123 retention identifies long-term human hematopoietic stem cells within the Lin-CD34+CD38- population. Blood *109*, 543-545.

Melnikov, A., Murugan, A., Zhang, X., Tesileanu, T., Wang, L., Rogov, P., Feizi, S., Gnirke, A., Callan, C.G., Jr., Kinney, J.B.*, et al.* (2012). Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. Nat Biotechnol *30*, 271-277.

Mifsud, B., Tavares-Cadete, F., Young, A.N., Sugar, R., Schoenfelder, S., Ferreira, L., Wingett, S.W., Andrews, S., Grey, W., Ewels, P.A.*, et al.* (2015). Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. Nat Genet *47*, 598-606.

Miles, L.A., Garippa, R.J., and Poirier, J.T. (2016). Design, execution, and analysis of pooled in vitro CRISPR/Cas9 screens. Febs j *283*, 3170-3180.

Mohr, S., Bakal, C., and Perrimon, N. (2010). Genomic Screening with RNAi: Results and Challenges. Annual review of biochemistry *79*, 37-64.

Moolten, F.L., and Cupples, L.A. (1992). A model for predicting the risk of cancer consequent to retroviral gene therapy. Hum Gene Ther *3*, 479-486.

Morita, Y., Ema, H., and Nakauchi, H. (2010). Heterogeneity and hierarchy within the most primitive hematopoietic stem cell compartment. J Exp Med *207*, 1173-1182.

Müller-Sieburg, C.E., Townsend, K., Weissman, I.L., and Rennick, D. (1988). Proliferation and differentiation of highly enriched mouse hematopoietic stem cells and progenitor cells in response to defined growth factors. J Exp Med *167*, 1825-1840.

Müller-Sieburg, C.E., Whitlock, C.A., and Weissman, I.L. (1986). Isolation of two early B lymphocyte progenitors from mouse marrow: A committed Pre-Pre-B cell and a clonogenic Thy-1<sup>lo</sup> hematopoietic stem cell. Cell *44*, 653-662.

Notta, F., Doulatov, S., Laurenti, E., Poeppl, A., Jurisica, I., and Dick, J.E. (2011). Isolation of single human hematopoietic stem cells capable of long-term multilineage engraftment. Science *333*, 218-221.

Notta, F., Zandi, S., Takayama, N., Dobson, S., Gan, O.I., Wilson, G., Kaufmann, K.B., McLeod, J., Laurenti, E., Dunant, C.F.*, et al.* (2016). Distinct routes of lineage development reshape the human blood hierarchy across ontogeny. Science *351*, aab2116.

Nowak, J. (2008). Role of HLA in hematopoietic SCT. Bone Marrow Transplantation *42*, S71.

Ochs, H.D., and Notarangelo, L.D. (2005). Structure and function of the Wiskott-Aldrich syndrome protein. Current opinion in hematology *12*, 284-291.

Ogawa, M., Matsuzaki, Y., Nishikawa, S., Hayashi, S., Kunisada, T., Sudo, T., Kina, T., Nakauchi, H., and Nishikawa, S. (1991). Expression and function of c-kit in hemopoietic progenitor cells. J Exp Med *174*, 63-71.

Ong, C.T., and Corces, V.G. (2011). Enhancer function: new insights into the regulation of tissue-specific gene expression. Nat Rev Genet *12*, 283-293.

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

Ott, M.G., Schmidt, M., Schwarzwaelder, K., Stein, S., Siler, U., Koehl, U., Glimm, H., Kuhlcke, 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, R.P., Hiatt, J.B., Witten, D.M., Kim, M.J., Smith, R.P., May, D., Lee, C., Andrie, J.M., Lee, S.I., Cooper, G.M.*, et al.* (2012). Massively parallel functional dissection of mammalian enhancers in vivo. Nat Biotechnol *30*, 265-270.

Pollard, K.S., Hubisz, M.J., Rosenbloom, K.R., and Siepel, A. (2010). Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res *20*, 110-121.

Prelich, G. (2012). Gene Overexpression: Uses, Mechanisms, and Interpretation. Genetics *190*, 841-854.

Rao, S.S., Huntley, M.H., Durand, N.C., Stamenova, E.K., Bochkov, I.D., Robinson, J.T., Sanborn, A.L., Machol, I., Omer, A.D., Lander, E.S*., et al.* (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell *159*, 1665-1680.

Rieger, M.A., and Schroeder, T. (2012). Hematopoiesis. Cold Spring Harbor perspectives in biology *4*.

Ritz, C., Baty, F., Streibig, J.C., and Gerhard, D. (2015). Dose-Response Analysis Using R. PLoS One *10*, e0146021.

Rodriguez-Fraticelli, A.E., Wolock, S.L., Weinreb, C.S., Panero, R., Patel, S.H., Jankovic, M., Sun, J., Calogero, R.A., Klein, A.M., and Camargo, F.D. (2018). Clonal analysis of lineage fate in native haematopoiesis. Nature *553*, 212.

Romano, O., Peano, C., Tagliazucchi, G.M., Petiti, L., Poletti, V., Cocchiarella, F., Rizzi, E., Severgnini, M., Cavazza, A., Rossi, C*., et al.* (2016). Transcriptional, epigenetic and retroviral signatures identify regulatory regions involved in hematopoietic lineage commitment. Sci Rep *6*, 24724.

Rosenberg, S.A., Aebersold, P., Cornetta, K., Kasid, A., Morgan, R.A., Moen, R., Karson, E.M., Lotze, M.T., Yang, J.C., Topalian, S.L*., et al.* (1990). Gene transfer into humans--immunotherapy of patients with advanced melanoma, using tumor-infiltrating lymphocytes modified by retroviral gene transduction. N Engl J Med *323*, 570-578.

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.

Sanjuan-Pla, A., Macaulay, I.C., Jensen, C.T., Woll, P.S., Luis, T.C., Mead, A., Moore, S., Carella, C., Matsuoka, S., Bouriez Jones, T*., et al.* (2013). Platelet-biased stem cells reside at the apex of the haematopoietic stem-cell hierarchy. Nature *502*, 232-236.

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.

Schofield, E.C., Carver, T., Achuthan, P., Freire-Pritchett, P., Spivakov, M., Todd, J.A., and Burren, O.S. (2016). CHiCP: a web-based tool for the integrative and interactive visualization of promoter capture Hi-C datasets. Bioinformatics *32*, 2511-2513.

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 U S A *110*, 12036-12041.

Shin, J.Y., Hu, W., Naramura, M., and Park, C.Y. (2014). High c-Kit expression identifies hematopoietic stem cells with impaired self-renewal and megakaryocytic bias. The Journal of Experimental Medicine *211*, 217-231.

Shlyueva, D., Stampfel, G., and Stark, A. (2014). Transcriptional enhancers: from properties to genome-wide predictions. Nature Reviews Genetics *15*, 272.

Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S., *et al.* (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res *15*, 1034-1050.

Siminovitch, L., McCulloch, E.A., and Till, J.E. (1963). THE DISTRIBUTION OF COLONY-FORMING CELLS AMONG SPLEEN COLONIES. Journal of cellular and comparative physiology *62*, 327-336.

Simonis, M., Klous, P., Splinter, E., Moshkin, Y., Willemsen, R., de Wit, E., van Steensel, B., and de Laat, W. (2006). Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture–on-chip (4C). Nature Genetics *38*, 1348.

Slatkin, M. (2008). Linkage disequilibrium — understanding the evolutionary past and mapping the medical future. Nature reviews Genetics *9*, 477-485.

Smale, S.T., and Kadonaga, J.T. (2003). The RNA polymerase II core promoter. Annual review of biochemistry *72*, 449-479.

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

Stein, S., Ott, M.G., Schultze-Strasser, S., Jauch, A., Burwinkel, B., Kinner, A., Schmidt, M., Kramer, A., Schwable, 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.

Stocking, C., Bergholz, U., Friel, J., Klingler, K., Wagener, T., Starke, C., Kitamura, T., Miyajima, A., and Ostertag, W. (1993). Distinct classes of factor-independent mutants can be isolated after retroviral mutagenesis of a human myeloid stem cell line. Growth factors (Chur, Switzerland) *8*, 197-209.

Sultana, T., Zamborlini, A., Cristofari, G., and Lesage, P. (2017). Integration site selection by retroviruses and transposable elements in eukaryotes. Nat Rev Genet *18*, 292-308.

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.

Sur, I., and Taipale, J. (2016). The role of enhancers in cancer. Nature Reviews Cancer *16*, 483.

Suzuki, T., Shen, H., Akagi, K., Morse, H.C., Malley, J.D., Naiman, D.Q., Jenkins, N.A., and Copeland, N.G. (2002). New genes involved in cancer identified by retroviral tagging. Nat Genet *32*, 166-174.

Temin, H.M., and Mizutani, S. (1970). RNA-dependent DNA polymerase in virions of Rous sarcoma virus. Nature *226*, 1211-1213.

Thomas, E.D., Lochte, H.L., Jr., Cannon, J.H., Sahler, O.D., and Ferrebee, J.W. (1959). Supralethal whole body irradiation and isologous marrow transplantation in man. J Clin Invest *38*, 1709-1716.

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

Thurman, R.E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M.T., Haugen, E., Sheffield, N.C., Stergachis, A.B., Wang, H., Vernot, B*., et al.* (2012). The accessible chromatin landscape of the human genome. Nature *489*, 75-82.

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

Varemo, L., Nielsen, J., and Nookaew, I. (2013). Enriching the gene set analysis of genome-wide data by incorporating directionality of gene expression and combining statistical hypotheses and methods. Nucleic Acids Res *41*, 4378-4391.

Velten, L., Haas, S.F., Raffel, S., Blaszkiewicz, S., Islam, S., Hennig, B.P., Hirche, C., Lutz, C., Buss, E.C., Nowak, D*., et al.* (2017). Human haematopoietic stem cell lineage commitment is a continuous process. Nat Cell Biol *19*, 271-281.

Weiss, R.A., and Vogt, P.K. (2011). 100 years of Rous sarcoma virus. The Journal of Experimental Medicine *208*, 2351-2355.

Weksberg, D.C., Chambers, S.M., Boles, N.C., and Goodell, M.A. (2008). CD150- side population cells represent a functionally distinct population of long-term hematopoietic stem cells. Blood *111*, 2444-2451.

Wilson, A., Laurenti, E., Oser, G., van der Wath, R.C., Blanco-Bose, W., Jaworski, M., Offner, S., Dunant, C.F., Eshkind, L., Bockamp, E., *et al.* (2008). Hematopoietic stem cells reversibly switch from dormancy to self-renewal during homeostasis and repair. Cell *135*, 1118-1129.

Wilson, N.K., Foster, S.D., Wang, X., Knezevic, K., Schutte, J., Kaimakis, P., Chilarska, P.M., Kinston, S., Ouwehand, W.H., Dzierzak, E., *et al.* (2010). Combinatorial transcriptional control in blood stem/progenitor cells: genome-wide analysis of ten major transcriptional regulators. Cell Stem Cell *7*, 532-544.

Wu, A.M., Till, J.E., Siminovitch, L., and McCulloch, E.A. (1967). A cytological study of the capacity for differentiation of normal hemopoietic colony-forming cells. J Cell Physiol *69*, 177-184.

Wu, C., Li, B., Lu, R., Koelle, S.J., Yang, Y., Jares, A., Krouse, A.E., Metzger, M., Liang, F., Lore, 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 *300*, 1749-1751.

Wu, X., Luke, B.T., and Burgess, S.M. (2006). Redefining the common insertion site. Virology *344*, 292-295.

Wünsche, P., Eckert, E.S.P., Holland-Letz, T., Paruzynski, A., Hotz-Wagenblatt, A., Fronza, R., Rath, T., Gil-Farina, I., Schmidt, M., von Kalle, C., *et al.* (2018). Mapping Active Gene-Regulatory Regions in Human Repopulating Long-Term HSCs. Cell Stem Cell *23*, 132-146.e139.

Yamada, Y., Warren, A.J., Dobson, C., Forster, A., Pannell, R., and Rabbitts, T.H. (1998). The T cell leukemia LIM protein Lmo2 is necessary for adult mouse hematopoiesis. Proc Natl Acad Sci U S A *95*, 3890-3895.

Yamamoto, R., Morita, Y., Ooehara, J., Hamanaka, S., Onodera, M., Rudolph, K.L., Ema, H., and Nakauchi, H. (2013). Clonal analysis unveils self-renewing lineage-restricted progenitors generated directly from hematopoietic stem cells. Cell *154*, 1112-1126.

Yu, V.W., Yusuf, R.Z., Oki, T., Wu, J., Saez, B., Wang, X., Cook, C., Baryawno, N., Ziller, M.J., Lee, E., *et al.* (2016). Epigenetic Memory Underlies Cell-Autonomous Heterogeneous Behavior of Hematopoietic Stem Cells. Cell *167*, 1310-1322 e1317.

Zhang, Y., Stehling-Sun, S., Lezon-Geyda, K., Juneja, S.C., Coillard, L., Chatterjee, G., Wuertzer, C.A., Camargo, F., and Perkins, A.S. (2011). PR-domain-containing Mds1-Evi1 is critical for long-term hematopoietic stem cell function. Blood *118*, 3853-3861.

Zhao, H., and Dean, A. (2004). An insulator blocks spreading of histone acetylation and interferes with RNA polymerase II transfer between an enhancer and gene. Nucleic Acids Res *32*, 4903-4919.

Zhou, V.W., Goren, A., and Bernstein, B.E. (2011). Charting histone modifications and the functional organization of mammalian genomes. Nat Rev Genet *12*, 7-18.

Zufferey, R., Dull, T., Mandel, R.J., Bukovsky, A., Quiroz, D., Naldini, L., and Trono, D. (1998). Self-inactivating lentivirus vector for safe and efficient in vivo gene delivery. J Virol *72*, 9873-9880.

7 CONTRIBUTIONS

The concept of this study was developed together with Elias Eckert, Claudia Ball, Friederike Herbst and Hanno Glimm. Several essential components of the barcoded lentiviral overexpression approach were designed and established with Elias Eckert, which mainly involves the design of the multiplexing cassette, the preparation of the multiplexing library and the testing of the multiplexing cassette. Conceptualization and implementation of statistical test was jointly performed with Tim Holland-Letz. Support with the HUSAR server was provided by Agnes Hotz-Wagenblatt. The CISOG plugin was kingly provided by Raffaele Fronza. The integration sited data from the WAS GT study was kindly provided by Manfred Schmidt, Anna Paruzynski, Raffaele Fronza, Tim Rath, Irene Gil-Farina, Christof von Kalle and Christoph Klein. The lentiviral packaging plasmids LV101, LV102 and LV103 were kindly provided by Luigi. The Cellecta shRNA screening library module 1 was kindly provided by Cellecta. Agnes Hotz-Wagenblatt and Alexander Jethwa provided expertise in the deconvolution of barcode reads. The ClusterCount function was kindly provided by Ulrich Abel. The implementation of the IS data into the CHiCP-browser was kindly performed by Olly Burren. The analysis of TAD boundaries and CTCF sites was performed with the help of Aleksander Jankowski. Mario Huerta helped with the analysis of the RNA-seq expression data and Sylvia Fessler, Nina Hofmann and Tim Kindinger supported the thesis with genotyping of mice and general lab support. Additional scientific input was provided through Nina Cabezas-Wallscheid and Alexander Jethwa as well as the thesis advisory committee assembled through Michael Boutros, Jan Lohmann and Hanno Glimm. Funding was acquired through Friederike Herbst, Claudia Ball and Hanno Glimm. The Helmholtz International Graduate School for Cancer Research kindly provided me with a fellowship for the first three years during my PhD studies.

8 Acknowledgements

First and foremost, I would like to thank Prof. Hanno Glimm for giving me the opportunity to work in his group on such an exciting project. Thank you for the perfect combination of trust, faith and freedom to peruse my ideas as well as the support and guidance to bring this challenging project to such a successful end.

Secondly, I would like to thank Dr. Friederike "Fee" Herbst. I have always admired your pervasive interest in science and your never-ending curiosity. Without your technical and theoretical advice and all the scientific discussion, this project would have not been the same.

I would also like to thank Prof. Michael Boutros and Prof. Jan Lohmann for being part of my thesis advisory- as well as the examination committee. Your valuable inputs and suggestions have significantly contributed to this thesis. Thanks also to Dr. Karin Müller-Decker for taking her time to be part of the examination committee.

Of course, I would like to thank Elias Eckert, who played a pivotal role in this thesis. I am happy that I got to know you and to have shared this experience with you. I am pretty confident that this project would have crushed me without your help, confidence and enthusiasm. A trouble shared is a trouble halved!

In line with this, I would like to thank Alexander Jethwa, the best office-buddy one could ask for. This project would not have not come to such a happy ending if it wasn't for our countless scientific discussions!

Of course, the start, the progress, the kickbacks, the suffering, the success and finally the end of a PhD journey is always and at any time a truly collaborative effort. I am grateful for having the honor to have worked with so many nice, helpful and supportive people in the past 4 ½ years. My greatest appreciation to Nina Hofmann and Tim Kindinger for their technical support and good laughter, Sylvia Fessler for keeping the lab together and organized, the AG Glimm depends on you! My greatest appreciation also to the countless people that crossed my path during the time at NCT and contributed to this project on so many different levels: Oksana Zavidij, Shayda Hemmati, Tonio Lang, Jens "Jörnsens" Langstein, Svenja Zielke, Florian Grünschläger, Klara Gießler, Taronish Dubash, Roland Ehrenberg, Sebastian Dieter, Martina Zowada, Karin Laaber, Mario Huerta, Lisa Pfitzer, Sophia Wild, Laura Schwarzmüller, Lino Möhrmann, Hendrik Strakerjahn, Erik “Erich” Schulz, Paula Codo, Stephanie Frank, Mariam Haffa, Christine Siegl, Mikolaj Slabicki, Tatjana Walther, Lena Wagner, Vineet Dalal and of course the unforgettable JP (Jianpeng Gao)! Surely, I would also like to thank the rest of the G100 and anyone else that I might have forgotten.

Of course, a big shout out to Kwang Seok Lee – WG buddy, science enthusiast, awesome friend, chef de cuisine, Mr. Butterfinger and so much more. Thank you for being there; you made this trip extra-special!

I would also like to thank Basti, who has welcomed me in Heidelberg from the very first minute and ever since stood by my side and supported me through all these turbulent 6 ½ years. The same holds true for the Heidelberg climbing community, which has always been my haven in stormy weathers. You are the sweetest bunch of people I ever came across.

Of course, also a big thanks to Hannah! Although our paths only crossed recently, I appreciate all your support and love throughout the last year. Thank you for taking so much weight of my shoulders during these busy times!

Last but not least, I would like to express my deepest gratitude to my family! You raised me to the person I am today, without you I would not be where I am, without you I would not write these lines and with you on my side, I always know where to go. My special gratitude goes to my dad, who sparked my interest in science already as a little kid. I miss you and I wish you could witness this moment.

9 APPENDIX

9.1 List of figures

Figure 1 | **Hierarchical organization of the murine and human hematopoietic system.** 5

Figure 2 | **Classical and modern models of the hematopoietic tree.** 7

Figure 3 | **Model of clonal dynamics after HSPC transplantation.** 8

Figure 4 | **Simplified view of the transcription initiation.** 10

Figure 5 | **Gene regulation through enhancer mediated transcription.** 12

Figure 6 | **Different modes of action of TADs in genome regulation.** 14

Figure 7 | **TAD and TAD boundary properties including enrichment of various genomic features.** 15

Figure 8 | **Overview of Roadmap Epigenomics 18-state expanded model ChromHMM chromatin states.** 19

Figure 9 | **Cell and nuclear entry paths of LV and γRV.** 24

Figure 10 | **Specific chromatin states are deterministic of the location of viral ISs (Legend continued on next page).** 25

Figure 11 | **Overview of retroviral insertional mutagenesis mechanisms.** 27

Figure 12 | **Genetic correction of diseased hematopoietic stem cells and subsequent monitoring of the patients in a clinical gene therapy trial.** 32

Figure 13 | **Schematic overview of genetic loci containing γRV ISs and terminology/parameters used for the characterization of clusters and statistic filtering for potential regulatory genes.** 33

Figure 14 | **Top 100 largest CIS and their associated gene with indicated proportion of every patient.** 34

Figure 15 | **Expression pattern of candidate genes.** 36

Figure 16 | **Lentiviral-mediated candidate gene overexpression.** 38

Figure 17 | **Detailed representation of the HT-seq barcode cassette with multiplexing PCR step.** 39

Figure 18 | **Matrix experiment to evaluate any potential index bleeding (Legend continued on next page).** 40

Figure 19 | **Evaluation of index bleeding.** 41

Figure 20 | **Relation between construct size and virus titer.** 42

Figure 21 | **Assessment of GFP and BC stability over time *in vitro*.** 43

Figure 22 | **Serial replating of transduced LSK-SLAM cells in semisolid medium (Legend continued on next page).** 46

Figure 23 | **CellTrace assay for the detection of changes in proliferation upon GOI OE (Legend continued on next page).** 48

Figure 24 | **Relative proportions of read counts for each gene (mean of two BCs) per mouse over time.** 51

Figure 25 | **Relative proportions of GFP control BCs per mouse over time, related to Figure 24.** 52

Figure 26 | **Wilcoxon-test for all possible barcode combinations.** 53

Figure 27 | **ISs mainly cluster around protein-coding gene TSS and histone marks associated with active promoters or enhancers** 56

Figure 28 **| Positive association matrices from patient 1, 2, 4, 5, 8, and 9 indicate a switch from short-term to long-term hematopoiesis after 6-20 months.** 58

Figure 29 | **Similarity between IS pattern is highest among patients** 60

Figure 30 | **Rainfall plots of ISs from WAS patients and CD34+ cells visualize commonalities and difference between the two datasets (Legend continued on next page).** 62

Figure 31 **| Mean cluster dimensions [bp] and sizes of patient and *in vitro* IS datasets as well as early and late patient ISs are highly similar.** 64

Figure 32 **| Statistical analysis of occurrence of ISs over time highlights known leukemogenic drivers.** 65

Figure 33 | **Conceptual outline.** 65

Figure 34 | **Correlation between WAS ISs and gene expression in 13 primary cell types.** 68

Figure 35 | **Signature ATAC-seq signal intensity at ISs from patients is highest in HSCs, while ISs from CD34$^+$ cells mostly enrich for CMP specific ATAC-seq peaks.** 69

Figure 36 | **ATAC-seq signal intensity at ISs from two independent studies is also highest in HSCs after transplantation. (Legend continued on next page)** 70

Figure 37 | “**Circular plots showing loci of reported HSPC regulators with CHi-C interaction and WAS IS.”** 72

Figure 38 | “**Gene set enrichment of differentially tagged genes in WAS patients vs. un-selected ISs from CD34$^+$ cells highlight long-term HSC specific genes (Legend continued on next page).”** 74

Figure 39 | “**Differential IS patterning at MYC enhancer correlates with selective activity of enhancer modules in HSCs and progenitor populations.”** 76

Figure 40 | **”Integration of long-range interaction data links ISs to enhancers and disease variants of their putative target gene.”** 78

Figure 41 | “**p-values of GWAS SNP enrichment at sites of ISs along with categorical enrichment across IS datasets.”** 80

Figure 42 | “**Global phastCons and phyloP conservation scores at ISs […]” point towards functional roles of IS-tagged regulatory regions.** 82

Figure 43 | **TAD boundaries but not CTCF sites show significantly more ISs than expected by chance** 83

Figure 44 | **Schematic representation showing the concept and strategy of mapping the gene-regulatory regions in human repopulating long-term HSCs.** 84

Figure 45 | **Effects of insufficient library representation during several rounds of passaging.** 88

9.2 List of tables

Table 1 | **Overview of protein-coding candidate genes in this study** 35

Table 2 | **Overview of performed transplantation experiments** 50

Table 3 | **Overview of γRV ISs and complementary datasets used in this study. Taken from Wünsche et al. (2018)** 55

Table 4 | **Overview of ISs per patient, separated into total, early and late sequenced ISs according to indicated switch. Taken from Wünsche et al. (2018)** 59

Table 5 | **Comparison of purification, cultivation and transduction between different studies** 67

Table 6 | **Overview of technical equipment and devices** 103

Table 7 | **Commercial kits** 104

Table 8 | **Reagents** 104

Table 9 | **Consumables** 106

Table 10 | **Plasmids** 107

Table 11 | **Westernblot Antibodies** 107

Table 12 | **Antibodies used for sorting of LSK and LSK-SLAM cells** 108

Table 13 | **Antibodies used for sorting of HSCs** 108

Table 14 | **Lineage cocktail used for the HSC sort staining** 108

Table 15 | **Antibodies used for detection or sort of progenitor cells** 109

Table 16 | **Antibodies used for analysis of HSCs with presents of GFP$^+$ cells** 109

Table 17 | **Antibodies used for analysis and sort of differentiated lymphoid and myeloid cells** 109

Table 18 | **Genotyping primers for B6.Cg-Gt(ROSA)26Sor tm1(rtTA*M2)Jae/J mice** 110

Table 19 | **Primers for amplifying the barcode cassette from the pRSI9 cellecta library and introducing the ClaI restriction enzyme cutsite** 110

Table 20 | **qPCR primers for endogenous expression of candidate genes** 111

Table 21 | **qPCR primers for expression of codon optimized candidate genes** 112

Table 22 | **Barcoding sequencing primers** 113

Table 23 | **High throughput multiplexing primers** 114

Table 24 | **Cellecta barcode sequences** 115

Table 25 | **Mouse strains** 116

Table 26 | **Software** 116

Table 27 | **R packages and Unix programs** 117

Table 28 | **Master Mix composition and PCR conditions for genotyping PCR of B6.Cg-Gt(ROSA)26Sor tm1(rtTA*M2)Jae/J mice** 118

Table 29 | **Master Mix composition and PCR conditions for qPCR reactions** 119

Table 30 | **Reagents and concentrations for the production blotting buffer as well as the resolving and stacking gel for Western blot** 121

Table 31 | **Volumes used for resolving and stacking gel** 122

Table 32 | **1st PCR during nested PCR for barcode amplification** 123

Table 33 | **2nd PCR during nested PCR for barcode amplification** 124

Table 34 | **µg plasmids used per 15cm dish.** 125

Table 35 | **Surface marker combinations for the identification of immunophenotypic cell populations** 128

9.3 List of equations

Equation 1 | **Calculation of virus titer** 126

Equation 2 | **Calculation of odds ratios between patient samples** 132

Equation 3 | **Calculation of area under the curve** 133

9.4 Abbreviations

Abbreviation	Complete term
3C	chromatin conformation capture
4C	chromosome conformation capture-on-chip
5C	chromosome conformation capture carbon copy
AB	Antibody
ADA-SCID	Adenosin-Desaminase - Severe combined immunodeficiency
AML	Acute myeloid leukemia
AmpR	Ampicillin resistance
ATAC-seq	Assay for Transposase-Accessible Chromatin with high throughput sequencing
AUC	Area under the curve
B cell	B lymphocyte
B-ALL	B-cell acute lymphoblastic leukemia
BC	Barcode
BENC	Blood enhancer cluster
BM	Bone marrow
bp	Basepair
CD	Cluster of differentiation
CD4 T cell	CD4^{+} T lymphocyte
CD8 T cell	CD8^{+} T lymphocyte
cDNA	Complement DNA
CFU	Colony forming unit
CHi-C	Capture Hi-C
ChIP-seq	Chromatin immunoprecipitation combined with sequencing
ChromHMM	Machine-learning multivariate hidden Markov model
CI	Confidence interval
CIS	Common integration site
CLOUD-HSPC	Continuum of low-primed undifferentiated hematopoietic stem and progenitor cells
CLP	Common lymphoid progenitor
CMP	Common myeloid progenitor
CMV	Cytomegalovirus promoter
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
CRM	*cis*-regulatory module
CTCF	CCCTC-binding factor
d	Days

DCE	Downstream core element
DHS	DNase I-hypersensitive sites
DKFZ	German Cancer Research Centre
DNA	Deoxyribonucleic acid
DOX	Doxycycline
Env	Envelope
Ery	Erythrocytes
ET domain	Extraterminal domain
ETP	Earliest thymic progenitors
FACS	Fluorescence-activated cell sorting
FBS	Fetal bovine serum
FDR	False discovery rate
FSC	Forward scatter
gag	Group specific antigen
G-CSF	Granulocyte-colony stimulating factor
gDNA	Genomic DNA
GFP	Green fluorescent protein
GMP	Granulocyte–monocyte progenitor
GOI	Gene of interest
GT	Gene therapy
GTF	general transcription factor
GvHD	Graft versus Host Disease
GWAS	Genome-wide association studies
HBS	Hank’s balanced salt solution
Hi-C	High-throughput sequencing chromosome conformation capture
HIV	Human Immunodeficiency virus
HLA	human leukocyte antigen
HSC	Hematopoietic stem cell
HSPC	Hematopoietic stem and progenitor cell
HSZ	Hämatopoetischen Stammzelle
HT-seq	High- throughput sequencing
IL	Interleukin
IN	Integrase protein
Inr	Initiator element
IRES	Internal ribosomal entry site
IS	Integration site
IU	Infectious units
kb	Kilobases

LAM-PCR	Linear amplification mediated polymerase chain reaction
LD	linkage disequilibrium
Lin$^-$	Lineage negative
lincRNA	Long intergenic non-coding RNA
LMPP	Lymphoid-primed multipotent progenitor
LSK	Lineage negative, Sca-1 positive, c-Kit positive
LSK-SLAM	LSK, CD48 negative, CD150 positive
LT-HSC	Long-term hematopoietic stem cell
LTR	Long terminal repeat
LV	Lentivirus
MACS	Magnetic cell separation
MEP	Megakaryocyte–erythroid progenitor
miRNA	micro RNA
MLV	Murine leukemia virus
MOI	Multiplicity of infection
Mono	Monocyte
MPP	Multipotent progenitor
mRNA	Messenger RNA
MTW	Motif ten element
ncRNA	Non-coding RNA
NCT	National Centre for Tumor Diseases
ND	Not disclosed / Not determined
NGS	Next generation sequencing
NK cell	Natural killer cell
NPC	Nuclear pore complex
NSG	NOD-Scid Il2γc$^{-/-}$
nt	Nucleotide
OE	Overexpression
PB	Peripheral blood
PBS	Phosphate-buffered saline
PCA	Principle component analysis
PCR	Polymerase chain reaction
PIC	Pre-integration complex
PrIC	Pre-initiation complex
qPCR	Quantitative PCR
RBC	Relative barcode count
RLD	Relative lineage differentiation
RNA	Ribonucleic acid

RNAP I, II, III	RNA polymerases I, II, III
RRE	Rev-responsive element
RSV	Rous sarcoma virus
RT	Room temperature
RT-PCR	Reverse transcription polymerase chain reaction
rtTA	reverse tetracycline-controlled trans-activator
SCF	Stem cell factor
SD	Standard deviation
SEM	Standard error of mean
shRNA	Short hairpin RNA
SINE	Short interspersed element
SNP	Single nucleotide polymorphisms
SSC	Side scatter
STARR-seq	self-transcribing active regulatory region sequencing
ST-HSC	Short-term hematopoietic stem cell
T7	T7-RNA-polymerase promotor
TAD	Topologically associated domains
T-ALL	T-lymphoblastic leukemia/lymphoma
TBP	TATA-binding protein
TF	Transcription factor
TFIIB	Transcription factor II B
TFIIH	Transcription factor II H
TPM	Transcript per million
TPO	Thrombopoietin
TSS	Transcription start site
TX	Transplantation experiment
UbiC	Ubiquitin C promoter
UTR	Untranslated region
WAS	Wiskott-Aldrich-Syndrome
WPRE	Woodchuck Hepatitis Virus Posttranscriptional Regulatory Element
X-SCID	X-linked severe combined immunodeficiency
γRV	γ-retrovirus

9.5 ClusterCount function for R

```
function (pos.is, d) {

  # ClusterCount

  # Counts the number of clusters contained in a vector of IS positions.
  # ClusterCount has to be performed on single chromosomes.
  # Clusters comprising identical IS are counted as well.
  # A cluster must contain at least 2 IS.

  # Args:
  # pos.is:      A vector of IS positions
  # d:           Threshold for differences in neighboring IS position used for
  #                 defining clusters (unit=b)

  # Returns:
  # A list with five elements:
  # 1, the number of clusters
  # 2, the number of IS contained in clusters.
  # 3, a vector which elements are the length (number of IS) of each cluster
  # 4, a vector which elements are the dimension of each cluster
  # 5, a list containing vectors representing all clusters in pos.is

  #-----------------------------------------------------------------------------
  # Check input:
  if (length(pos.is) == 0) stop("The IS vector must have length >0.")

  # Initialize:
  pos.is <- sort(pos.is)
  nis <- length(pos.is)
  cluster.new <- pos.is[1]
  list.cluster <- list()
  length.new <-1
  # length of present cluster (at this point, clusters may contain only one IS)
  # N.B. "length means number of IS!
  v.length <- numeric(0)
  v.dimension <- numeric(0)

  if (nis > 1) {
    # Loop over IS positions (This code generates all clusters except the last one):

    for (i in (2:nis))  {
      if ((pos.is[i] - pos.is[i - 1]) <= d) {
        length.new <- length.new + 1
        cluster.new <- c(cluster.new, pos.is[i])
      }
      else {
        if (length.new > 1) list.cluster <- c(list.cluster, list(cluster.new))
        v.dimension <- c(v.dimension, max(cluster.new) - min(cluster.new))
        cluster.new <- pos.is[i]
        # pos.is[i] starts a new cluster, possibly of length 1
        v.length <- c(v.length, length.new)
        length.new <- 1
      }
    }
  }

  # Add last cluster:
  v.length <- c(v.length, length.new)
  v.dimension <- c(v.dimension, max(cluster.new) - min(cluster.new))
  v.dimension <- v.dimension[v.length > 1]
  v.length <- v.length[v.length > 1]
  if (length.new > 1) list.cluster <- c(list.cluster, list(cluster.new))

  # Evaluation:
  n.cluster <- length(v.length)
  n.is.cluster <- sum(v.length)

  return(list(n.cluster, n.is.cluster, v.length, v.dimension, list.cluster))
}
```

9.6 Condon optimized cDNA sequences of candidate genes

9.6.1 Amica1

```
GGATCCGCCACCATGCTGTGCCTGCTGAAGCTGATCGTGATCCCCGTGATCCTGGCCCCCGTGGGATATCCTCAGGGA
CTGCCTGGCCTGACCGTGTCCTCTCCACAGCTGAGAGTGCACGTGGGCGAGAGCGTGCTGATGGGCTGTGTGGTGCAG
AGAACCGAAGAGAAGCACGTGGACAGAGTGGACTGGCTGTTCAGCAAGGACAAGGACGACGCCAGCGAGTACGTGCTG
TTCTACTACAGCAACCTGAGCGTGCCCACCGGCAGATTCCAGAACAGATCTCACCTCGTGGGCGACACCTTCCACAAC
GACGGAAGCCTGCTGCTGCAGGACGTGCAGAAGGCTGACGAGGGCATCTACACATGCGAGATCAGACTGAAGAACGAG
AGCATGGTCATGAAGAAACCCGTGGAACTGTGGGTGCTGCCCGAGGAACCCAAGGACCTGCGCGTCAGAGTGGGCGAT
ACCACCCAGATGAGATGCAGCATCCAGTCCACCGAAGAAAAACGCGTGACCAAAGTGAACTGGATGTTCTCCAGCGGC
AGCCACACCGAAGAGGAAACCGTGCTGAGCTACGACTCCAACATGAGAAGCGGCAAGTTCCAGAGCCTGGGCAGGTTC
AGAAACAGGGTGGACCTGACCGGCGACATCAGCAGAAACGACGGCAGCATCAAGCTGCAGACCGTGAAAGAGAGCGAC
CAGGGAATCTACACCTGTAGCATCTACGTGGGCAAGCTGGAAAGCAGAAAGACCATCGTGCTGCACGTGGTGCAGGAC
GAGTTCCAGCGGACCATCAGCCCTACCCCCCCTACAGATAAGGGCCAGCAGGGCATCCTGAACGGCAATCAGCTCGTG
ATCATCGTGGGAATCGTGTGTGCCACCTTTCTGCTGCTGCCCGTGCTGATCCTGATCGTGAAGAAAGCCAAGTGGAAC
AAGAGCAGCGTGTCCAGCATGGCCAGCGTGAAGTCCCTGGAAAACAAAGAGAAGATCAACCCCGAGAAGCACATCTAC
AGCAGCATCACCACCTGGGAGACAACCGAGAGAGGCATCAGCGGCGAGTCCGAGGGAACCTACATGACAATGAACCCC
GTGTGGCCCAGCAGCCCCAAGGCTAGTTCTCTCGTGCGAAGCAGCGTGCGGAGCAAGTGACCTGCAGG
```

BamHI cutsite | Kozac sequence | Start codon | Stop codon | Sbf1 cutsite

9.6.2 Ccnd3

```
GGATCCGCCACCATGGAACTGCTGTGTTGCGAGGGCACCAGACACGCCCCTAGAGCTGGCCCTGATCCTAGACTGCTG
GGCGACCAGAGAGTGCTGCAGAGCCTGCTGAGACTGGAAGAAAGATACGTGCCCAGAGCCAGCTACTTCCAGTGCGTG
CAGAAAGAAATCAAGCCCCACATGAGAAAGATGCTGGCCTACTGGATGCTGGAAGTGTGCGAGGAACAGAGATGCGAA
GAGGACGTGTTCCCCCTGGCCATGAACTACCTGGACAGATACCTGAGCTGCGTGCCCACCAGAAAGGCCCAGCTGCAG
CTGCTGGGCACCGTGTGTCTGCTGCTGGCCTCCAAGCTGAGAGAGACAACCCCCCTGACCATCGAGAAGCTGTGCATC
TACACCGACCAGGCCGTGGCCCCTTGGCAGCTGAGGGAATGGGAAGTGCTGGTGCTGGGAAAGCTGAAGTGGGACCTG
GCCGCCGTGATCGCCCACGATTTTCTGGCTCTGATTCTGCACAGACTGAGCCTGCCCAGCGACAGACAGGCCCTCGTG
AAGAAGCACGCCCAGACCTTTCTGGCCCTGTGCGCCACCGACTACACCTTCGCCATGTACCCCCCCAGCATGATCGCC
ACCGGCTCTATCGGAGCAGCCGTGCTGGGACTGGGCGCCTGTTCTATGTCTGCCGACGAGCTGACCGAGCTGCTGGCT
GGCATCACAGGCACCGAGGTGGACTGCCTGAGAGCCTGCCAGGAACAGATCGAGGCCGCCCTGAGAGAGTCTCTGAGA
GAGGCCGCTCAGACCGCCCCAAGCCCTGTGCCTAAAGCTCCTAGAGGCAGCAGCTCCCAGGGCCCTAGCCAGACCAGC
ACACCTACAGACGTGACCGCCATCCACCTGTGACCTGCAGG
```

BamHI cutsite | Kozac sequence | Start codon | Stop codon | Sbf1 cutsite

9.6.3 Evl

GGATCCGCCACCATGAGCGAGCAGAGCATCTGTCAGGCCAGAGCCAGCGTGATGGTGTACGACGACCAGCAAGAAA
TGGGTGCCCATCAAGCCCGGCCAGCAGGGCTTCAGCAGAATCAACATCTACCACAACACCGCCAGCAGCACCTTCAGA
GTCGTGGGCGTGAAGCTGCAGGACCAGCAGGTCGTGATCAACTACAGCATCGTGAAGGGCCTGAAGTACAACCAGGCC
ACCCCCACCTTTCACCAGTGGCGGGATGCCAGACAGGTGTACGGCCTGAACTTCGCCAGCAAAGAGGAAGCCACCACC
TTCAGCAACGCCATGCTGTTCGCCCTGAACATCATGAACAGCCAGGAAGGCGGCCCTAGCACCCAGAGACAGGTGCAG
AACGGCCCCAGCCCCGAGGAAATGGACATCCAGCGGCGCCAAGTGATGGAACAGCAGCACAGACAGGAAAGCCTGGAA
AGAAGAATCAGCGCCACCGGCCCCATCCTGCCACCTGGACATCCTAGCTCTGCCGCCAGCACCACACTGAGCTGTAGC
GGACCTCCTCCCCCTCCACCACCACCTGTGCCTCCACCTCCAACAGGCAGCACACCTCCCCCACCCCCCCCACTGCCA
GCAGGCGGAGCACAGGGAACAAACCACGACGAGTCTAGCGCCAGCGGCCTGGCTGCTGCTCTGGCTGGCGCAAAGCTG
AGAAGAGTGCAGAGGCCTGAGGACGCTAGCGGCGGCAGTAGCCCTTCTGGCACAAGCAAGAGCGACGCCAACAGAGCC
TCTTCCGGCGGAGGCGGAGGGGGACTGATGGAAGAGATGAACAAGCTGCTGGCCAAGAGAAGAAAGGCCGCCTCCCAG
ACCGACAAGCCCGCCGACAGAAAAGAGGACGAGAGCCAGACCGAGGACCCCAGCACATCTCCTAGCCCTGGCACCAGA
GCCACCAGCCAGCCTCCAAACTCTAGCGAGGCCGGCAGAAAGCCCTGGGAGAGAAGCAACAGCGTGGAAAAGCCCGTG
TCCAGCCTGCTGAGCAGAACCCCTAGCGTGGCCAAGTCCCCTGAGGCCAAGAGCCCTCTGCAGTCCCAGCCTCACAGC
AGAGTGAAGCCTGCCGGCTCCGTGAACGACGTGGGACTGGATGCCCTGGACCTGGACAGAATGAAGCAGGAAATTCTG
GAAGAGGTCGTGCGCGAGCTGCACAAAGTGAAAGAGGAAATCATCGACGCCATCCGGCAGGAACTGAGCGGCATCAGC
ACAACCTGACCTGCAGG

BamHI cutsite | Kozac sequence | Start codon | Stop codon | Sbf1 cutsite

9.6.4 Fbxl18

GGATCCGCCACCATGGCTAGCAGCGGCGAGGACATCAGCAACGACGACGATGACATGCACCCTGCCGCCGCTGGAATG
GCCGACGGAGTGCATCTGCTGGGCTTCAGCGACGAGATCCTGCTGCACATCCTGAGCCACGTGCCCAGCACCGACCTG
ATCCTGAACGTGCGGAGAACCTGCAGAAAGCTGGCCGCCCTGTGCCTGGACAAGAGCCTGATCCACACCGTGCTGCTG
CAGAAGGACTACCAGGCCAGCGAGGACAAAGTGCGGCAGCTCGTGAAAGAGATCGGCAGAGAGATCCAGCAGCTGAGC
ATGGCCGGCTGCTACTGGCTGCCTGGCTCTACCGTGGAACACGTGGCCAGATGCAGATCCCTCGTGAAAGTGAACCTG
AGCGGCTGCCACCTGACCAGCCTGAGACTGAGCAAGATGCTGAGCGCCCTGCAGCACCTGAGAAGCCTGGCCATCGAT
GTGTCCCCAGGCTTCGACGCCAGCCAGCTGTCTAGCGAGTGCAAGGCCACCCTGAGCAGAGTGCGCGAGCTGAAGCAG
ACCCTGTTCACCCCTAGCTACGGCGTGGTGCCTTGCTGCACCAGCCTGGAAAAGCTGCTGCTGTACTTTGAGATCCTG
GACAGAACCAGAGAGGGCGCCATCCTGTCCGGCCAGCTGATGGTGGGACAGAGCAACGTGCCCCACTACCAGAACCTG
AGAGTGTTCTACGCCAGACTGGCCCCTGGCTACATCAACCAGGAAGTCGTGCGGCTGTACCTGGCCGTGCTGAGCGAC
AGAACCCCCCAGAATCTGCACGCCTTTCTGATCAGCGTGCCCGGCAGCTTCGCTGAGTCTGGCGCCACAAAGAACCTG
CTGGACAGCATGGCCAGAAACGTGGTGCTGGACGCTCTGCAGCTGCCCAAGTCTTGGCTGAACGGCAGCTCCCTGCTG
CAGCACATGAAGTTCAACAACCCCTTCTACTTCAGCTTCAGCCGGTGCACCCTGTCTGGCGGACACCTGATTCAGCAA
GTGATCAACGGCGGCAAGGACCTGAGATCCCTGGCCTCCCTGAACCTGTCCGGATGCGTGCACTGTCTGAGCCCCGAC
AGCCTGCTGAGAAAGGCCGAGGACGACATCGACAGCAGCATCCTGGAAACCCTGGTGGCCAGCTGCTGCAACCTGAGA
CACCTGAATCTGTCTGCCGCCCACCACCACAGCTCTGAGGGACTGGGCAGACACCTGTGTCAGCTGCTGGCCAGACTG
AGACATCTGCGGAGCCTGAGCCTGCCCGTGTGTTCTGTGGCCGACTCTGCCCCTAGAGCCGATAGAGCACCAGCCCAG
CCTGCCATGCACGCTGTGCCTAGAGGCTTCGGCAAGAAAGTGCGCGTGGGCGTGCAGTCCTGCCCCAGCCCTTTTAGC
GGACAGGCTTGCCCTCAGCCCAGCTCCGTGTTTTGGTCCCTGCTGAAGAATCTGCCCTTCCTGGAACACCTGGAACTG
ATCGGCAGCAACTTCAGCAGCGCCATGCCTAGAAACGAGCCCGCCATCAGAAACAGCCTGCCCCCTTGTAGCAGAGCC
CAGAGCGTGGGCGATTCTGAGGTGGCCGCTATCGGGCAGCTGGCTTTCCTGAGGCATCTGACCCTGGCCCAGCTGCCA
AGTGTGCTGACAGGCAGCGGCCTCGTGAACATCGGCCTGCAGTGTCAGCAGCTGCGGTCCCTGTCTCTGGCCAACCTG
GGCATGATGGGAAAGGTGGTGTACATGCCCGCCCTGTCCGACATGCTGAAGCACTGCAAGAGACTGAGGGACCTGAGG
CTGGAACAGCCTTACTTCAGCGCCAACGCCCAGTTCTTCCAGGCCCTGAGCCAGTGTCCTAGCCTGCAGAGACTGTGT
CTGGTGTCCAGAAGCGGCACCCTGCAGCCTGATGCTGTGCTGGCCTTCATGGCCCGGTGTCTGCAGGTCGTGATGTGC
CACCTGTTCACAGGCGAGAGCCTGGCTACCTGCAAAAGCCTGCAGCAGAGCCTGCTGCGGTCTTTCCAGGCCGAAAGA
CCCGCTCTGAACGTCGTGATCTTCCCACTGCTGCACGAGGGCCTGACCGACGTGATCAGAGATGTGCCCCTGGTGCAC
CTGGACGAGATCACACTGTTCAAGTCCAGAGTGGCCGAGGAACCCCCTAACCTGTGGTGGTAGCCTGCAGG

BamHI cutsite | Kozac sequence | Start codon | Stop codon | Sbf1 cutsite

9.6.5 Igf2bp2

```
GGATCCGCCACCATGATGAACAAGCTGTACATCGGCAACCTGAGCCCTGCCGTGACAGCCGACGATCTGAGACAGCTG
TTCGGCGACAGAAAGCTGCCCCTGGCTGGACAGGTGCTGCTGAAGTCTGGCTACGCCTTCGTGGACTACCCCGACCAG
AACTGGGCCATCAGAGCCATCGAGACACTGAGCGGCAAGGTGGAACTGCACGGCAAGATCATGGAAGTGGACTACAGC
GTGTCCAAGAAGCTGAGGTCCAGAAGAATCCAGATCCGGAACATCCCCCCACATCTGCAGTGGGAGGTGCTGGATGGA
CTGCTGGCCGAGTACGGCACCGTGGAAAACGTGGAACAAGTGAACACCGACACCGAGACAGCCGTCGTGAACGTGACC
TACATGACCAGAGAGGAAGCCAAGCTGGCTATCGAGAAGCTGTCCGGCCACCAGTTCGAGGACTACTCCTTCAAGATC
AGCTACATCCCCGACGAGGAAGTGTCCAGCCCCAGCCCTCCTCACAGAGCTAGAGAGCAGGGACACGGCCCTGGCAGC
AGCTCTCAGGCCAGACAGATCGACTTCCCACTGAGAATCCTGGTGCCCACCCAGTTCGTGGGCGCCATCATCGGCAAA
GAGGGCCTGACCATCAAGAACATCACCAAGCAGACCCAGAGCAGAGTGGACATCCACAGAAAAGAGAACAGCGGCGCT
GCCGAGAAGCCCGTGACAATCCACGCTACCCCTGAGGGCACAAGCGAGGCCTGCAGAATGATCCTGGAAATCATGCAG
AAAGAGGCCGACGAGACAAAGCTGGCCGAAGAGGTGCCCCTGAAGATCCTGGCCCACAACGGCTTCGTGGGCAGACTG
ATCGGAAAAGAAGGCCGGAACCTGAAGAAGATCGAGCACGAGACAGGCACCAAGATTACAATCAGCTCTCTGCAGGAC
CTGAGCATCTACAACCCCGAGAGAACCATCACCGTGCGGGGCACCATCGAGGCTTGTGCCAACGCCGAGATCGAGATC
ATGAAGAAACTGAGAGAGGCCTTCGAGAACGACATGCTGGCCGTGAACCAGCAGGCCAACCTGATCCCAGGCCTGAAC
CTGTCTGCCCTGGGCATCTTCAGCACCGGCCTGTCAGTGCTGCCACCTCCTGCTGGACCTAGAGGCGTGCCACCTAGC
CCTCCCTACCACCCTTTCGCCACACACAGCGGCTACTTCAGCTCCCTGTACCCCCACCACCACTTCGGCCCATTCCCT
CACCACCACAGCTACCCCGAGCAGGAAACCGTGTCTCTGTTCATCCCAACCCAGGCCGTGGGAGCTATCATTGGCAAG
AAGGGCGCCCACATCAAGCAGCTGGCCAGATTCGCTGGCGCCTCCATCAAGATCGCCCCTGCTGAAGGCCCTGACGTG
TCCGAGAGAATGGTCATCATCACCGGCCCTCCCGAGGCTCAGTTCAAGGCTCAGGGCAGAATCTTCGGCAAGCTGAAA
GAGGAAAACTTCTTCAACCCCAAAGAAGAAGTGAAGCTGGAAGCCCACATCCGGGTGCCAAGCAGCACAGCCGGAAGA
GTGATTGGCAAGGGCGGCAAGACCGTGAACGAGCTGCAGAACCTGACCAGCGCCGAAGTGATCGTGCCCAGGGACCAG
ACCCCTGACGAGAATGAGGAAGTGATTGTGCGGATCATCGGCCACTTTTTCGCCAGCCAGACCGCCCAGAGAAAGATC
CGCGAGATCGTGCAGCAAGTGAAGCAGCAGGAACAGAGATACCCCCAGGGCGTGGCCCCCCAGAGATCCAAATGACCT
GCAGG
```

BamHI cutsite | Kozac sequence | Start codon | Stop codon | Sbf1 cutsite

9.6.6 Irf2bpl

```
GGATCCGCCACCATGTCTGCTGCCCAGGTGTCCAGCAGCAGAAGGCAGAGCTGCTACCTGTGCGACCTGCCCAGAATG
CCCTGGGCCATGATCTGGGACTTCAGCGAGCCTGTGTGCAGAGGCTGCGTGAACTACGAGGGCGCCGACAGAATCGAG
TTCGTGATCGAGACAGCCAGACAGCTGAAGAGAGCCCACGGCTGCTTCCAGGACGGCAGATCTCCTGGACCTCCTCCA
CCCGTGGGCGTGAAAACAGTGGCCCTGTCTGCCAAAGAGGCCGCTGCTGCAGCTGCTGCCGCCCAACAACAACAACAA
CAACAACAGCAACAGCAGCAGCAGCTGAACCACGTGGACGGCAGCACAAAGCCTGCCGTGCTGGCTGCTCCTAGCGGC
CTGGAAAGATACGGCCTGTCTGCAGCCGCCGCTGCCGCCGCAGCAGCCGCTGCAGTGGAACAGAGAAGCAGATTCGAG
TACCCCCCTCCCCCTGTGTCCCTGGGCTCTAGCTCTCACGCTGCCAGACTGCCTAACGGCCTGGGCGGACCTAACGGC
TTCCCTAAGCCTGCCCCTGAGGAAGGCCCTCCCGAGCTGAACAGACAGAGCCCCAACTCTAGCAGCGCCGCCACAAGC
GTGGCCAGCAGAAGAGGCACACACTCCGGCCTCGTGACCGGCCTGCCTAATCCTGGCGGAGGCGGAGGACCTCAGCTG
ACCGTGCCTCCAAATCTGCTGCCTCAGACCCTGCTGAACGGCCCTGCTTCTGCAGCTGTGCTGCCTCCTCCTCATGGA
CTGGGCGGCTCTAGAGGCCCTCCTACACCAGCTCCTCCAGGCGCACCTGGCGGACCTGCTTGTCTGGGAGGACCACCT
GGCGTGTCCGCCACAGTGTCTAGCGCCCCTAGCAGCACAAGCAGCACCGTGGCTGAAGTGGGCGTGGGCGCTGCTGGC
AAAAGACCTGGCTCTGTGTCCTCCACCGACCAGGAAAGAGAGCTGAAAGAAAAGCAGAGAAACGCCGAGGCCCTGGCC
GAGCTGTCTGAGAGCCTGAGAAACAGAGCCGAGGAATGGGCCAACAAGCCCAAGATGGTGCGAGACACACTGCTGACA
CTGGCCGGCTGCACCCCTTACGAAGTGCGGTTCAAGAAGGACCACAGCCTGCTGGGCAGAGTGTTCGCCTTCGACGCC
GTGTCCAAGCCCGGCATGGACTACGAGCTGAAGCTGTTCATCGAGTATCCCACCGGCTCCGGCAACGTGTACTCTAGC
GCTTCTGGGGTGGCCAAGCAGATGTACCAGGACTGCATGAAGGACTTCGGCAGAGGCCTGAGCAGCGGCTTCAAGTAC
CTGGAATACGAGAAGAAGCACGGCTCTGGCGATTGGAGACTGCTGGGCGACCTGCTGCCAGAGGCTGTGCGGTTCTTC
AAAGAAGGCGTGCCAGGCGCCGATATGCTGCCCCAGCCTTACCTGGACGCCAGCTGCCCTATGCTGCCTACCGCTCTG
GTGTCCCTGAGCAGAGCCCCTTCTGCTCCTCCTGGAACAGGCGCTCTGCCACCAGCTGCACCTACTGGAAGGGGAGCC
GCCAGCTCCCTGAGAAAGAGAAAGGCCAGCCCCGAGCCTCCTGACTCTGCCGAGTCTGCTCTGAAGCTGGGCGAGGAA
CAGCAGAGACAGCAGTGGATGGCCAACCAGTCTGAGGCCCTGAAGCTGACCATGAGCGCTGGCGGATTTGCCGCCCCT
GGACATTCTGCAGGCGGACCTCCACCCCCTCCACCTCCACTGGGACCTCACTCCAACAGAACCACCCCCCCTGAGAGC
GCCCCTCAGAACGGACCTTCTCCTATGGCCGCCCTGATGAGCGTGGCCGACACACTGGGAACAGCCCACAGCCCTAAG
GACGGCTCTAGCGTGCACAGCACAACAGCCAGCGCCAGAAGAAACAGCTCCAGCCCAGTGTCCCCTGCCTCTGTGCCT
GGACAGAGAAGGCTGGCCTCCAGAAACGGCGACCTGAATCTGCAGGTGGCCCCACCACCACCTAGCGCTCACCCTGGA
ATGGACCAGGTGCACCCCCAGAACATCCCCGACAGCCCCATGGCTAACAGCGGCCCTCTGTGCTGCACCATCTGCCAC
GAGAGACTGGAAGATACCCACTTCGTGCAGTGCCCCAGCGTGCCCAGCCACAAGTTCTGCTTCCCTTGCAGCAGAGAG
TCCATCAAGGCTCAGGGCGCCACCGGCGAGGTGTACTGTCCTTCTGGCGAGAAGTGCCCCCTCGTGGGCAGCAATGTG
CCTTGGGCTTTCATGCAGGGCGAGATCGCCACAATCCTGGCCGGCGACGTGAAAGTGAAGAAAGAGCGGGACCCCTGA
CCTGCAGG
```

BamHI cutsite | Kozac sequence | Start codon | Stop codon | Sbf1 cutsite

9.6.7 Lair 1 long isoform

```
GGATCCGCCACCATGTCTCTGCACCCCGTGATCCTGCTGGTGCTGGTGCTGTGTCTGGGCTGGAAGATCAACACCCAG
GAAGGCAGCCTGCCCGACATCACCATCTTCCCCAACAGCAGCCTGATGATCAGCCAGGGCACCTTCGTGACCGTCGTG
TGCAGCTACAGCGACAAGCACGACCTGTACAACATGGTGCGACTGGAAAAGGACGGCAGCACCTTCATGGAAAAGAGC
ACCGAGCCCTACAAGACCGAGGACGAGTTCGAGATCGGCCCCGTGAACGAGACAATCACCGGCCACTACAGCTGCATC
TACAGCAAGGGCATCACTTGGAGCGAGAGAAGCAAGACCCTGGAACTGAAAGTGATCAAAGAAAACGTGATCCAGACC
CCTGCCCCTGGCCCTACCAGCGACACAAGCTGGCTGAAAACCTACAGCATCTACATCTTCACCGTGGTGTCCGTGATC
TTCCTGCTGTGCCTGAGCGCCCTGCTGTTCTGCTTCCTGAGACACAGACAGAAGAAGCAGGGCCTGCCCAACAACAAG
AGACAGCAGCAGAGGCCCGAGGAAAGACTGAACCTGGCCACCAACGGCCTGGAAATGACCCCCGACATCGTGGCCGAC
GACAGACTGCCTGAGGACAGATGGACCGAGACATGGACACCCGTGGCCGGCGATCTGCAGGAAGTGACCTACATTCAG
CTGGACCACCACAGCCTGACCCAGAGGGCTGTGGGCGCTGTGACAAGCCAGAGCACAGACATGGCCGAGAGCAGCACC
TACGCCGCCATCATCAGACACTGACCTGCAGG
```

BamHI cutsite | Kozac sequence | Start codon | Stop codon | Sbf1 cutsite

9.6.8 Lair1 short isoform

GGATCCGCCACCATGTCTCTGCACCCCGTGATCCTGCTGGTGCTGGTGCTGTGTCTGGGCTGGAAGATCAACACCCAG
GAAGATACCAGCTGGCTGAAAACCTACAGCATCTACATCTTCACCGTGGTGTCCGTGATCTTCCTGCTGTGCCTGAGC
GCCCTGCTGTTCTGCTTCCTGAGACACAGACAGAAGAAGCAGGGCCTGCCCAACAACAAGAGACAGCAGCAGAGGCCC
GAGGAAAGACTGAACCTGGCCACCAACGGCCTGGAAATGACCCCCGACATCGTGGCCGACGACAGACTGCCTGAGGAC
AGATGGACCGAGACATGGACACCCGTGGCCGGCGATCTGCAGGAAGTGACCTACATTCAGCTGGACCACCACAGCCTG
ACCCAGAGGGCTGTGGGCGCTGTGACAAGCCAGAGCACAGACATGGCCGAGAGCAGCACCTACGCCGCCATCATCAGA
CACTGACCTGCAGG

BamHI cutsite | Kozac sequence | Start codon | Stop codon | Sbf1 cutsite

9.6.9 Lrrc33

GGATCCGCCACCATGCAGGAACCTCTGGAAACCGGCAGCATCGAGAGCAGCGGCACAGGCAACGTGGTGGTGTCTCAC
CAGAGGGCCGTGCCCGAGATGGAATTCCCTCCTCTGTGGCTGTGCCTGGGCTTCCACTTCCTGATCGTGGAATGGCGC
AGCGGCCCTGGCACTGCTACAGCTGCTTCTCAGGGCGGCTGCAAGGTGGTGGATGGCGTGGCAGACTGCAGAGGCCTG
AACCTGGCCTCTGTGCCTAGCAGCCTGCCCCCCACAGCAGAATGCTGATCCTGGACGCCAACCCCCTGAAGGACCTG
TGGAACCACTCTCTGCAGGCCTACCCCAGACTGGAAAACCTGAGCCTGCACAGCTGCCACCTGGACAGAATCAGCCAC
TACGCCTTCAGAGAGCAGGGCCACCTGAGAAACCTGGTGCTGGCCGACAACAGACTGAGCGAGAACTACAAAGAGAGC
GCCGCTGCCCTGCACACCCTGCTGGGACTGAGAAGGCTGGACCTGAGCGGCAACAGCCTGACCGAGGATATGGCCGCA
CTGATGCTGCAGAACCTGAGCAGCCTGGAAGTGGTGTCCCTGGCCAGAAACACCCTGATGAGACTGGACGACAGCATC
TTCGAGGGCCTGGAACACCTGGTGGAACTGGACCTGCAGAGGAACTACATCTTTGAGATCGAGGGCGGAGCCTTCGAC
GGCCTGACAGAACTGCGGAGACTGAATCTGGCCTACAACAACCTGCCTTGCATCGTGGACTTTAGCCTGACCCAGCTG
AGATTCCTGAACGTGTCCTACAATATCCTGGAATGGTTCCTGGCTGCCAGAGAAGAGGTGGCCTTCGAGCTGGAAATC
CTGGACCTGTCCCACAACCAGCTGCTGTTCTTCCCACTGCTGCCCCAGTGCGGCAAGCTGCATACACTGCTGCTGCAG
GACAACAACATGGGCTTCTACAGAGAGCTGTACAACACCAGCAGCCCCCAGGAAATGGTGGCCCAGTTTCTGCTGGTG
GACGGCAACGTGACCAACATCACCACCGTGAACCTGTGGGAGGAATTCAGCAGCAGCGACCTGTCCGCCCTGCGGTTC
CTGGACATGAGCCAGAACCAGTTCAGACATCTGCCCGACGGCTTTCTGAAGAAAACCCCCAGCCTGAGCCACCTGAAT
CTGAACCAGAACTGCCTGAAAATGCTGCACATCCGCGAGCACGAGCCTCCAGGCGCTCTGACAGAGCTGGATCTGAGC
CACAATCAGCTGGCCGAGCTGCACCTGGCCCCTGGACTGACAGGCTCTCTGAGGAACCTGAGAGTGTTCAACCTGTCC
TCTAATCAGCTGCTGGGCGTGCCCACCGGCCTGTTCGATAACGCCAGCAGCATCACCACAATCGACATGTCTCACAAT
CAGATCAGCCTGTGCCCCCAGATGGTGCCCGTGGATTGGGAGGGACCTCCTAGCTGCGTGGACTTCAGAAACATGGGC
AGCCTGAGATCCCTGTCCCTGGACGGCTGTGGCCTGAAGGCTCTGCAGGACTGCCCATTTCAAGGCACCTCCCTGACC
CATCTGGATCTGTCCAGCAACTGGGGCGTGCTGAACGGCTCCATCAGCCCTCTGTGGGCCGTGGCTCCTACACTGCAG
GTGCTGAGCCTGAGAGATGTGGGCCTGGGATCTGGCGCCGCTGAGATGGACTTCTCCGCCTTCGGCAACCTGAGGGCC
CTGGATCTGTCTGGCAACTCCCTGACCAGCTTCCCCAAGTTCAAGGGCTCCCTGGCCCTGAGGACCCTGGACCTGAGA
AGAAACTCTCTGACCGCCCTGCCCCAGAGGGTGGTGTCAGAACAGCCTCTGAGAGGACTGCAGACCATCTACCTGTCT
CAGAACCCCTACGACTGCTGCGGCGTGGAAGGATGGGGAGCACTGCAGCAGCACTTCAAGACCGTGGCCGACCTGAGC
ATGGTCACCTGTAACCTGTCTAGCAAGATCGTGCGGGTGGTGGAACTGCCCGAGGGACTGCCTCAGGGCTGCAAGTGG
GAACAGGTGGACACCGGACTGTTCTATCTGGTGCTGATTCTGCCCTCCTGTCTGACCCTGCTGGTGGCCTGTACCGTG
GTGTTCCTGACCTTCAAGAAACCCCTGCTGCAAGTGATCAAGTCCAGATGCCACTGGTCCAGCATCTACTGACCTGCA
GG

BamHI cutsite | Kozac sequence | Start codon | Stop codon | Sbf1 cutsite

9.6.10 Mbnl1

```
GGATCCGCCACCATGGCCGTGTCTGTGACCCCCATCAGAGACACCAAGTGGCTGACCCTGGAAGTGTGCAGAGAGTTC
CAGAGAGGCACCTGTAGCAGACCCGACACCGAGTGCAAGTTCGCCCACCCCAGCAAGAGCTGCCAGGTGGAAAACGGC
AGAGTGATCGCCTGCTTCGACAGCCTGAAGGGCAGATGCAGCAGAGAGAACTGCAAGTACCTGCACCCCCCTCCCCAC
CTGAAAACCCAGCTGGAAATCAACGGCCGGAACAACCTGATCCAGCAGAAAAACATGGCTATGCTGGCCCAGCAGATG
CAGCTGGCCAACGCCATGATGCCTGGCGCTCCTCTGCAGCCCGTGCCCATGTTTTCTGTGGCCCCTAGCCTGGCCACA
AGCGCCTCTGCTGCCTTCAACCCTTACCTGGGCCCTGTGTCCCCTTCCCTGGTGCCTGCTGAGATCCTGCCTACCGCC
CCCATGCTCGTGACAGGCAATCCTGGCGTGCCAGTGCCAGCTGCTGCCGCTGCTGCTGCCCAGAAACTGATGAGAACC
GACAGACTGGAAGTGTGCCGCGAGTACCAGCGGGGCAACTGCAACAGAGGCGAGAACGACTGCAGATTCGCTCACCCC
GCCGACAGCACCATGATCGACACCAACGACAACACCGTGACCGTGTGCATGGACTACATCAAGGGCCGGTGCTCCCGC
GAAAAGTGCAAGTACTTCCACCCTCCCGCCCATCTGCAGGCCAAGATCAAGGCCGCTCAGTACCAAGTGAACCAGGCC
GCTGCAGCCCAGGCTGCTGCTACTGCTGCAGCTATGGGCATCCCTCAGGCCGTGCTGCCCCCCCTGCCTAAAAGACCT
GCCCTGGAAAAGACCAACGGCGCCACCGCCGTGTTCAACACCGGCATCTTCCAGTACCAGCAGGCCCTGGCCAACATG
CAGCTGCAGCAGCACACCGCCTTTCTGCCCCCTGGCAGCATCCTGTGTATGACCCCTGCCACCAGCGTGGTGCCTATG
GTGCATGGCGCTACCCCAGCCACAGTGTCTGCCGCCACAACAAGCGCCACCTCTGTGCCTTTCGCCGCCACCGCTACA
GCCAACCAGATCCCCATCATCAGCGCCGAGCACCTGACCAGCCACAAATACGTGACCCAGATGTGACCTGCAGG
```

BamHI cutsite | Kozac sequence | Start codon | Stop codon | Sbf1 cutsite

9.6.11 Ninj2

```
GGATCCGCCACCATGGAAAGCGACAGAGAGACAATCCATCTGCAGCACAGACACAGCATGAGAGGCGGCAACCAGAGA
ATCGACCTGAACTTCTACGCCACCAAGAAAAGCGTGGCCGAGAGCATGCTGGACGTGGCCCTGTTCATGAGCAACGCC
ATGAGACTGAAGTCCGTGCTGCAGCAGGGCCCCTTCGCCGAGTACTACACCACCCTCGTGACCCTGATCATCGTGTCC
CTGCTGCTGCAGGTCGTGATCTCTCTGCTGCTGGTGTTTATCGCCATCCTGAACCTGAACGAGGTGGAAAACCAGAGG
CACCTGAACAAGCTGAACAACGCCGCCACAATCCTGGTGTTCATCACCGTCGTGATCAACATCTTCATCACAGCCTTC
GGCGCCCACCACGCCGCCTCTATGGCTGCCAGAACAAGCAGCAACCCAATCTGACCTGCAGG
```

BamHI cutsite | Kozac sequence | Start codon | Stop codon | Sbf1 cutsite

9.6.12 Nrip1

```
GGATCCGCCACCATGACACACGGCGAGGAACTGGGCTCTGACGTGCACCAGGACAGCATCGTGCTGACCTACCTGGAA
GGCCTGCTGATGCACCAGGCTGCTGGCGGCTCTGGCACCGCCATCAACAAGAAGTCTGCCGGCCACAAAGAAGAGGAC
CAGAACTTCAACCTGAGCGGCAGCGCCTTCCCCAGCTGTCAGTCTAACGGCCCTACCGTGTCCACCCAGACCTACCAG
GGCAGCGGAATGCTGCACCTGAAGAAGGCCAGACTGCTGCAGAGCAGCGAGGACTGGAACGCCGCCAAGAGAAAGAGA
CTGAGCGACTCCATCGTGAACCTGAACGTGAAGAAAGAGGCCCTGCTGGCCGGCATGGTGGACTCTGTGCCTAAGGGC
AAGCAGGACTCCACACTGCTGGCCTCCCTGCTGCAGTCCTTCAGCAGCAGACTGCAGACCGTGGCCCTGAGCCAGCAG
ATCAGACAGAGCCTGAAAGAGCAGGGCTACGCCCTGTCCCACGAGTCCCTGAAGGTGGAAAAGGACCTGAGATGCTAC
GGCGTGGCCAGCTCCCACCTGAAAACCCTGCTGAAGAAGTCCAAGACCAAGGATCAGAAGTCCGGCCCCACCCTGCCT
GACGTGACCCCCAACCTGATCAGAGACAGCTTCGTGGAAAGCAGCCACCCCGCCGTGGGCCAGTCTGGCACAAAAGTG
ATGAGCGAGCCCCTGAGCTGCGCCGCTAGACTGCAGGCTGTGGCTTCCATGGTGGAAAAAAGAGCCAGCCCTGCCGCC
AGCCCCAAGCCTTCTGTGGCTTGTTCTCAGCTGGCACTGCTGCTGTCCAGCGAGGCCCATCTGCAGCAGTACAGCAGA
GAGCACGCCCTGAAAACACAGAACGCCCACCAGGTGGCCAGCGAGAGGCTGGCTGCTATGGCTAGGCTGCAGGAAAAC
GGCCAGAAAGACGTGGGCTCCAGCCAGCTGTCTAAGGGCGTGTCCGGCCACCTGAACGGACAGGCTAGAGCCCTGCCT
GCCTCTAAGCTGGTGGCCAACAAGAACAACGCCGCTACCTTCCAGAGCCCCATGGGCGTGGTGCCTAGCAGCCCTAAG
AACACCAGCTACAAGAACAGCCTGGAACGGAACAACCTGAAGCAGGCTGCCAACAACAGCCTGCTGCTGCATCTGCTG
AAGTCTCAGACCATCCCCACCCCCATGAACGGCCACAGCCAGAACGAGAGGGCCAGCAGCTTCGAGAGCAGCACCCCT
ACCACCATCGACGAGTACAGCGACAACAACCCCAGCTTCACCGACGACAGCAGCGGCGACGAGTCCAGCTACTCCAAC
TGCGTGCCCATCGACCTGTCCTGCAAGCACAGAATCGAGAAGCCCGAGGCCGAGAGGCCCGTGTCCCTGGAAAACCTG
ACCCAGAGCCTGCTGAACACCTGGGACCCCAAGATCCCCGGCGTGGACATCAAAGAGGATCAGGACACCAGCACCAAC
AGCAAGCTGAACAGCCACCAGAAAGTGACTCTGCTGCAGCTGCTGCTGGGCCACAAGAGCGAGGAAACCGTGGAAAGA
AACGCCTCCCCCCAGGACATCCACAGCGACGGCACAAAGTTCAGCCCCCAGAACTACACCAGAACCAGCGTGATCGAG
AGCCCCTCCACCAACAGAACCACCCCTGTGTCCACACCCCCCCTGTACACAGCCTCTCAGGCCGAGTCCCCTATCAAC
CTGTCCCAGCACTCCCTCGTGATCAAGTGGAACAGCCCCCCCTACGCCTGTAGCACCCCTGCTTCCAAGCTGACCAAC
ACCGCCCCCAGCCACCTGATGGACCTGACCAAGGGCAAAGAGAGCCAGGCCGAGAAGCCTGCCCCTTCTGAAGGCGCC
CAGAACAGCGCCACATTCAGCGCCTCAAAGCTGCTGCAGAACCTGGCCCAGTGTGGGCTGCAGAGTTCTGGCCCTGGC
GAAGAACAGCGGCCTTGCAAACAGCTGCTGAGCGGAAACCCCGACAAGCCCCTGGGCCTGATCGACAGACTGAATAGC
CCCCTGCTGAGCAACAAGACAAACGCTGCCGAGGAAAGCAAGGCCTTCAGCTCCCAGCCAGCCGGACCTGAACCTGGA
CTGCCTGGATGCGAGATCGAGAACCTGCTGGAAAGACGGACCGTGCTGCAGCTGCTGCTGGGAAACAGCAGCAAGGGC
AAGAATGAGAAGAAAGAAAAGACCCCCGCCAGGGACGAGGCCCCTCAGGAACATTCTGAGAGGGCCGCCAACGAGCAG
ATCCTGATGGTCAAGATCAAGTCCGAGCCCTGCGACGACTTCCAGACCCACAACACCAACCTGCCCCTGAACCACGAC
GCCAAGAGCGCCCCATTTCTGGGCGTGACACCCGCCATCCACAGAAGCACAGCTGCCCTGCCAGTGTCCGAGGACTTC
AAGTCTGAGCCTGCCAGCCCTCAGGACTTCAGCTTCAGCAAGAACGGCCTGCTGTCCCGGCTGCTGAGACAGAACCAG
GAAAGCTACCCTGCCGACGAGCAGGACAAGTCCCACAGAAACAGCGAGCTGCCTACCCTGGAATCCAAGAACATCTGC
ATGGTGCCCAAGAAGCGGAAGCTGTACACCGAGCCTCTGGAAAATCCCTTCAAGAAGATGAAGAACACCGCCGTGGAC
ACCGCCAACCACCACTCTGGACCAGAGGTGCTGTACGGATCACTGCTGCACCAGGAAGAACTGAAGTTCAGCAGAAAC
GAGCTGGACTACAAGTACCCAGCCGGCCACTCTAGCGCCTCTGACGGCGATCACAGAAGCTGGGCCAGAGAGTCCAAG
AGCTTCAACGTGCTGAAACAGCTGCTGCTGTCCGAGAACTGCGTGCGGGATCTGAGCCCCCACAGATCCGACAGCGTG
CCCGACACCAAGAAGAAGGGCCACAAAAACAACGCTCCCGGCAGCAAGCCCGAGTTCGGCATCTCTTCCCTGAATGGC
CTGATGTACAGCTCCCCTCAGCCCGGCTCTTGCGTGACCGACCACAGAACCTTCAGCTACCCCGGAATGGTCAAAACC
CCCCTGAGCCCTCCATTCCCCGAGCACCTGGGATGCGTGGGAAGCAGACCAGAGCCCGGACTGCTGAACGGCTGTTCT
GTGCCTGGCGAGAAGGGCCCCATCAAATGGGTCATCGCCGACATGGACAAGAACGAGTACGAGAAGGACAGCCCCAGA
CTGACAAAGACCAACCCCATCCTGTACTACATGCTGCAGAAAGGCGGCGGAAACAGCGTGACCACCCAGGAAACCCAG
GACAAGGACATTTGGAGAGAGCCCGCCTCCGCCGAGAGCCTGTCTCAAGTGACCGTGAAAGAGGAACTGCTGCCAGCC
GCCGAGACAAAGGCCAGCTTCTTTAACCTGAGAAGCCCCTACAACAGCCACATGGGCAACAACGCCAGCAGACCCCAC
AGCACAAACGGCGAGGTGTACGGGCTGCTGGGGAACGCCCTGACCATCAAGAAAGAATCCGAGTGACCTGCAGG
```

BamHI cutsite | Kozac sequence | Start codon | Stop codon | Sbf1 cutsite

9.6.13 Plcb4

```
GGATCCGCCACCATGGCCAAGCCCTACGAGTTCAACTGGCAGAAAGAGGTGCCCAGCTTTCTGCAGGAAGGCGCCGTG
TTCGACAGATACGAAGAGGAATCCTTCGTGTTCGAGCCCAACTGCCTGTTCAAGGTGGACGAGTTCGGATTCTTCCTG
ACCTGGAAGTCCGAGGGCAAAGAGGGCCAGGTGCTGGAATGCAGCCTGATCAACAGCATCAGACAGGCCGCCATCCCC
AAGGACCCCAAGATCCTGGCTGCCCTGGAAGCTGTGGGCAAGAGCGAGAACGATCTGGAAGGCAGAATCCTGTGCGTG
TGCAGCGGCACCGACCTCGTGAACATCGGCTTCACCTACATGGTGGCCGAGAACCCCGAAGTGACCAAGCAGTGGGTG
GAAGGCCTGAGATCCATCATCCACAACTTCAGAGCCAACAACGTGTCCCCCATGACCTGCCTGAAGAAACACTGGATG
AAGCTGGCCTTCCTGACAAACACCACCGGCAAGATCCCCGTGCGGAGCATCACCAGAACATTCGCCAGCGGCAAGACA
GAGAAAGTGATCTTCCAGGCCCTGAAAGAGCTGGGCCTGCCCTCCGGCAAGAACGACGAGATCGAGCCTGCCGCCTTC
ACATACGAGAAGTTCTACGAGCTGACCCAGAAGATCTGCCCCAGAACCGACATCGAGGATCTGTTCAAGAAGATCAAC
GGCGACAAGACCGACTACCTGACCGTGGATCAGCTGGTGTCCTTCCTGAACGAGCACCAGAGGGACCCCAGACTGAAC
GAGATCCTGTTCCCATTCTACGACGCCAAGAGAGCCATGCAGATCATCGAGATGTACGAGCCCGACGAGGAACTGAAG
AAGAAGGGCCTGATCAGCTCCGACGGCTTCTGCAGATACCTGATGAGCGACGAGAACGCCCCCGTGTTCCTGGACAGA
CTGGAACTGTACCAGGAAATGGACCACCCCCTGGCCCACTACTTCATCAGCAGCAGCCACAACACCTACCTGACAGGC
AGACAGTTCGGCGGCAAGAGCAGCGTGGAAATGTACAGACAGGTGCTGCTGGCCGGCTGCAGATGCGTGGAACTGGAC
TGTTGGGACGGCAAGGGCGAGGACCAGGAACCCATCATCACACACGGCAAGGCCATGTGCACCGACATCCTGTTTAAG
GACGTGATCCAGGCCATCAAAGAAACCGCCTTCGTGACCAGCGAGTACCCCGTGATCCTGAGCTTCGAGAACCACTGC
AGCAAGTACCAGCAGTACAAGATGAGCAAGTACTGCGAGGACCTGTTCGGCGACCTGCTGCTGAAGCAGGCCCTGGAA
TCCCACCCTCTGGAACCCGGCAGACCTCTGCCTAGCCCCAACGACCTGAAGAGAAAGATCCTGATCAAGAACAAGCGG
CTGAAGCCCGAGGTGGAAAAGAAGCAGCTGGAAGCCCTGAAGTCCATGATGGAAGCCGGCGAGTCTGCCGCCCCTGCC
AGCATTCTGGAAGATGACAACGAGGAAGAGATCGAGAGCGCCGACCAGGAAGAGGAAGCCCACCCCGAGTACAAGTTC
GGCAACGAGCTGTCCGCCGACGACTACAGCCACAAAGAAGCCGTGGCCAACAGCGTGAAGAAAGGCCTCGTGACCGTG
GAAGATGAGCAGGCCTGGATGGCCAGCTACAAATACGTGGGCGCCACCACCAACATCCACCCCTACCTGAGCACCATG
ATCAACTACGCCCAGCCCGTGAAGTTCCAGGGCTTTCACGTGGCCGAGGAAAGAAACATCCACTACAACATGAGCAGC
TTCAACGAGTCCGTGGGCCTGGGCTACCTGAAAACCCACGCCATCGAGTTCGTGAACTACAACAAGAGACAGATGAGC
CGGATCTACCCCAAGGGCGGCAGGGTGGACAGCAGCAACTATATGCCCCAGATCTTTTGGAACGCTGGCTGCCAGATG
GTGTCCCTGAACTACCAGACACCCGACCTGGCCATGCAGCTGAACCAGGGCAAGTTCGAGTACAACGGCAGCTGCGGC
TACCTGCTGAAACCCGACTTCATGAGAAGGCCCGACAGAACCTTCGACCCCTTCAGCGAGACACCCGTGGATGGCGTG
ATCGCCGCCACATGTAGCGTGCAAGTGATCAGCGGCCAGTTCCTGAGCGACAAGAAAATCGGCACCTACGTGGAAGTG
GATATGTACGGCCTGCCCACCGACACCATCAGAAAAGAATTCAGAACCCGGATGGTCATGAACAACGGCCTGAACCCC
GTGTACAACGAAGAGTCTTTCGTGTTCCGCAAAGTGATCCTGCCAGACCTGGCCGTGCTGAGAATCGCCGTGTACGAC
GACAACAACAAGCTGATCGGCCAGAGAATCCTGCCCCTGGACGGACTGCAGGCTGGCTACAGACACATCAGCCTGAGA
AACGAGGGCAACAAGCCCCTGAGCCTGCCTACCATCTTCTGCAACATCGTGCTGAAAACCTACGTGCCAGACGGCTTC
GGCGACATCGTGGACGCTCTGAGCGACCCTAAGAAGTTCCTGTCCATCACCGAGAAGCGGGCCGACCAGATGAGGGCC
ATGGGCATCGAGACATCCGATATCGCCGACGTGCCAAGCGACACCTCTAAGAACGACAAGAAGGGCAAGGCTAACCCC
GCCAAGGCCAACGTGACACCCCAGTCTAGCAGCGAGCTGAGGCCTACCACAACAGCCGCTCTGGGCTCTGGCCAGGAA
GCCAAGAAGGGAATCGAGCTGATCCCCCAAGTGCGGATTGAGGACCTGAAGCAGATGAAGGCCTATCTGAAGCACCTG
AAAAAGCAGCAGAAAGAACTGAACTCTCTGAAGAAAAAGCACGCCAAAGAACACAGCACCATGCAGAAGCTGCACTGC
ACCCAGGTGGACAAGATCGTGGCCCAGTACGACAAAGAGAAGTCCACCCACGAGAAGATTCTGGAAAAGGCCATGAAG
AAGAAAGGCGGCTCTAACTGCCTGGAAATCAAGAAAGAGACTGAGATCAAGATCCAGACCCTGACCACCGACCACAAG
AGCAAAGTGAAAGAAATCGTGGCTCAGCATACCAAAGAATGGAGCGAGATGATCAACACCCACAGCGCCGAGGAACAG
GAAATCAGGGACCTGCACCTGAGCCAGCAGTGCGAGCTGCTGAGAAAGCTGCTGATTAACGCCCACGAGCAGCAGACC
CAGCAGCTGAAGCTGTCCCACGACCGCGAGAGCAAAGAGATGCGGGCTCACCAGGCCAAGATCAGCATGGAAAACTCC
AAGGCCATCAGCCAGGACAAGAGCATTAAGAACAAGGCCGAGCGCGAGCGGAGAGTGCGCGAGCTGAACAGCTCCAAC
ACCAAAAAGTTTCTGGAAGAACGGAAGCGGCTGGCCATGAAGCAGTCCAAAGAGATGGACCAGCTGAAGAAGGTGCAG
CTGGAACACCTGGAATTTCTGGAAAAGCAGAACGAGCAGGCCAAAGAAATGCAGCAGATGGTCAAGCTGGAAGCCGAG
ATGGACAGACGGCCTGCTACCGTGGTGTAACCTGCAGG
```

BamHI cutsite | Kozac sequence | Start codon | Stop codon | Sbf1 cutsite

9.6.14 Prkcb

GGATCCGCCACCATGGCTGATCCTGCTGCTGGCCCTCCACCTAGCGAGGGCGAAGAAAGCACAGTCAGATTCGCCAGA
AAGGGCGCTCTGAGACAGAAAAACGTGCACGAAGTGAAGAACCACAAGTTCACCGCCCGGTTCTTCAAGCAGCCCACC
TTCTGCAGCCACTGCACCGACTTCATCTGGGGCTTCGGCAAGCAGGGATTCCAGTGCCAAGTGTGCTGCTTCGTGGTG
CACAAGAGATGCCACGAGTTCGTGACCTTCAGCTGCCCTGGCGCCGATAAGGGCCCTGCCTCTGACGACCCTAGAAGC
AAGCACAAGTTTAAGATCCACACCTACAGCTCCCCAACCTTCTGTGACCACTGCGGCAGCCTGCTGTACGGCCTGATC
CACCAGGGCATGAAGTGCGACACCTGTATGATGAACGTGCACAAACGCTGCGTGATGAATGTGCCCAGCCTGTGCGGC
ACCGACCACACCGAGAGAAGAGGCAGAATCTACATCCAGGCCCACATCGACCGCGAGGTGCTGATTGTGGTCGTGCGG
GACGCCAAGAACCTGGTGCCCATGGACCCTAACGGCCTGAGCGACCCCTACGTGAAGCTGAAGCTGATCCCCGACCCC
AAGAGCGAGAGCAAGCAGAAAACAAAGACCATCAAGTGCAGCCTGAACCCCGAGTGGAACGAGACATTCAGATTCCAG
CTGAAAGAGAGCGACAAGGACAGACGGCTGAGCGTGGAAATCTGGGACTGGGACCTGACCAGCAGAAACGACTTCATG
GGCAGCCTGAGCTTCGGCATCAGCGAGCTGCAGAAAGCTGGCGTGGACGGCTGGTTCAAGCTGCTGTCTCAGGAAGAG
GGCGAGTACTTCAACGTGCCCGTGCCTCCTGAGGGCAGCGAGGGAAACGAGGAACTGAGGCAGAAGTTCGAGAGAGCC
AAGATCGGCCAGGGCACCAAGGCCCCCGAGGAAAAGACCGCCAACACCATCAGCAAGTTCGACAACAACGGCAACAGG
GACAGAATGAAGCTGACAGACTTCAATTTCCTGATGGTGCTGGGCAAGGGCTCCTTCGGCAAAGTGATGCTGAGCGAG
AGAAAGGGCACCGACGAGCTGTACGCCGTGAAGATCCTGAAGAAAGACGTCGTGATCCAGGACGACGACGTGGAATGT
ACCATGGTGGAAAAGAGAGTGCTGGCTCTGCCCGGCAAGCCCCCATTCCTGACACAGCTGCACAGCTGCTTCCAGACC
ATGGACAGACTGTACTTCGTGATGGAATACGTGAACGGCGGCGACCTGATGTACCACATCCAGCAAGTGGGCAGATTC
AAAGAACCCCACGCCGTGTTCTACGCCGCCGAGATCGCTATCGGCCTGTTCTTCCTGCAAAGCAAGGGCATCATCTAC
AGGGACCTGAAGCTGGACAACGTGATGCTGGACAGCGAGGGCCACATCAAGATCGCCGACTTCGGCATGTGCAAAGAG
AACATCTGGGACGGCGTGACCACCAAGACATTCTGCGGCACCCCCGACTATATCGCCCCCGAGATCATTGCCTACCAG
CCCTACGGCAAGTCCGTGGATTGGTGGGCTTTCGGCGTGCTGCTGTATGAGATGCTGGCTGGCCAGGCCCCTTTCGAG
GGCGAGGATGAGGATGAGCTGTTCCAGAGCATCATGGAACACAACGTGGCCTACCCTAAGAGCATGAGCAAAGAAGCC
GTGGCCATCTGCAAGGGCCTGATGACCAAGCACCCCGGCAAGAGACTGGGCTGTGGACCCGAAGGCGAGAGAGATATC
AAAGAGCACGCCTTCTTCCGGTACATCGACTGGGAGAAGCTGGAACGGAAAGAGATCCAGCCCCCCTACAAGCCCAAG
GCCTGTGGCAGAAACGCCGAGAACTTCGACAGATTCTTCACCAGACACCCCCCCGTGCTGACCCCCCCAGATCAGGAA
GTGATCAGAAACATCGACCAGAGCGAGTTCGAGGGCTTTAGCTTCGTGAACAGCGAGTTCCTGAAGCCTGAAGTGAAG
TCCTGACCTGCAGG

BamHI cutsite | Kozac sequence | Start codon | Stop codon | Sbf1 cutsite

9.6.15 Slx4ip long isoform

GGATCCGCCACCATGGCCAGCAAGAAATTCGCCGTGAAGTGCGGCAACTTCGCCGTGCTGGTGGACCTGCATGTGCTG
CCTCAGGGCAGCAACAGAGACAGCAGCTGGTTCAGCGAGCAGAAAAAAGAGGAAGTGTGCCTGCTGCTGAAAGAGACA
ATCGACAGCCGCGTGAAAGAATACGTGGGCATCTACAAGCAGAGAAAGCCCAGCAGCGCCGAGTTCACCAGAAGCAGC
CCTCTGAGCCTGAAGGGCTACGGCTTCCAGATCACCGCCTACTTTCTGAAGAGAGGCATCCATCTGCACTGCATCCAG
AACAGCCAGAACACCGAGCTGAGAGTGTTCCCCGAGAGATTCGTCGTGTGCGTGTCCCAGCTGGCCTTCGGCCACGAT
ATCTGGGCCAACCAGAACGAGAAGTCCACCAAGAAAGCCCTGCACGGCGTGTCCGACTACTTCCCTGAGTGTGCCGAG
AGCAGCCCTAGCCCTGGCACCAAGCTGAAGAGAAACGCCCTGAAAGAAATCGTGCGGAGGACCAAGAGCAAGGGCACC
GACGTGTCCAAGCCTCAGCCTAGCGGAGATCTCGTGGGCAGATCCAGCGACAGCGTGATCACCGTGGTGCCTTGGAGA
AGAGATGCCAGCGCCATCCTGCTGAGCGAGTCTGTGGGACAGGCCCAGGACGATATCAGAGCCGCCAAGAGCCACCAG
GAACTGCCCGTGCAGAAACTGGAAAATGTGTCCCAGACCCAGCCCGGCGACACCAGATCACAGCAGCAGCTGCATCCT
GGCGAGTGGCTGAAAACCGGCCTGCTGTCTAGAAGCCCCGCCTACAACTACGAGAGCGCCAGCCCAGGCCCTAAGCAG
TCTCTGAGAGCCGCTAAGACCCAGCAGAAGCACAGAAACTGCGGCAGCGTGGAAGATTGCGACCACCGCAGAAGAGTG
TCCCTGGGCAACGAGGGACTGGTGCCTGAGGACGCTGACCGCGAGAGATCTACAGCTGTGCGGGTGCTGCCTGCCCTG
GAACTGTCTGATCCTGGACTGCTGCTGAAGCAGGACCTGGCCAAGGCCAAGGCTAAAGAGGAACTGCACGCCCTGGAA
AACCTGAGCAGCAGACACCTCGTGACCAACAACCCAGGCCAGGCCCAGCAGAGCGATAGCGCTGCTATCACCGAGCAG
CTGGCCACAGATCAGGGCGGACCTAGCAAGAAGAGAAAGAAGCTGCAGAGCTACAACAGAGGCTGCAGCGGCAAGAAG
AACTGACCTGCAGG

BamHI cutsite | Kozac sequence | Start codon | Stop codon | Sbf1 cutsite

9.6.16 Slx4ip short isoform

```
GGATCCGCCACCATGGCCAGCAAGAAATTCGCCGTGAAGTGCGGCAACTTCGCCGTGCTGGTGGACCTGCATGTGCTG
CCTCAGGGCAGCAACAGAGACAGCAGCTGGTTCAGCGAGCAGAAAAAAGAAGTGATGGCCTTCAGGTCCCAGCTGATC
TCCAGCAGAGAGGGCTACACCTTCACCGTGTCCAGAACCCCCAGAATCCTGACCAAGAAAGCCCTGCACGGCGTGTCC
GACTACTTCCCTGAGTGTGCCGAGAGCAGCCCTAGCCCTGGCACCAAGCTGAAGAGAAACGCCCTGAAAGAAATCGTG
CGGAGGACCAAGAGCAAGGGCACCGACGTGTCCAAGCCTCAGCCTAGCGGAGATCTCGTGGGCAGATCCAGCGACAGC
GTGATCACCGTGGTGCCTTGGAGAAGAGATGCCAGCGCCATCCTGCTGAGCGAGTCTGTGGGACAGGCCCAGGACGAT
ATCAGAGCCGCCAAGAGCCACCAGGAACTGCCCGTGCAGAAACTGGAAAACGTGTCCCAGACCCAGCCCGGCGACACC
AGATCTCAGCAGCAGCTGCATCCTGGCGAGTGGCTGAAAACCGGCCTGCTGTCTAGAAGCCCCGCCTACAACTACGAG
AGCGCCAGCCCAGGCCCTAAGCAGTCTCTGAGAGCCGCTAAGACCCAGCAGAAGCACAGAAACTGCGGCAGCGTGGAA
GATTGCGACCACCGCAGAAGAGTGTCCCTGGGCAACGAGGGACTGGTGCCTGAGGACGCTGACCGCGAGAGATCTACA
GCTGTGCGGGTGCTGCCTGCCCTGGAACTGTCTGATCCTGGCCTGCTGCTGAAACAGGACCTGGCCAAGGCCAAGGCT
AAAGAGGAACTGCACGCCCTGGAAAACCTGAGCAGCAGACACCTCGTGACCAACAACCCAGGCCAGGCCCAGCAGAGC
GATAGCGCTGCTATCACAGAGCAGCTGGCCACCGATCAGGGCGGACCTAGCAAGAAGAGAAAGAAGCTGCAGAGCTAC
AACAGAGGCTGCAGCGGCAAGAAGAACTGACCTGCAGG
```

BamHI cutsite | Kozac sequence | Start codon | Stop codon | Sbf1 cutsite

9.6.17 Swap70

```
GGATCCGCCACCATGAGAGGCCTGAAGGACGAGCTGCTGAAGGCCATCTGGCACGCCTTCACAGCCCTGGACCTGGAC
AGATCCGGCAAGGTGTCCAAGAGCCAGCTGAAGGTGCTGAGCCACAACCTGTGCACCGTGCTGAAAGTGCCCCACGAC
CCTGTGGCCCTGGAAGAACACTTCAGGGACGACGATGAGGGCCCCGTGTCCAACCAGGGCTACATGCCCTACCTGAAC
AAGTTCATCCTGGAAAAGGTGCAGGACAACTTCGACAAGATCGAGTTCAACAGGATGTGCTGGACCCTGTGCGTGAAG
AAGAACCTGACCAAGAGCCCCCTGCTGATCACCGAGGACGACGCCTTCAAAGTGTGGGTCATCTTCAACTTTCTGAGC
GAGGACAAGTACCCCCTGATCATCGTGCCCGAGGAAATCGAGTACCTGCTGAAGAAACTGACCGAGGCCATGGGCGGA
GGCTGGCAGCAGGAACAGTTCGAGCACTACAAGATCAACTTCGATGACAACAAGGACGGCCTGAGCGCCTGGGAGCTG
ATCGAACTGATCGGCAACGGCCAGTTCAGCAAGGGCATGGACAGACAGACCGTGTCCATGGCCATCAACGAGGTGTTC
AACGAGCTGATCCTGGACGTGCTGAAGCAGGGCTATATGATGAAGAAGGGCCACAAGAGGAAGAACTGGACCGAGCGG
TGGTTTGTGCTGAAACCCAACATCATCAGCTACTACGTGTCCGAGGATCTGAAGGACAAGAAGGGCGACATCCTGCTG
GACGAGAACTGCTGCGTGGAAAGCCTGCCCGACAAGGATGGCAAGAAGTGCCTGTTCCTGATCAAGTGCTTCGATAAG
ACCTTCGAGATCAGCGCCAGCGACAAGAAAAAGAAACAGGAATGGATTCAGGCCATCTACAGCACCATCCATCTGCTG
AAGCTGGGAAGCCCCCCACCCCACAAAGAGGCCAGACAGAGGCGGAAAGAGCTGAGAAGAAAGCTGCTGGCCGAGCAG
GAAGAACTGGAAAGACAGATGAAGGAACTGCAGGCCGCCAACGAGAACAAACAGCAGGAACTGGAATCCGTGCGGAAG
AAGCTGGAAGAGGCCGCCTCTAGAGCCGCCGACGAGGAAAAGAAGAGACTGCAGACCCAGGTGGAACTGCAGACCAGA
TTCAGCACCGAGCTGGAAAGAGAGAAGCTGATCAGACAGCAGATGGAAGAACAGGTGGCCCAGAAGTCCAGCGAACTG
GAACAGTACCTGCAGAGAGTGCGCGAGCTGGAAGATATGTACCTGAAGCTGCAGGAAGCTCTGGAGGACGAGAGACAG
GCCAGGCAGGATGAGGAAACAGTGCGCAAGCTGCAGGCCAGACTGCTGGAAGAAGAGTCCAGCAAGAGGGCTGAGCTG
GAAAAGTGGCACCTGGAACAGCAGCAGGCCATCCAGACCACCGAGGCCGAAAAACAGGAACTGGAACAGCAGAGAGTG
ATGAAGGAACAGGCTCTGCAGGAAGCCATGGCCCAGCTGGAGCAGCTGGAACTGGAACGGAAGCAGGCCCTGGAACAG
TATGAGGGCGTGAAGAAAAAGCTGGAAATGGCCACCCACATGACCAAGTCCTGGAAGGACAAAGTGGCCCACCACGAG
GGACTGATCAGGCTGATCGAGCCCGGCAGCAAGAACCCTCACCTGATCACCAACTGGGGCCCTGCCGCTTTCACACAG
GCCGAACTGGAAGAGAGGGAAAAGTCTTGGAAAGAAAAGAAAACCACCGAGTGACCTGCAGG
```

BamHI cutsite | Kozac sequence | Start codon | Stop codon | Sbf1 cutsite

9.6.18 Xbp1

```
GGATCCGCCACCATGGTGGTGGTGGCTGCTGCACCTTCTGCCGCCACAGCTGCTCCTAAGGTGCTGCTGCTGTCTGGC
CAGCCTGCTAGCGGAGGCAGAGCACTGCCACTGATGGTGCCTGGCCCTAGAGCTGCTGGCTCTGAGGCTTCTGGCACC
CCCCAGGCCAGAAAGAGACAGAGACTGACCCACCTGAGCCCCGAGGAAAAGGCCCTGAGAAGAAAGCTGAAGAACAGA
GTGGCCGCCCAGACCGCCAGAGACAGAAAGAAAGCCAGAATGAGCGAGCTGGAACAGCAGGTGGTGGACCTGGAAGAG
GAAAACCACAAACTGCAGCTGGAAAACCAGCTGCTGAGAGAAAAGACCCACGGCCTGGTGGTGGAAAATCAGGAACTG
AGAACCAGACTGGGCATGGACACCCTGGACCCTGACGAGGTGCCAGAGGTGGAAGCTAAGGGATCTGGCGTGCGGCTG
GTGGCCGGATCTGCTGAATCTGCCGCCCTGAGACTGTGCGCCCCTCTGCAGCAGGTGCAGGCTCAGCTGAGTCCCCCC
CAGAACATCTTCCCTTGGACACTGACCCTGCTGCCCCTGCAGATCCTGAGCCTGATCAGCTTCTGGGCCTTCTGGACC
AGCTGGACACTGTCCTGCTTCAGCAACGTGCTGCCCCAGAGCCTGCTCGTGTGGCGGAACAGCCAGAGAAGCACCCAG
AAAGACCTGGTGCCCTACCAGCCCCCATTCCTGTGTCAGTGGGGACCCCACCAGCCCAGCTGGAAGCCTCTGATGAAC
AGCTTCGTGCTGACCATGTACACCCCCTCACTGTAACCTGCAGG
```

BamHI cutsite | Kozac sequence | Start codon | Stop codon | Sbf1 cutsite

9.6.19 Xbp1S

```
GGATCCGCCACCATGGTGGTGGTGGCTGCTGCACCTTCTGCCGCCACAGCTGCTCCTAAGGTGCTGCTGCTGTCTGGC
CAGCCTGCTAGCGGAGGCAGAGCACTGCCACTGATGGTGCCTGGCCCTAGAGCTGCTGGCTCTGAGGCTTCTGGCACC
CCCCAGGCCAGAAAGAGACAGAGACTGACCCACCTGAGCCCCGAGGAAAAGGCCCTGAGAAGAAAGCTGAAGAACAGA
GTGGCCGCCCAGACCGCCAGAGACAGAAAGAAAGCCAGAATGAGCGAGCTGGAACAGCAGGTGGTGGACCTGGAAGAG
GAAAACCACAAACTGCAGCTGGAAAACCAGCTGCTGAGAGAAAAGACCCACGGCCTGGTGGTGGAAAATCAGGAACTG
AGAACCAGACTGGGCATGGACACCCTGGACCCTGACGAGGTGCCAGAGGTGGAAGCTAAGGGATCTGGCGTGCGGCTG
GTGGCCGGATCTGCTGAATCTGCTGCTGGCGCTGGCCCCGTCGTGACATCTCCTGAGCATCTGCCCATGGACAGCGAC
ACCGTGGCCAGCAGCGACAGCGAGAGCGATATCCTGCTGGGCATCCTGGACAAGCTGGACCCCGTGATGTTCTTCAAG
TGCCCCAGCCCTGAGAGCGCCAGCCTGGAAGAACTGCCCGAGGTGTACCCTGAGGGCCCTAGCTCTCTGCCTGCCAGC
CTGAGTCTGAGCGTGGGCACAAGCAGCGCCAAGCTGGAAGCCATCAACGAGCTGATCAGATTCGACCACGTGTACACC
AAGCCCCTGGTGCTGGAAATCCCCAGCGAGACAGAGTCCCAGACCAACGTGGTCGTGAAGATCGAGGAAGCCCCCCTG
AGCAGCAGCGAAGAGGACCACCCTGAGTTCATCGTGTCCGTGAAGAAAGAACCCCTGGAAGATGACTTCATCCCCGAG
CTGGGAATCAGCAACCTGCTGAGCAGCTCCCACTGCCTGAGGCCTCCAAGCTGTCTGCTGGACGCCCACAGCGACTGT
GGCTACGAGGGAAGCCCTAGCCCCTTCAGCGACATGTCTAGCCCTCTGGGCACCGACCACAGCTGGGAGGACACATTC
GCTAACGAGCTGTTCCCCCAGCTGATCTCAGTGTGACCTGCAGG
```

BamHI cutsite | Kozac sequence | Start codon | Stop codon | Sbf1 cutsite

9.6.20 Znf217

```
GGATCCGCCACCATGCCTACACAGAGCCTGCTGGTGTACATGGACGGCCCCGAGGTGCTGTCTAGCAGCCTGGGCTCT
CAGATGGAAGTGGACGACGCCGTGCCCATCAAGGGCCCTGTGGCTGTGCCTTTCAGAGCCGCCCAGGAAAAGTCCATG
GCCGTGGCTGAGGGCCACATGCCCCTGGACTGCATGTTCTGTAGTCAAGTGTTCAGCCAGGCCGAGGACCTGAGCCAG
CATGTGCTGCTGCAGCATAGACCCACCCTGTGCGAACCCGCCGTGCTGAGAGTGGAAGCCGAGTACCTGTCCCCCCTG
GACAAGGCCCTGGAACCTACAGAGCCCGCTCTGGAAAAGAGCGGCGAGGACCCTGAGGAACTGAGCTGCGACGTGTGC
GGCCAGACATTCCCTGTGGCCTTCGACGTGGAATCCCACATGAAGAAGCACAAGGACAGCTTCACCTACGGCTGCTCC
ATGTGCGGCAGAAGATTCAAAGAGCCCTGGTTCCTGAAGAACCACATGAGAACCCACAACGGCAAGAGCGGCACCAGA
AGCAAACTGCAGCAGGGCATGGAAAGCCCCGTGACCATCAACGAGGTGGTGCAGCCTCACGCCCCTGGCAGCATCAGC
ACCCCCTACAAGATCTGTATGGTGTGCGGCTTCCTGTTCCCCAACAAGCAGAGCCTGATCGAGCACAGCAAGGTGCAC
GCCAAAGAAACCGTGCCCAGCGCCTCTAACGTGGCCCCTGACGACCACAGAGAGGAACCCACCAGCCCCAGAGAAGAA
CTGCTGCAGTTCCTGAACCTGAGGCCCAGAAGCACCGCCGGCAGCACCGTGAAGCCTATGACCTGCATCCCCCAGCTG
GACCCCTTCACCACCTACCAGGCTTGGCAGCTGGCCACCAAGGGAAAGGTGGCAGTGGCTCAGGAAGAAGTGAAAGAG
TCCGGCCAGGAAGGCTCCACCGACAACGACGACAGCTGCAGCGAGAAAGAGGAACTGGGCGAGATCTGGGTGGGAGGC
AAGGCTGAGGGAAGCGGCAAGTCCAAGACCAGCAAGAGCAGCTGCCCTGGCCTGTCCCAGGACAAAGAGAAGCCCAGA
CACGCCAACAGCGAGGTGCCAAGCGGCGACAGCGACCCTAAGCTGAGCAGCAGCAAAGAAAAGCCTACCCACTGCTCC
GAGTGCAGCAAGGCCTTCAGAACCTACCATCAGCTGGTGCTGCACAGCAGAGTGCACAGAAAGGACAGAAGAACCGAC
GCCCTGAGCCCCACCATGGCTGTGGATGCAAGACAGCCCGGCACCTGTAGCCCTGACCTGAGCACCACCCTGGAAGAT
AGCGGCGCTGGCGATAGAGAGGGCGGAAGCGAGGACGGCTCTGAGGATGGACTGCCTGATGGCCTGCACCTGGATAAG
AACGACGACGGCGGCAAGGCTAAGCCCCTGCCTAGCAGCAGAGAGTGCAGCTACTGCGGCAAGTTCTTCCGCAGCAAC
TACTACCTGAACATCCACCTGAGGACACACACCGGCGAGAAGCCCTACAAGTGCGAGTTCTGCGAGTACGCCGCTGCC
CAGAAAACCAGCCTGAGATACCACCTGGAAAGACACCACAAGGATAAGCAGCCCGTGGACGCTGCCGCCGAGTCTAAG
TCTGAGGGCAGAAGCCAGGAACCCCAGGACGCCCTGCTGACAGCCGCTGATTCTGCCCAGACCAAGAACCTGAAGAGA
TTCCTGGACGGCGCCAAGGACGTGAAGGGCAGCCCTCCTGCCAAGCAGCTGAAAGAAATGCCCAGCGTGTTCCAGTCC
GTGCTGTCCCCTGCCCACAGCAACGACACCCAGGACTTCCACAAGCACGCCGCCGACTCTGCCGAGAAGGCCAGAAAG
TCTCCCGCCCCTACCTACCTGGACATGCAGAGAAAGAAGGCCGGCGAGCCTCAGGCCAGCAGCCCTGTGTGTAGACTG
GAAGGCGTGGGCAGCCTGGCTAGAGAGGCTGGCCACAGAGAAAAGATGGACCAGGATGCCGACTACAGACACAAGCCT
GGCGCCGACTGCCAGGACAGACCTCTGAACCTGTCTCTGGGCCCTCTGCACGCCTGTCCTGCCATCAGCCTGAGCAAG
TGCCTGATCCCCAGTATCGCCTGCCCCTTCTGCACCTTCAAGACCTTCTACCCCGAAGTGCTGATGATGCACCAGAGG
CTGGAACACAGATACAACCCCGACCCCCACAAGAACGGCAGCTCCAAGAGCGTGCTGAGGAACAGAAGGACCGGCTGC
CCTCCAGCTCTGCTGGGCAAAGATGTGCCTCCTCTGAGCGGCCTGCACAAGCCCAAGGCCAAGACCGCTTTCAGCCCT
CACAGCAAGTCCCTGCACAGCGAGAAGGCTAGACAGGGCGCCAGCGGCCCTTCTAAGGCCCCTCAGACAAGCGGCCCT
GACAACAGCACACTGGCCCCCAGCAACCTGAAGTCCCACAGATCCCAGCCTAACGCTGGCGGCACAAGCGCCACAAGA
CAGCAGCAGTCCGAGCTGTTCCCAAAGGGCGGAGTGCCTGCCGCTATGGACAAAGTGAAGAGGCCCGAGCCCAAGCTG
AAGTCTCTGCCTGCCAGCCCTAGCCAGAGCCCCCTGTCCAGCAACAACAGCAACGGCAGCGTGGAATACCCCGTGAAG
GTGGACGGACCTTGGGCCCAGCAGGGAAGAGACTACTACTGCCACAGAAACTCCGGCAGCGCCGCAGCTGAGTACAGC
GAGCCACACCCCAAGAGACTGAAGTCCAGCGCCGTGTCCCTGGACACAGAGCACGCTGGCACAAACGGCAGACGGGGC
TTCGAGCTGCCCAAGTATCACGTCGTGCGGAGCATCACCAGCCTGCTGCCACCTGAATGCGTGCGGCCTCCTCCTGTG
CTGCCACACAAGGCCAGATTTCTGAGCCCTGGCGAGGTGGAATCTCCTAGCGTGCTGGCCGTGCAGAAGCCTTACTCT
GCTAGCGGCCCACTGTACACCTGTGGCCCAGTGGGACATGCCGGCGGATCTCCAGCACTGGAAGGGAAGAGGCCTGTG
TCCCACCAGCACCTGAGCAACTCCATGCTGCAGAAGAGAAGCTACGAGAACTTCATCGGCAACACCCACTACAGACCC
AACGACAAGAAGCCCTGACCTGCAGG
```

BamHI cutsite | Kozac sequence | Start codon | Stop codon | Sbf1 cutsite