Technische Universität München, Fakultät für Medizin

Genome-wide *in vivo* screens link regulatory principles and phenotype evolution in T cell leukemia

Anja Fischer

Vollständiger Abdruck der von der Fakultät für Medizin der Technischen Universität München zur Erlangung einer

Doktorin der Naturwissenschaften (Dr.rer.nat)

genehmigten Dissertation.

Vorsitz: Prof. Dr. Jürgen Ruland

Prüfer*innen der Dissertation:

1. Prof. Dr. Radu Roland Rad

2. Prof. Angelika Schnieke, Ph.D.

3. Prof. Dr. Irmela Jeremias

Die Dissertation wurde am 07.11.2022 bei der Technischen Universität München eingereicht und durch die Fakultät für Medizin am 16.05.2023 angenommen.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

°C	Degree Celcius
aCGH	Array-based comparative genomic hybridization
ALL	Acute lymphoblastic leukemia
AML	Acute myeloid leukemia
ARCIS	Annotation pipeline for Regulatory Common Insertion Sites
AS	Antisense
ATP2	Activating transposon 2
ATAC-seq	Assay for Transposase-Accessible Chromatin using sequencing
B-ALL	B cell acute lymphoblastic leukemia
Bp	Base pairs
Chr	Chromosome
CIS	Common insertion site
cDNA	Complementary DNA
ChIP-seq	Chromatin immunoprecipitation sequencing
CIMPL	Common Insertion site Mapping PLatfrom
CLP	Common lymphoid progenitor
CMP	Common myeloid progenitor
CNV	Copy number variation
CRISPR	Clustered regularly interspaced short palindromic repeats
Ctrl	Control
DMEM	Dulbecco's modified eagle's medium
DMSO	Dimethyl sulfoxide
DNA	Desoxyribonucleic acid
DN	Double negative
Dox	Doxycycline
DP	Double positive
ETP	Early T cell Precursor
FACS	Fluorescence Activated Cell Sorting
FCS	Fetal calf serum
FDR	False Discovery Rate
FFPE	Formalin-fixed paraffin-embedded
FWER	Family Wise Error Rate
gCIS	Gene overlapping a common insertion site
GEP	Gene expression profile
GFP	Green fluorescent protein
GKC	Gaussian Kernel Convolution
GMP	Granulocyte-macrophage progenitor
GSEA	Gene set enrichment analysis
GRO-seq	Global run on sequencing
GWAS	Genome-wide association studies
H&E	Hematoxylin eosin
H3K27ac	Acetylation of lysine 27 histone 3
H3K4me1	Mono-methylation of lysine 4 histone 3
HEK	Human embryonic kidney
Het	Heterozygous
HiC	Chromosome conformation capture method
Hom	Homozygous

IHC	Immunohistochemistry
Kb	Kilobase
KO	Knockout
Indel	Insertions and/or deletions
Linc	Long intergenic non-coding
Lnc	Long non-coding
Mi	Micro
Mb	Megabase
MPP	Multipotential progenitor
MSCV	Murine stem cell virus
MTL	Mature T cell lymphoma
NA	Not available
Nc	Non-coding
Ncruc	Noncoding region upstream of *Cdkn2a*
ngCIS	Gene not overlapping a common insertion site
NGS	Next-generation sequencing
P/S	Penicillin/streptomycin
pA	Poly adenylation site
PB	PiggyBac
PBS	Phosphate buffered saline
PCA	Principal component analysis
PCR	Polymerase chain reaction
QiSeq	Quantitative insertion site sequencing
qPCR	Quantitative polymerase chain reaction
R26	Rosa26 locus
RNA	Ribonucleic acid
RT	Room temperature
rtTA	Reverse tetracycline-controlled transactivator
SA	Splice acceptor
SB	Sleeping Beauty
SD	Splice donor
SE	Super-enhancer
sgRNA	Single guide RNA
SPF	Specific-pathogen-free
STL	Small T cell lymphoma
T-ALL	T cell acute lymphoblastic leukemia
T-LBL	T cell lymphoblastic lymphoma
TCR	T cell receptor
TES	Transcription end site
TF	Transcription factor
TRE	Tetracycline responsive element
TSO	Template switch oligo
TSS	Transcription start site
UMI	Unique molecular identifier
UV	Ultraviolet
WGS	Whole genome sequencing
WHO	World health organization
WT	wildtype

Abstract

The past decades of cancer research were primarily focused on identifying and understanding the role of protein-coding mutations. However, cancer development is characterized by extensive changes of regulatory landscapes, which are poorly understood at the functional level. Even more challenging than decrypting the tumor's regulome, is the discovery of the cancer's cell of origin or of subtle regulatory processes during cancer evolution.

Likewise, although advances have been made in the genetic characterization of T cell acute lymphoblastic leukemia (T-ALL), genetic subtypes are clinically not yet considered as different disease entities due to a lack of understanding of subtype evolution and biology.

Building on new mouse models, genetic tools, screening technologies and data analysis pipelines, this thesis addressed these challenges at different levels enabling high-throughput functional mapping of regulatory landscapes and evolutionary principles in oncogenesis.

In the first part of this thesis, a genome-wide transposon screen was performed. To uncover regulatory landscapes of T-ALL, a method for the systematic perturbation of the non-protein coding regulatory genome was developed. Thereby, hundreds of regulatory elements and cancer genes involved in T-ALL were identified creating a comprehensive resource. Further, the evolutionary principles were interrogated in a prospective manner by mapping the stage and cell of origin at which transposon insertions occur. This approach enabled the discovery of molecular determinants of phenotypic diversification (T-ALL subentities). Modelling human T-ALL heterogeneity in mice revealed tumor subtype-specific clonal structures, driver genes, pathway hierarchies, genetic interactions and sequentialities. Unlike early T precursor leukemias, tumors developing from committed T cells display a dominance of insertions in regulatory elements, indicating context-specific roles of subtle gene regulation.

The second part focused on functional analysis of newly discovered regulatory elements and the validation of their relevance in human T-ALL. To this end, CRISPR-based perturbation of non-protein coding elements such as lncRNAs or enhancers was performed in cell lines and their link to potential target genes, such as *Pten*, *Ikzf1* or *Zeb1,* established. To investigate human relevance, cancer risk variants described in human genome-wide association studies (GWAS) were intersected with the gene list of regulatory elements identified in the screen. Of note, target genes of regulatory transposon insertions were highly significantly enriched for human GWAS hotspots highlighting subtle gene regulatory effects.

Taken together, this study created comprehensive functional maps of the genetic and regulatory principles orchestrating T-ALL evolution and phenotypic diversification. The developed concepts, analytical methods and computational tools as well as connected

pathological insights describe the first survey of its kind for any cancer type and can be applied to other tumor types in the future.

Zusammenfassung

In den vergangenen Jahrzehnten konzentrierte sich die Krebsforschung in erster Linie auf die Identifizierung und das Verständnis der Rolle von Protein-kodierenden Mutationen. Die Krebsentwicklung ist jedoch durch weitreichende Veränderungen der regulatorischen Landschaften gekennzeichnet, die auf funktioneller Ebene nur unzureichend verstanden sind. Eine noch größere Herausforderung als die Entschlüsselung des Tumor-Reguloms ist die Entdeckung der Krebs-Ursprungszelle oder subtiler regulatorischer Prozesse während der Krebsentwicklung.

Obwohl bei der genetischen Charakterisierung der akuten lymphoblastischen T-Zell-Leukämie (T-ALL) Fortschritte erzielt wurden, werden die genetischen Subtypen klinisch noch nicht als unterschiedliche Entitäten betrachtet, da die Evolution und Biologie der Subtypen nicht verstanden ist.

Basierend auf neuen Mausmodellen, genetischen Werkzeugen, Screening-Technologien und Datenanalysepipelines wurden diese Herausforderungen in der vorliegenden Arbeit auf verschiedenen Ebenen angegangen, um eine funktionelle Hochdurchsatzkartierung von regulatorischen Landschaften und evolutionären Prinzipien in der Onkogenese zu ermöglichen.

Im ersten Teil dieser Arbeit wurde ein genomweiter Transposon-Screen durchgeführt. Um die regulatorischen Landschaften der T-ALL aufzudecken, wurde eine Methode zur systematischen Untersuchung des nicht-proteinkodierenden regulatorischen Genoms entwickelt. Auf diese Weise wurden Hunderte von regulatorischen Elementen und Krebsgenen, die an T-ALL beteiligt sind, identifiziert und eine umfassende Daten-Ressource geschaffen. Darüber hinaus wurden die Prinzipien der Tumorevolution in einer prospektiven Weise untersucht, indem das Stadium und die Ursprungszelle, in denen Transposon-Insertionen auftreten, kartiert wurden. Dieser Ansatz ermöglichte die Entdeckung der molekularen Determinanten der phänotypischen Diversifizierung (T-ALL-Subentitäten). Die Modellierung der menschlichen T-ALL-Heterogenität in Mäusen ergab tumorsubtypspezifische klonale Strukturen, Treibergene, Signalweghierarchien, genetische Interaktionen und Sequentialitäten. Im Gegensatz zu frühen T-Vorläuferleukämien weisen Tumoren, die sich aus reifen T-Zellen entwickeln, eine Dominanz von Insertionen in regulatorischen Elementen auf, was auf eine kontextspezifische Rolle subtiler Genregulation hinweist.

Der zweite Teil konzentrierte sich auf die funktionale Analyse der neu entdeckten regulatorischen Elemente und die Validierung ihrer Bedeutung für die menschliche T-ALL. Zu diesem Zweck wurde eine CRISPR-basierte Perturbation von nicht-proteinkodierenden Elementen wie lncRNAs oder Enhancern in Zelllinien durchgeführt und ihre Verbindung zu

potenziellen Zielgenen wie *Pten*, *Ikzf1* oder *Zeb1* hergestellt. Um die Relevanz für den Menschen zu untersuchen, wurden Krebsrisikovarianten, die in genomweiten Assoziationsstudien (GWAS) beim Menschen beschrieben wurden, mit den hier entdeckten Zielgenen der regulatorischen Elemente, überlappt. Bemerkenswerterweise waren diese Zielgene der regulatorischen Hits hoch signifikant angereichert in Regionen menschlicher GWAS-Hotspots, was subtile genregulatorische Effekte betont.

Insgesamt wurden in dieser Arbeit umfassende funktionelle Karten der genetischen und regulatorischen Prinzipien erstellt, die die Evolution und phänotypische Diversifizierung der T-ALL steuern. Die entwickelten Konzepte, Analysemethoden und computergestützten Systeme sowie die damit verbundenen pathologischen Erkenntnisse stellen die erste Erhebung dieser Art für eine Krebsart dar und können in Zukunft auch auf andere Tumorarten angewendet werden.

1. Introduction

Cancer is the first or second leading cause of death in most countries of the world (Bray et al., 2021) with an increasing tendency due to the aging population. In 2020, almost 10 million cancer death occurred worldwide confirming that cancer is a crucial barrier to increasing life expectancy (Sung et al., 2021). Although tremendous advances have been made in the field of targeted and specific therapies for some entities, other cancer types are still treated with the therapeutic approaches (high-dose chemotherapy) discovered decades ago. Bottlenecks for improved diagnostics and personalized therapy include the multifaceted molecular processes and the interplay of multiple factors underlying cancer evolution.

Already five decades ago, cancer was shown to be an evolutionary process with parallels to Darwinian natural selection (Nowell, 1976). Cancer evolves by a multistep process of clonal expansion and genetic diversification. This dynamic tumor progression is complex and characterized by highly variable patterns of clonal architecture (Greaves and Maley, 2012). The importance of this process became more and more clear as tumor heterogeneity was found to be a major cause of therapeutic resistance.

Tumors are characterized by specific features, described as “hallmarks of cancer” by Hanahan and Weinberg (2000; 2011). These features include, among others, increased proliferative signaling, limitless replicative potential, evading apoptosis and genome instability. Very recently, the hallmarks were updated and now also include non-mutational epigenetic reprogramming and phenotypic plasticity (Hanahan, 2022). This latest update underlines the complexity behind the molecular principles driving cancer and demonstrates the importance of the epigenome/non-protein-coding genome and the phenotypic evolution in cancer research.

While in recent years, many innovative methods and techniques for cancer gene discovery have brought important advances in cancer genetics, there is still a need for methods and tools to functionalize the (epi-)genome in the course of cancer evolution.

1.1 Sequencing and screening technologies in cancer research

Next generation sequencing (NGS) enabled large-scale studies and created enormous catalogues of mutated genes for all major cancer entities (Alexandrov et al., 2020; Alexandrov et al., 2013; Lawrence et al., 2014; Stratton et al., 2009). Genomic approaches include whole exome sequencing (WES) to detect mutations in protein-coding (PC) genes, while transcriptomic approaches such as RNA sequencing (RNA-Seq) reveal which genes are dysregulated, but not necessarily mutated. As sequencing costs decreased substantially, also whole genome sequencing (WGS) approaches were applied to detect all classes of genetic

alterations, including single nucleotide variants in non-exonic regions and structural changes such as copy number variations (CNVs).

Now, that for most cancer entities a plethora of samples were sequenced and these extensive efforts are close to completion, it is realized that we are still far from fully understanding the underlying pathogenic mechanisms. It became clear that the processes driving tumorigenesis are difficult to capture at the molecular, cellular and organismal level. Moreover, we still struggle, for example, to (i) interpret the extensive changes of the non-coding regulatory landscapes in cancer development and (ii) assess the temporal order of mutation acquisition.

As many of these fundamental questions cannot be addressed systematically on a genome-wide scale by “omics” based approaches, unbiased genetic screening became an attractive solution to overcome these limitations (reviewed in Weber et al. (2020)). Genome-wide forward genetic screening in a model organism can be seen as a complementary cancer gene discovery approach to classical sequencing.

Screening approaches can be differentiated into library-based screening and mutagenesis screens (Weber et al., 2020). CRISPR/Cas9 knockout screening is the most widely used library-based screening technology. However, one major disadvantage of CRISPR screening includes the limit of guide RNAs that can be used. Covering the non-protein-coding genome, which is around 50 times larger than the coding part, exceeds the capacities of guide libraries. This thesis focuses on the application of insertional mutagenesis screening to identify non-protein-coding, epigenetic driver alterations in cancer evolution.

1.2 The non-protein-coding genome in cancer

1.2.1 Elements in the non-coding genome

The non-protein-coding (nPC) genome is defined as the collection of nucleotides not belonging to protein-coding (PC) sequence (exons) and represents the majority of the genome (~98-99%). Within the nPC genome, regulatory elements can be differentiated by specific histone modifications representing epigenetic indicators of chromatin state (Kimura, 2013). Examples include histone 3 lysine 27 (H3K27) acetylation and H3K4 monomethylation at enhancers, H3K4 trimethylation at promoters, H3K36 trimethylation at gene bodies and H3K27 trimethylation at repressed sequences (Kimura, 2013).

Cis-regulatory elements represent all DNA sequences that regulate gene expression (enhancer, promoter, insulator, silencer) by recruiting specific proteins and thereby influencing the 3D structure of the genome and target gene expression. The cis-regulatory code is also known as the genome’s second code providing necessary information to read regulatory information (Zeitlinger, 2020). The most abundant cis-regulatory sequences are enhancers.

Enhancers represent short DNA sequences and regulate gene expression from a distance through binding of transcription factors and juxtaposition of enhancer and nearby promoter DNA (Shlyueva et al., 2014). Enhancers are located in the intergenic area or in intronic sequence. Super-enhancers (SEs) are defined as an especially large and important enhancer cluster containing multiple transcription factor binding sites and high H3K27 acetylation levels (Pott and Lieb, 2015; Whyte et al., 2013). The location of potential (super-)enhancers can be inferred from histone modifications or the binding of transcription factors in open chromatin (Shlyueva et al., 2014). However, the genome-wide prediction based on chromatin states should be interpreted with caution. None of the features is perfectly predictive and their functional role still needs to be validated as enhancers are cell type specific and context dependent (Long et al., 2016).

Moreover, it has been demonstrated that the genome is pervasively transcribed producing, in addition to mRNA, many different types of non-protein-coding RNA (Kapranov et al., 2007). Despite their important roles in translation (rRNA, tRNA), many non-coding transcripts harbor gene regulatory function. The class of long non-coding (lnc) RNAs can be further subdivided as lincRNAs (long intergenic), antisense or sense intronic RNA depending on their genomic location in the intergenic area or antisense/sense to a protein-coding gene, respectively. In addition to classical non-coding RNAs, RNA transcription is also observed at enhancers although the function of this enhancer RNA (eRNA) is still under investigation (Li et al., 2016).

lncRNAs show tissue specific expression patterns with levels generally much lower than protein-coding genes. Their widespread roles in cell homeostasis and gene regulation is mediated through interactions with DNA, mRNA or proteins and can affect multiple stages including chromatin modification/structure and protein biogenesis (Mattick and Rinn, 2015; Quinn and Chang, 2016). Thus, lncRNAs are important molecules in transcriptional regulation in *cis* and *trans*. High-throughput characterization of functional lncRNAs is, however, hindered by several challenges: (i) low sequence conservation among species hinders the establishment of animal models, (ii) the low expression levels require sequencing approaches with an extremely high coverage (iii) the functional effect might be rather the act of transcription than the lncRNA itself (not detectable in *in vitro* genetic engineering experiments).

A second class of non-coding transcripts are micro RNAs (miRNA) a well-defined cohort of small RNAs. As lncRNAs are defined as transcripts with more than 200 nucleotides, miRNAs only harbor around 20 nucleotides. MiRNAs have a well-established role in gene regulation influencing transcript translation and degradation (Ameres and Zamore, 2013).

As both, cis-regulatory elements and non-protein-coding transcripts can contribute to the regulation of gene expression, they are referred to as ‘regulatory elements (REs)’ in this study.

The definition ‘regulome’ in analogy to genome comprises regulatory elements but also genes important for gene regulation such as transcription factors.

1.2.2 Reasons to study the non-coding genome

Chromatin remodeling plays a major regulatory role in cell-type specific function and differentiation (Ho and Crabtree, 2010). In development, differentiation is a gradual transition from open to condensed chromatin states, first described in detail in the hematopoietic system (Lara-Astiaso et al., 2014). Cancer evolution is characterized by extensive changes of these regulatory landscapes.

Non-protein coding regions make up almost 99% of the genome. However, so far there is little consensus regarding the percentage of functional elements. Large-scale studies such as the ENCODE project suggest that the majority of ncDNA is functional (ENCODE Project Consortium (2012)), while mutational load analysis indicated that a maximum of 25% harbors functional elements (Graur, 2017).

Non-coding mutations affecting the regulatory genome are frequent in cancer (Weinhold et al., 2014). However, the analysis of several WGS datasets revealed that the interpretation still remains challenging (Elliott and Larsson, 2021). Due to the non-recurrent nature of non-coding mutations, it is still difficult to distinguish drivers from the high number of passengers, or real mutations from background noise (Rheinbay et al., 2020; Stratton et al., 2009).

The majority of trait-associated genetic variants (>90%) map to the non-protein-coding genome (Gallagher and Chen-Plotkin, 2018; Maurano et al., 2012) underlining the importance of the nPC genome and the great significance to solve the problem of deciphering the cis-regulatory code (Zeitlinger, 2020). Likewise, the transcription of non-coding regions is very complex, with up to 90% of the genome being transcribed (Lee, 2012) increasing the need for the annotation of functional non-coding regions.

The effect of variations in regulatory elements is most likely only subtle compared to mutations directly affecting the protein-coding genome. For tumor-suppressors, it was shown that only one allele or even a more subtle dysregulation is enough for interference with the tumor suppressing function (Alimonti et al., 2010). Therefore, regulatory alterations might contribute to this subtle gene regulation of tumor suppressors. Moreover, also for oncogenes it was reported that the dosage is very important for tumor progression (Berger et al., 2011).

1.2.3 Tools to study the non-coding genome

In the last decades a wide range of NGS techniques for epigenomics was developed. These technologies can be distinguished into different categories analyzing the accessibility of the

chromatin (ATAC, DNAse-Seq) or histone modifications and transcription factor binding analysis (ChIP-Seq) (reviewed in Elkon and Agami (2017)).

While regulatory elements (REs) can be identified easily by the above mentioned techniques, a major challenge includes the determination which RE controls which gene due to large distances between these elements in the genome. Chromosomal conformation capture (3C) techniques are crucial to understand the link between nuclear structure and function (reviewed in (Bonev and Cavalli, 2016; Davies et al., 2017)).

Global run-on sequencing (GRO-Seq) represents a method for nascent RNA sequencing mapping the position, amount and orientation of RNA polymerases transcribing DNA (Core et al., 2008). Using labelled nucleotides, GRO-Seq provides a snapshot of genome-wide transcription by capturing all kinds of transcripts, also instable ones such as enhancer RNA (eRNA) (Core et al., 2008; Kaikkonen et al., 2013).

In this study, regulatory activity is defined as a signal in one of the epigenomic assays such as ChIP-Seq (H3K27ac, H3K4me1), ATAC/DNase-Seq or GRO-Seq. However, all described techniques are only descriptive. This leads to a fast annotation of the non-coding genome but functional consequences or relevance of changes are still unexplored due to the lack of tools to systematically assess which regions/transcripts are functional. There is a lack of scalable methods capable to perturb the epigenome and capture the relevant phenotype: cellular transformation in an organism. Major bottlenecks in the analysis of the technologies mentioned in Table 1 include the restriction of many of these tools to the protein-coding or transcribed genome.

Table 1: Selection of sequencing and screening tools used in cancer research. *Technologies used in cancer research. Methods used in this study are indicated. * Publicly available GWAS datasets, **Publicly available datasets used; ***used for functional validation, not for screening. Only methods relevant to this study are mentioned.*

Technique	Analysis	Interpretation	Used in this study
Genomics			
Whole exome sequencing (WES)	Point mutations, indels	Protein-coding genes	No
Whole genome sequencing (WGS)	Structural variation, nc mutations	Alteration in the complete genome	Yes*
Transcriptomics			
RNA-Seq	Differential expression	Dysregulated genes	Yes
GRO-Seq	Nascent RNA transcription	Regulatory activity	Yes
Epigenomics			
ATAC/DNase-Seq	Open chromatin	Chromatin accessibility	Yes**
ChIP-Seq	Histone modifications, Transcription factor binding	Chromatin activity	Yes**

HiC	3D chromatin capture	Interacting genomic areas in 3D space	Yes**
Genetic screening			
CRISPR	Mostly gene inactivation, essential genes	Mostly focused on PC genome	Yes***
Transposon	Activation and inactivation	Whole genome	Yes

1.3 *PiggyBac* transposon screening

The challenges of genome-wide genetic screening for cancer gene discovery were described in 1.1. Mutagenesis screens represents a versatile tool to overcome the limitations of "omics"- and library-based screening approaches (missing functional readout and impossibility to cover the complete genome, respectively) and include irradiation or chemical approaches as well as insertional mutagenesis (reviewed in (Weber et al., 2020)). Insertional mutagenesis is defined by virus- or transposon-mediated creation of mutations by the addition of a specific sequence to the genome. Viruses or transposons integrating into the genome are attractive mutagens due to their molecular fingerprint allowing a simple recovery of insertion sites by splinkerette-based PCR approaches.

1.3.1 Transposons used for insertional mutagenesis

Transposons are mobile genetic elements, which were first discovered as 'jumping genes' by Barbara McClintock in maize in the 1940s (McClintock, 1950). Their ability to change the position within the genome through mobilization by an enzyme called "transposase" can be exploited for insertional mutagenesis screening approaches. Transposon recognition and binding by the transposase for mobilization occurs at specific sites within the inverted terminal repeats (ITRs). In nature, the sequence between the ITRs encodes for the transposase. Transposable elements were used in invertebrates for the discovery of key pathways in development (Thibault et al., 2004).

Due to inactivation in vertebrate genomes millions of years ago, transposons were not available in mice until the late 1990s (Ding et al., 2005; Dupuy et al., 2001; Ivics et al., 1997). The development of transposon technologies that are active in vertebrates was a key advance for genetic screening in mammals. In engineered mice, the transposase gene is encoded separately (*in trans*) and the genomic sequence in between the ITRs can be replaced with any DNA cargo. Upon bringing the systems (transposase and transposon) together by mouse crossing, transposition starts and the transposon "jumps" across the genome. Transposon mobilization is a continuous process: Whenever the transposon hits a position leading to a growth advantage, the insertion will be selected and a tumor evolves from a pre-malignant clone. The generation of bifunctional transposons containing a promoter for gene activation

and gene trapping elements (splice acceptors and polyadenylation sites) for gene inactivation enables the possibility to screen for oncogenes and tumor suppressor genes at the same time. Gene inactivation is achieved by trapping/truncation (upon insertion in introns) or frameshift (upon insertion in exons) whereas gene activation is realized through the promoter and the splice donor that were introduced into the transposon. Gene expression can be driven by the promoter if the transposon integrates upstream of the gene. Importantly, the precise mutagenic effect depends on the integration pattern, transposon orientation and spatial relationship to functional genetic elements (Friedrich et al., 2017; Rad et al., 2010; Weber et al., 2020). This represents a benefit over the detection of single nucleotide variants in next generation sequencing, where the functional consequences of such point mutations are often very difficult to extrapolate.

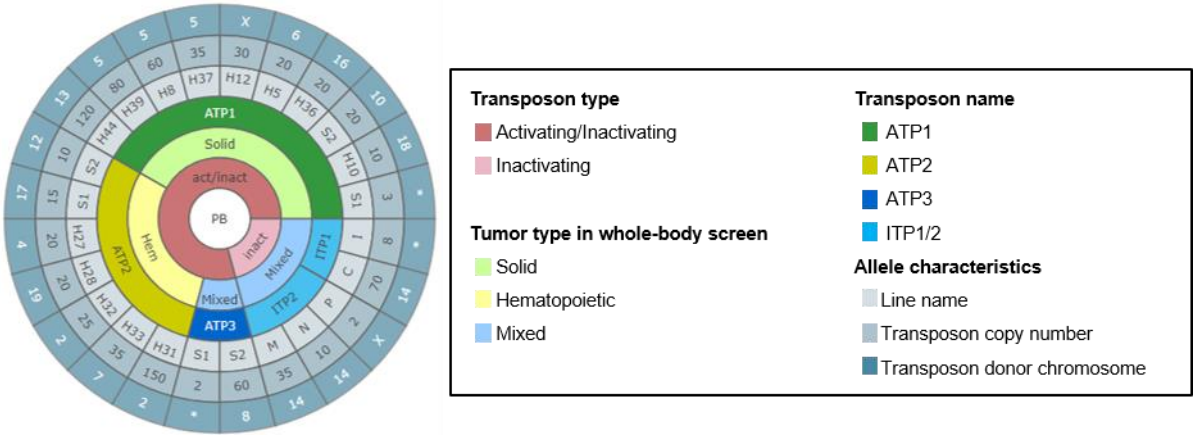

Figure 1: Available PiggyBac transposon mouse lines for cancer gene discovery*. Bifunctional (ATP) and only inactivating (ITP) mouse lines for PiggyBac transposon screening. This study focuses on ATP2 mouse lines containing a promoter derived from the MSCV 5'-LTR inducing predominantly hematopoietic malignancies. ATP, activating transposon; ITP, inactivating transposon; MSCV, murine stem cell virus; LTR, long terminal repeat. adapted from Weber et al., 2020, Nature Reviews Cancer.*

For genetic engineering of mouse lines, transposon sequences from fish and insect genomes were used. *Sleeping Beauty* (SB) was the first transposon engineered in mice and used for cancer gene discovery (Collier et al., 2005; Dupuy et al., 2005; Ivics et al., 1997). *PiggyBac* (PB) originates from the cabbage looper moth *Trichoplusia ni* and was also modified to be active in mammalian cells (Ding et al., 2005). Allan Bradley's group was the first to use the PB transposons to develop genetic tool kits for insertional mutagenesis in mice including 19 different activating transposon (ATP) mouse lines (Rad et al., 2010). Depending on the type of promoter used in individual ATP transposon lines, double transgenic Rosa26^{PB};ATP mice develop solid and/or hematopoietic cancers (Figure 1). In this study, the Rosa26^{PB};ATP2 cohort was expanded and a large collection of transposon-induced hematopoietic tumors was analyzed.

1.3.2 Characteristics of *PiggyBac* transposition

PiggyBac is characterized by a largely unbiased coverage of the genome. Genomic integration of *PiggyBac* is dependent on the minimal sequence TTAA with only 2% of insertions identified in non-TTAA sequence (Li et al., 2013). During excision, PB leaves no footprint mutations (Yusa et al., 2011), a difference compared to the widely-used *Sleeping Beauty* transposons.

PiggyBac shows an integration bias towards open chromatin and the transcriptional start site (TSS) of genes (de Jong et al., 2014; Wang et al., 2008). Additionally, it was shown that *PiggyBac* is biased to transcriptional units, actively transcribed loci as well as interfaces of topologically associated domains in previous studies (de Jong et al., 2014). Despite the bias for highly expressed genes, PB shows characteristics in line to use this system for genome-wide screening for open chromatin and active regulatory elements. This is in contrast to *Sleeping Beauty*, which only showed minimum preference to chromatin states and seems less suitable for chromatin accessibility screening based on a study in embryonic stem cells (Yoshida et al., 2017).

Thus, this *in vivo* screening approach is largely unbiased with respect to localization in the genome and a powerful and versatile tool for the discovery of functionally relevant open chromatin. The focus of this thesis was the application of transposon tools to study regulatory changes in cancer evolution.

1.3.3 Analysis of insertional mutagenesis data

1.3.3.1 Evolution of an individual tumor

Transposon screening is based on a random mutagenesis in large cell pools and subsequent clonal selection. Transposon mobilization starts in context to a specific phenotype and continues in premalignant clones leading subsequently to full-blown tumors. To identify the causative mutations, insertion site sequencing is applied. The improvement of the insertion site sequencing strategy and the drop of sequencing costs enabled semi-quantitative analysis (Friedrich et al., 2017; Klijn et al., 2013; Koudijs et al., 2011). Transposon insertion sites in a tumor can be identified using a high-throughput and high-resolution sequencing approach of PCR-amplified transposon-genome junctions. At the level of an individual cancer, semi-quantitative insertion site sequencing was a key innovation to understand the tumor's genetic complexity and intratumor heterogeneity (Friedrich et al., 2017). The sequence read coverage determined by QiSeq supporting one individual insertion can range from one to tens of thousands reflecting the frequency of this insertion within one cancer sample. This gives insights into the clonal architecture and evolutionary trajectory. A high read coverage can indicate an early occurrence of the insertion (=insertion at the trunk of the evolutionary tree) and the presence of selective pressure against remobilization indicating a biological relevance.

Linking genes to their position at a tumor's evolutionary tree allows inferring a more detailed understanding of the biological role of the respective gene in early or late tumor development. The high sequencing depth in QiSeq therefore enables to assess each tumor's clonality and dissect differences in evolutionary landscapes (Figure 2, (Friedrich et al., 2017)).

Although transposon insertions can be used as markers of clonal size and branched evolution, so far the predominant application rather was cancer gene discovery than evolutionary dynamics. However, deciphering the chronological order of genetic alterations in tumor evolution is crucial to understand stage- and context-specific roles of cancer genes and will be addressed in this thesis.

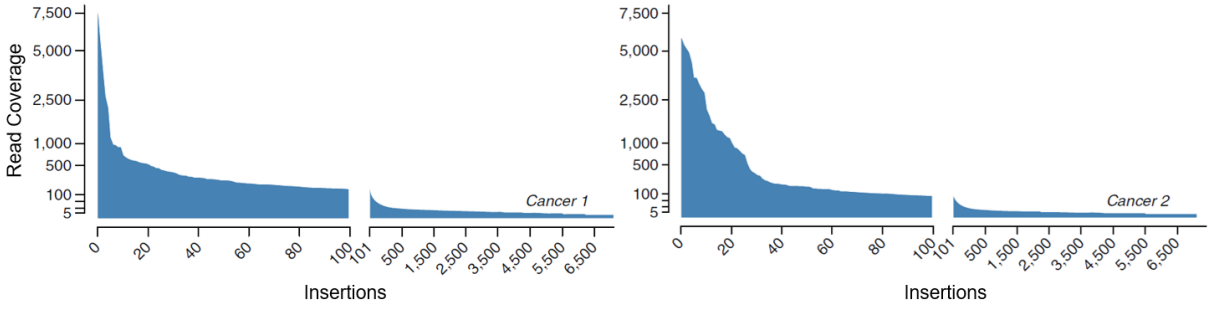

Figure 2: Transposon insertions reveal clonal architecture and tumor evolution*. Two pancreatic tumors are shown with different clonal architecture. Insertions are ranked by read coverage. In the left panel, the tumor is characterized by few high coverage insertions (high clonality) whereas the tumor in the right panel shows many high coverage insertions (multiclonal). Adapted from Friedrich et al., 2017*

1.3.3.2 Common insertion sites

The most common approach to pinpoint candidate cancer genes from insertion data is the analysis of insertions across multiple tumors. So called 'common insertion sites (CISs)' represent genomic regions harboring more insertions than expected by chance (regions that are hit by the transposon in multiple tumors). Statistical models based on Gaussian Kernel Convolution (de Ridder et al., 2006) or Poisson distribution (Bergemann et al., 2012; Sarver et al., 2012) are used to assess whether the insertion density at any position of the genome differs from random (background) distribution.

In standard analysis workflows, protein-coding genes located in or overlapping with CISs are taken as putative cancer drivers. There are even methods available, which only focus on protein coding genes (gene-centric approaches such as described in Brett et al. (2011)). However, insertions into DNA most likely interferes with regulatory function in several ways. The activation and disruption of non-coding transcripts might be similar to protein-coding genes. The interference with cis-regulatory elements is based on the disruption of either binding sites (of transcription factors or other DNA binding proteins in enhancers) or the 3D organization of the genome (Figure 3, (Weber et al., 2020)).

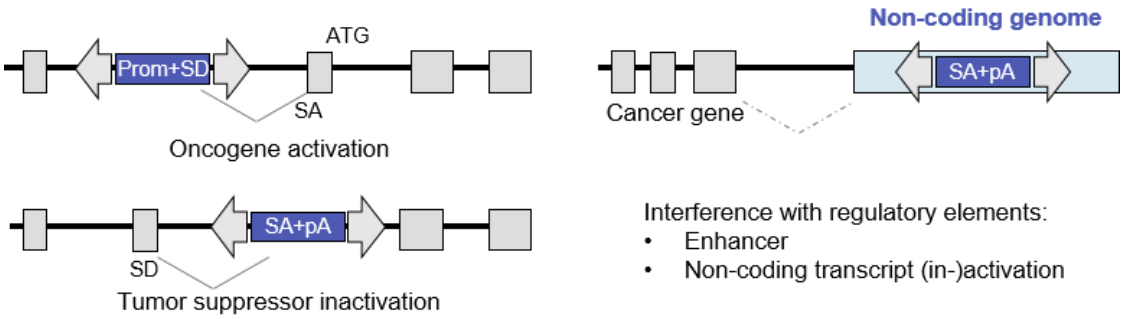

Figure 3: Effect of transposon insertions on the protein-coding and non-coding genome. *Insertion pattern, orientation of transposon insertions and genomic location are important to infer the effect on target gene expression. Oncogenes can be activated when the transposon integrates sense-orientated upstream of an exon with a translational start site (ATG). The transposon is spliced to the next exon using the splice donor (SD) on the transposon and the splice acceptor (SA) of the exon and drives expression through the promoter (Prom) on the transposon. Tumor suppressor gene inactivation is mediated by truncation using bidirectional polyA (pA) sites on the transposon, which are spliced to the transcript mediating gene disruption from intronic positions. However, the effect of transposon insertion in the intergenic area remains unclear. An interference with regulatory elements through steric DNA properties or an (in-) activation of non-coding transcripts represent possible mechanism for transposon-induced tumorigenesis.*

Due to the characteristics described above, the *PiggyBac* screening technology is suitable to screen for regulatory regions (intergenic insertions) and to study cancer evolution (sequencing coverage distinguishes early and late hits).

1.4 Acute leukemias

1.4.1 General classification of leukemias

Hematopoietic malignancies develop from different cells of the hematopoietic tree (Figure 4). Whereas aberrations in myeloid cells such as common myeloid progenitors (CMPs), granulocytes or monocytes/macrophages and their progenitors give rise to various subtypes of acute myeloid leukemia (AML), aberration in B and T cells can lead to the development of immature lymphoblastic neoplasms (acute lymphoblastic leukemia, ALL) or mature T or B cell lymphomas (Figure 4). Decisive for the classification of hematopoietic neoplasms are the morphology of the diseased cells as well as genetic characteristics.

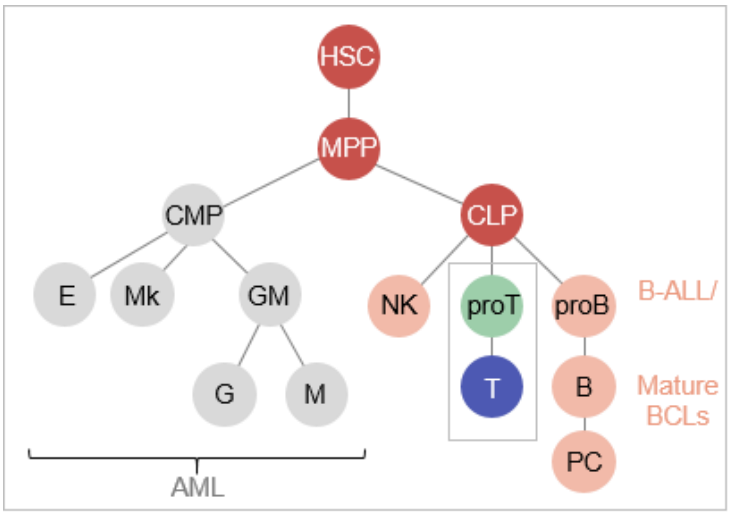

Figure 4: The hematopoietic system and connected malignancies. *The hematopoietic system can be divided into the myeloid and the lymphoid lineage. In a simplified manner, myeloid cells give rise to different subtypes of acute myeloid leukemia (AML) while lymphoid cells either induce immature acute lymphoblastic leukemias (ALL) or mature T or B cell lymphomas (BCL). The subtypes of T cell malignancies are displayed in Figure 6.*

Leukemia, first described in 1845 by Rudolf Virchow is the cancer of white blood cells and can be sub-grouped into ‘acute and chronic’ (associated with their clinical course) and into ‘myeloid and lymphoid’ (dependent on their cell of origin). The pathogenesis of acute leukemias can be described by different steps: (i) malignant transformation of hematopoietic stem cells (HSCs) or progenitor cells, (ii) impaired differentiation of immature blood cells, (iii) suppression of normal hematopoiesis, (iv) hematopoietic insufficiency like anemia, granulocytopenia, thrombocytopenia, and (v) clinical symptoms like paleness, infections, bleeding, sepsis and organ dysfunction due to infiltration.

Hematopoietic tumors are responsible for approximately eight percent of all cancer deaths in the Western World. The 5-year survival rate of leukemias differs between leukemia subtypes (Figure 5). Whereas for chronic leukemias targeted therapy approaches are available and used for a long time, most subtypes of acute leukemia are still treated with high-dose chemotherapy.

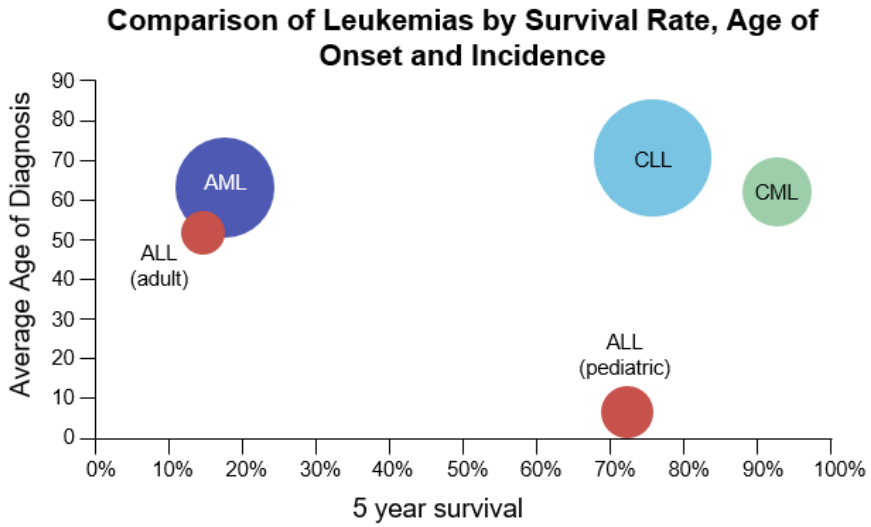

Figure 5: Different leukemia types and their associated survival rates, incidence and age of onset. *Average age of diagnosis in years and 5-year survival in percent is shown for different leukemia types. Circle size is proportional to incidence. Colors discriminate between the four major types of leukemia: chronic myeloid leukemia (CML), chronic lymphocytic leukemia (CLL), acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL). ALL is further separated into pediatric and adult cases. Adapted from Omar Abdel-Wahab, Overview of the Lymphoid Leukemias.*

Acute lymphoblastic leukemia (ALL) is a malignant transformation and proliferation of lymphoid precursors in the bone marrow. The incidence follows a bimodal distribution peak affecting children and adults around the age of 50 (Terwilliger and Abdul-Hay, 2017).

1.4.2 Epidemiology and genetic landscapes of T-ALL

ALL represents the most common malignancy of childhood and is differentiated in B cell (85%) and T cell (15%) lineage (Pui et al., 2008). T cell acute lymphoblastic leukemia (T-ALL) represents a clonal expansion of immature T cells and accounts for 10-15% of childhood and 20-25% of adult ALL cases (Ribera et al., 2007; Vadillo et al., 2018). T-ALL belongs to the precursor T cell neoplasms according to the latest WHO classification of hematolymphoid neoplasms (Alaggio et al., 2022). Precursor neoplasms include immature groups such as T cell lymphoblastic lymphoma (T-LBL)/T cell acute lymphoblastic leukemia (T-ALL) and Early T-precursor lymphoblastic leukemia/lymphoma (ETP-ALL). T-ALL and T-LBL only differ in the percentage of bone marrow infiltration but are considered as the same entity. ETP-ALL accounts for 10-13% of pediatric and 5-10% of adult T-ALL cases (Wenzinger et al., 2018). Despite important advances in the understanding of the genetic basis of T-ALL, in the latest WHO classification WHO-HAEM5 there still was not sufficient evidence for distinguishing clinically-relevant genetic subtypes of T-ALL (Alaggio et al., 2022). Adult T-ALL patients with relapsed disease have dismal outcomes with <10% of patients surviving long term and a 6-9 months median survival (Gökbuget et al., 2012).

The immunophenotype characteristic for T-ALL shows positivity for CD3 and TDT. Based on immunophenotyping, T-ALL can be further subdivided according to the stage of thymic maturation into early cortical, late cortical or mature T cell stage (Liu et al., 2017). Markers to differentiate between a pro/pre-T cell phenotype or rather a cortical or mature phenotype are CD7, CD2, CD1a and cyCD3 (Patel et al., 2012). CD4 and CD8 characterize mature T-ALLs and are usually not expressed in immature T-ALLs and ETP-ALL (Haydu and Ferrando, 2013; Wenzinger et al., 2018). The WHO subgroup ETP-ALL (early T cell precursor), has recently been described and is characterized by poor outcome, a distinct gene expression profile and immunophenotype similar to HSCs or myeloid progenitor cells (Coustan-Smith et al., 2009; Jain et al., 2016; Zhang et al., 2012). Within the ETP-ALL subtype, expression of oncogenic transcription factors as well as surface marker expression was highly variable (Coustan-Smith et al., 2009).

The number of somatic mutations is generally low in leukemias compared to solid cancers (Vogelstein et al., 2013). Despite this low number of mutations, leukemias are complex genetic diseases due to their multifaceted pattern of co-existing mutations, the functional interplay between mutated genes and the clonal heterogeneity and evolution (Papaemmanuil et al., 2016).

The genetic landscape of T-ALL is mainly characterized by gain-of-function mutations in genes of the NOTCH signaling pathway. More than 60% of T-ALL patient carry a *NOTCH1* mutation what is often accompanied by deletions of the *CDKN2A* locus (Liu et al., 2017; Weng et al., 2004). A genetic hallmark of T-ALL additionally is the activation of oncogenic transcription factors (discussed in chapter 1.5.2). Commonly affected pathways in T-ALL include transcriptional regulation (91%), cell cycle regulation (84%), NOTCH1 signaling (79%), epigenetic regulation (68%) and PI3K-AKT-mTOR signaling (29%) (Liu et al., 2017). The mutational spectrum of ETP-ALL is similar to myeloid tumors (Zhang et al., 2012) what can be explained by the retained ability of ETP cells to differentiate to the T cell and myeloid lineage (Wada et al., 2008). ETP-ALL is further characterized by less frequent *NOTCH1* mutations (compared to “classical” T-ALL) (Zhang et al., 2012).

Leukemias are often initiated by deregulation of transcriptional machinery including enhancers (Bhagwat et al., 2018). Human T-ALL cases harbor on average only six protein-coding, but almost 1000 non-coding mutations (Hu et al., 2017) showing the relevance for mutations in the non-coding, regulatory genome.

1.4.3 T cell development and cell of origin of genetic T-ALL subgroups

T cells and their very distinctive development in the thymus are well conserved and appear to be a signature feature of vertebrates (Rothenberg, 2019). T cell development is a segmented

process and starts with hematopoietic stem cells (HSCs) and multipotent progenitors (MPP) in the bone marrow. Common lymphoid progenitors (CLPs) and early T cell progenitors (ETPs) begin to migrate to the thymus but still keep features of their multilineage potential (Figure 6).

T cell fate in the thymus is predominantly promoted by the Notch pathway. The thymus consists of an outer cortex and an inner medulla region surrounded by a capsule and provides a very specific microenvironment for the different steps of T cell maturation. However, a complex interplay of gene regulatory networks and chromatin state changes is necessary that T cells gradually acquire their specific characteristics (Rothenberg, 2019). Especially transcription factors play a crucial and well-established role during this complicated process (explained in more detail in chapter 1.5.2). A very specific time point during T cell development is a step called ‘commitment’. While post-commitment T cells are functionally distinct from myeloid cells, pre-commitment T cell precursors express substantial levels of myeloid genes and still keep multilineage potential although discussed controversially (Schlenner and Rodewald, 2010). T cell development is characterized by pre-commitment stages where T cells lack both CD4 and CD8 expression, known as double negative (DN) cells. The double negative stage is further subdivided into four phenotypically distinct steps (DN1-DN4 by the expression of CD44, KIT/CD117 and CD25) (Rothenberg et al., 2008). Whereas in the early stages (DN1-DN2a) chromatin of multipotency sites is still open, at commitment (DN2b/DN3a) T cell sites open. Subsequently, these cells acquire CD4 and CD8 expression and become double positive cells (DP). Then, cells differentiate into CD4 and CD8 single positive cells (SP). Another characteristic of this very specific developmental step in the thymus is the extensive proliferation. Only few immature progenitor cells enter the thymus but they proliferate extensively.

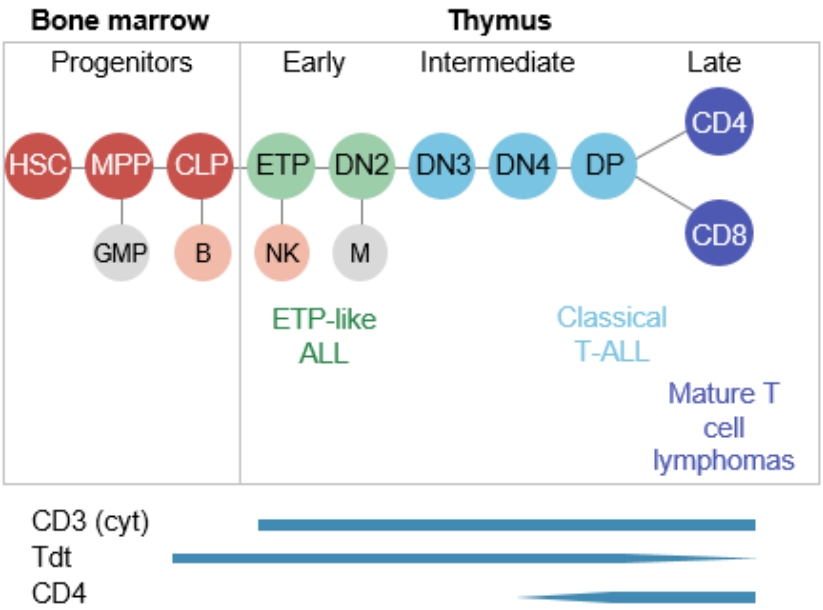

Figure 6: T cell development and T-ALL subtypes. *T cells develop from hematopoietic stem cells (HSCs) and multipotent progenitors (MPPs) in the bone marrow. Common lymphoid cells (CLPs) start to migrate to the thymus and further differentiate into double negative (DN) T cells. In the DN1, also called early T cell precursor (ETP), and DN2 stage, these immature T cells still show multilineage potential. Between DN2 and DN3 T cell commitment takes place. Further differentiation is characterized by CD4 and CD8 expression, first as double positive (DP) cells, later as single positive CD4 or CD8 cells. Early T cell precursor (ETP-) ALL develops from early DN stages, where cells are not yet committed to T cells. "Classical" T-ALLs develop from stages around or after T cell commitment. Mature T cell lymphomas originate from mature CD4 or CD8 cells. In immunohistochemistry, immature T cell malignancies are characterized by the expression of Tdt, while more mature forms of T-ALL express CD4 and/or CD8. Adapted from Fischer et al.*

Although later steps of T cell development contain crucial processes such as TCR rearrangement, positive and negative selection, activation and differentiation into diverse T cell subsets such as $CD4^+$ T helper cells or $CD8^+$ cytotoxic T cells, these concepts are beyond the scope of this thesis. The development of T cell leukemias is restricted to several steps before and shortly after T cell commitment.

Bringing this complex system of T cell development in context with disease remains, however, a major challenge. At the commitment stage, abrupt genome-wide changes of chromatin organization were found indicating a drastic change (Hu et al., 2018; Johnson et al., 2018). It is not yet clear how this chromatin reorganization contributes to T-ALL leukemogenesis, barring few examples (Kloetgen et al., 2020; Petrovic et al., 2019). The functional consequences of these genome-wide chromatin changes and the interplay of transcription factors still remain unclear and represent a focus of this study.

Different steps of T cell differentiation are also related to different cells of origin in T-ALL. Cell of origin (COO) is defined as the normal cell that acquires the first cancer-promoting mutation(s) (Visvader, 2011). Depending on the COO, tumors arising in the same organ can be very variable (intertumoral heterogeneity). This heterogeneity can be explained by either

different mutations occurring in the same cell or different subtypes developing from different cells of origin (Visvader, 2011). While for myeloid malignancies, the concept of leukemia initiating cells and the cell of origin are well-established, for ALL our understanding is limited (Lang et al., 2015). Especially for T-ALL, the cell of origin of different T-ALL subgroups is still discussed controversially (Berquam-Vrieze et al., 2011; Booth et al., 2018; Tan et al., 2017). The discovery of the ETP subgroup added an additional layer of complexity to this discussion. Model systems to analyze the cell of origin in the different subgroups of T-ALL are lacking. However, understanding the origins of T-ALL subtypes in more detail might be important for future treatment decisions.

1.5 Transcription factors in leukemogenesis

1.5.1 Transcriptional dysregulation in cancer development

Transcriptional dysregulation is a hallmark of nearly all types of cancer. While direct alterations to the DNA sequence are well reported due to enormous international sequencing efforts, the effect of genes not mutated, but dysregulated at the epigenetic or epitranscriptomic level, is still far from being understood. Cell identity and transcriptional homeostasis is controlled by the action of transcription factors (TFs), which directly interpret the genome (Lambert et al., 2018; Lee and Young, 2013). These genes specifically bind to genomic sequences thereby regulating gene expression. As transcription factors are (i) the core of developmental programs often hijacked by cancer cells, (ii) usually not mutated but dysregulated by other mechanisms and (iii) at the center of a cell's transcriptional circuit which can be exploited for therapy, they are crucial to understand cancer development but difficult to pinpoint with commonly used techniques. In previous screens, we observed a strikingly high number of CISs involved in transcriptional regulation (Weber et al., 2019).

1.5.2 Transcription factors in T cell and T-ALL development

T-cell development is characterized by a sequence of changes in transcriptional gene regulatory networks and chromatin states (as described in chapter 1.4.3). A core group of TFs has essential roles in this process by directly activating or repressing specific genes (Hosokawa and Rothenberg, 2021). Recent studies deciphered mechanisms of chromatin opening by transcription factor binding as well as differential co-binding and collaboration of transcription factors (reviewed in Rothenberg et al. (2019)). Specific expression patterns of TFs and successive chromatin changes that guide T-lineage commitment are beginning to be understood in more and more detail. As described above, T cells are dependent on a balance between precursor expansion and quality-controlled differentiation (Hosokawa and Rothenberg, 2021). Therefore, transcription factors have a highly stage-specific and context-

dependent role in T cell differentiation. In contrast to other hematopoietic lineages, for early T cell development there is no defined ‘master TF set’. Here, an ensemble of TFs rather acts in combination with each other and Notch signaling in a coordinated manner (Hosokawa and Rothenberg, 2021). Indispensable for an early phase of T cell development are factors including E2A, HEB, GATA3, TCF1, BCL11B, RUNX family, IKZF family and PU.1. The modular interaction of these TFs contributes to the irreversibility of T cell lineage commitment (Hosokawa and Rothenberg, 2021). However, the contribution of this complex system in context of T-ALL is not yet fully understood.

Finding novel, cancer-relevant but non-mutated or -translocated TFs is challenging. Although ALL is characterized by a rather small number of mutations (Hu et al., 2017), these disproportionally affect transcription factors (Inaba et al., 2013). In contrast to B-ALL, where mainly tumor-suppressive TFs are affected by translocations and mutations, the activation of oncogenic transcription factors is a hallmark of T-ALL. These include TAL1/2, LYL1, TLX1/3, NKX2-1/2-2/2-5, LMO1/2, MYB and MYC, which are commonly activated from rearrangement to T cell receptor loci (Liu et al., 2017). Genetic T-ALL subgroups are named according to their main TF rearrangement and can differ in their immunophenotype (Belver and Ferrando, 2016).

In human T-ALL, TAL1/SCL is one of the most prevalent oncogenic transcription factors (Brown et al., 1990). In the last decade the complex interplay between TAL1 and other T cell lineage transcription factors was beginning to be understood and core transcriptional regulatory circuits were revealed (Sanda et al., 2012) underlining the importance of TF networks in T-ALL pathogenesis.

1.5.3 SPIC as candidate leukemia transcription factor

The SPI-C Transcription Factor (*SPIC, SPI-1/PU.*1 Related) is an important paralog of *SPI-1* (PU.1) and belongs to the ETS (Erythroblast transformation specific) family of transcription factors. SPIC is mainly described to control the production of red-pulp macrophages (RPMs) in the spleen and required for red blood cell recycling and iron homeostasis (Kohyama et al., 2009). Monocyte differentiation into this specific type of macrophages is promoted by heme-mediated Spic induction (Haldar et al., 2014).

SPIC was also shown to influence B cell differentiation and immune response (DeKoter et al., 2010; Li et al., 2015). In early B cell development, SPIC is responsible for a RAG-induced transcriptional change (Soodgupta et al., 2019). In more detail, RAG induced double-strand breaks activate SPIC which recruits BCLAF1 to gene-regulatory elements (Soodgupta et al., 2019).

SPIC function is not described in the T cell lineage. However, the SPIC family member PU.1 (encoded by *SPI-1*) has a broad role in hematopoiesis. In T cells, PU.1 is expressed in early

stages of development and required for the generation of early T cell precursors (ETPs). At this stage, myeloid differentiation is restricted through NOTCH1 signaling in the thymus. However, PU.1 expression needs to be turned off for T cell lineage commitment (Ungerbäck et al., 2018).

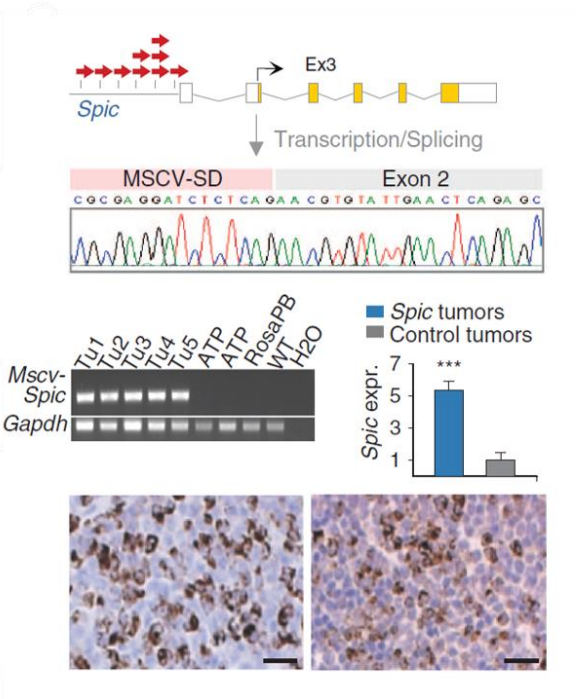

Figure 7: Spic as candidate oncogene in PiggyBac-induced acute myeloid leukemias*. Insertion pattern at the Spic locus. Insertions cluster sense-oriented close to the Spic promoter. Transposon-Spic fusion transcripts were confirmed by reverse transcription PCR and Sanger sequencing. Expression of Spic was analyzed using qPCR and compared to control tumors without Spic expression. Myeloid origin (MPO staining) of Spic driven tumors were shown in spleen and lymph nodes. Adapted from Rad et al., 2010*

Spic was found as common insertion site in the AML subgroup of the ATP2 transposon screen published earlier (Rad et al., 2010). Insertions were clustered close to the *Spic* promoter and transposon-Spic fusion transcripts as well as an increased *Spic* expression in samples with insertions was shown (Figure 7). In this published study, *Spic* insertions were exclusively associated with a myeloid origin of the leukemias.

In this thesis, *Spic* was validated as a leukemia oncogene using different mouse models.

1.6 Aims of this thesis

Despite crucial advances in the 'omics' field in last decades, tools to systematically study the function of the regulatory genome were still lacking. Accordingly, while significant progress has been made in the genetic characterization of T-ALL, patients are still treated with intensive chemotherapy protocols as genetic subgroups so far could not be related to the clinical course.

Transposon-based insertional mutagenesis screening became a powerful tool for the discovery of novel cancer genes. Although earlier studies described the preference of *PiggyBac* for open chromatin and punctuated analyses are available which identified regulatory elements in transposon screen, a systematic application was still missing. Additionally, *PiggyBac* screening was so far not applied to systematically study tumor evolution.

Therefore, this thesis aimed to investigate whether:

- the *PiggyBac* system is suitable for *in vivo* interrogation and functionalization of the non-protein-coding genome;
- a large cohort of hematopoietic tumors can be used to establish an annotation tool for functional regulatory common insertions sites;
- identified regulatory regions have functional relevance in leukemia cell lines;
- transposon insertions can be used to study T-ALL subtype-specific tumor evolution and differential sequentiality of driver genes;
- the screening system identifies novel leukemia subtype-specific transcription factors;
- engineered mouse models of a selected transcription factor recapitulate the leukemia phenotype observed in the screen.

The overall aim was to use transposon data to study subtle gene regulation and its cancer promoting role in a living organism. Different methodological bottlenecks were addressed and the first genome-wide *in vivo* screening approach interrogating the regulatory genome was described in this study. Studying T-ALL, large catalogues of cancer-relevant regulatory elements and non-coding transcripts were assembled – constituting the first survey of its kind for any cancer type. By thoroughly investigating the regulatory and evolutionary aspects in T cell leukemogenesis, this study aimed for a better understanding of T-ALL pathogenesis.

2. Material and Methods

2.1 Material

2.1.1 Technical Equipment

Table 2: Technical Equipment

Instruments	Source
BD FACSAria™ III High Sensitivity Flow Cytometer	BD Biosciences
Centrifuge 5424	Eppendorf
Centrifuge 5810 R	Eppendorf
Class II Biological Safety Cabinet	Thermo Fisher Scientific
CO_2-incubator Heracell™ VIOS 250i	Thermo Fisher Scientific
CyAn™ ADP Analyzer	Beckman Coulter
CytoFLEX LX Flow Cytometer	Beckman Coulter
Homogenisator Precellys® 24	Bertin Instruments
Incubator NCU-Line® IL 23	VWR International
MiSeq System	Illumina
NextSeq 550 System	Illumina
Nucleofector™ 2b Device	Lonza
NU-5500 Incubator	NuAire
Primovert Microscope	Carl Zeiss
Qubit® 2.0 Fluorometer	Thermo Fisher Scientific
scil Vet abc Plus™ Hematology Analyzer	Scilvet
StepOne Plus Real-Time PCR System	Applied Biosystems
Thermocycler TProfessional Basic 96	Biometra
ThermocyclerTProfessional Basic Gradient 96	Biometra
ThermoMixer® comfort 5355	Eppendorf
Ultra Low-Temperature Freezer Innova® U725	Eppendorf
UVsolo 2 Gel Documentation System	Analytik Jena
Vortex-Genie 2	Scientific Industries
Weighing Scale A120S	Sartorius

2.1.2 Consumables

Table 3: Consumables

Consumables	Source
ABgene Storage Plate, 96-well, 2.2 mL, square well, conical	Thermo Fisher Scientific
Adhesive PCR Plate Foils	Thermo Fisher Scientific
Biopsy/tissue embedding cassettes	Simport
Cell culture dishes (100 mm)	Greiner Bio-One
Cell culture flasks (50 mL, 250 mL, 550 mL)	Greiner Bio-One
Cell culture plates (6-well, 12-well, 24-well, 96-well)	Corning
Cell scrapers	Sarstedt
Cell strainers (70 µm, 100 µm)	Corning
Combitips advanced® (0.2 mL, 0.5 mL, 1 mL, 5 mL)	Eppendorf

Conical tubes (15 mL, 50 mL)	Greiner Bio-One
Cover slips	Gerhard Menzel B.V.
Cryotubes (1.6 mL)	Sarstedt
Disposable blades	Swann-Morton
Disposable reservoirs	Integra Biosciences
Disposable scalpels	B. Braun Melsungen
Disposable spatulas	Carl Roth
DNA LoBind Tubes (1.5 mL)	Eppendorf
Glass slides SuperFrost™ Plus	Thermo Fisher Scientific
Hard-Shell® 96-Well PCR Plates, high profile, semi skirted	Bio-Rad Laboratories
Hard-Shell® Low-Profile Thin-Wall 96-Well Skirted PCR Plate	Bio-Rad Laboratories
MicroAmp® optical 96-well reaction plate	Thermo Fisher Scientific
MicroAmp® Optical Adhesive Film	Thermo Fisher Scientific
Microtome blades S35	Feather Safety Razor
microTUBE AFA Fiber Snap-Cap 6x16mm Case	Covaris
Needles 27 gauge	Seidel medipool
Pasteur pipettes	Brand
PCR stripes (8 tubes)	Sarstedt
Petri dishes (100 mm)	Greiner Bio-One
Pipette tips (10 µL, 200 µL)	Biozym
Pipette tips with filter (10 µL, 100 µL, 200 µL, 300 µL, 1250 µL)	Biozym
Reaction tubes safe-seal (0.5 mL, 1.5 mL, 2 mL)	Sarstedt
Reaction tubes safe-seal (5 mL)	Eppendorf
S-Monovette®	Sarstedt
Serological pipettes (5 mL, 10 mL, 25 mL, 50 mL)	Greiner Bio-One
Syringes (1 mL, 30 mL)	B. Braun Melsungen

2.1.3 Reagents and enzymes

Table 4: Reagents and enzymes

Reagent/Enzyme	Source
1 kb DNA Ladder	New England Biolabs
100 bp DNA Ladder	New England Biolabs
2-Mercaptoethanol, 98%	Sigma-Aldrich
Acetic acid	Sigma-Aldrich
Agarose	Sigma-Aldrich
Ampicillin	Sigma-Aldrich
BbsI (10,000 units/mL)	New England Biolabs
BsaI-HF®v2 (20,000 units/ml)	New England Biolabs
Collagenase Type II	Worthington Biochemical
CutSmart Buffer	New England Biolabs
Deoxynucleotide Mix, 10 mM each	Sigma-Aldrich
Dimethyl sulfoxide (DMSO)	Carl Roth
DirectPCR Lysis Reagent (Cell)	Viagen Biotech

Doxycycline food 625 mg/kg	Ssniff
Eosine	Waldeck
Ethanol absolute	Carl Roth
Ethidium bromide	Sigma-Aldrich
Ethylenediaminetetraacetic acid (EDTA)	Sigma-Aldrich
Forene® isoflurane	Abbott
Formalin	Carl Roth
Gel Loading Dye, Purple (6x)	New England Biolabs
Glycerol	Sigma-Aldrich
Haematoxylin	Merck
Isopropanol absolute	Carl Roth
KAPA2G Fast Genotyping Mix	Sigma-Aldrich
LB-Agar (Luria/Miller)	Carl Roth
LB-Medium (Luria/Miller)	Carl Roth
NEBuffer 2	New England Biolabs
Phenol:Chloroform	Thermo Fisher Scientific
Phosphate buffered saline	Sigma-Aldrich
Polyethylene glycol 4000	Sigma-Aldrich
Propidium iodide	Thermo Fisher Scientific
Proteinase K	Sigma-Aldrich
Q5® High-Fidelity DNA Polymerase	New England Biolabs
RBC Lysis buffer (1x)	Thermo Fisher Scientific
RNAlater	Sigma-Aldrich
RNase-free DNase set	Qiagen
Roti®-Histofix 4%	Carl Roth
SuperScriptII	Thermo Fisher Scientific
SYBR® Select Master Mix	Thermo Fisher Scientific
T4 DNA Ligase	New England Biolabs
T4 DNA Ligase buffer	New England Biolabs
Taq DNA Polymerase	New England Biolabs
TaqMan™ Fast Advanced MM (1ml)	Thermo Fisher Scientific

2.1.4 Cell culture reagents

Table 5: Cell culture reagents

Reagent	Source
DMEM, high-glucose	Sigma-Aldrich
DPBS, no calcium, no magnesium	Thermo Fisher Scientific
FBS Superior	Biochrom
Penicillin-Streptomycin (5,000 U/ml)	Thermo Fisher Scientific
RPMI 1640 Medium	Thermo Fisher Scientific
Trypsin-EDTA (0.5%)	Thermo Fisher Scientific

2.1.5 Oligonucleotides

All oligonucleotides were synthesized by Eurofins Genomics.

Table 6: qPCR primer

qPCR Primer	Forward sequence	Reverse Sequence
Gapdh-murine	TGTGTCCGTCGTGGATCTGA	CACCACCTTCTTGATGTCATCATAC
Bcl11b-murine	GCCAGTGTGAGTTGTCAGGTAAA	GAACCAGGCGCTGTTGAAG
Pten-murine	TAACTGCAGAGTTGCACAGTATCC	CTTTACAGTGAATTGCTGCAACAT
Zeb1-murine	AGGTGATCCAGCCAAACG	GGTGGCGTGGAGTCAGAG
GAPDH-human	TGCACCACCAACTGCTTAGC	GGCATGGACTGTGGTCATGAG
RNLS-human	TGCAGCTTCAAGGTGACATC	CCCAGAGCATATCGAGAGGA
PTEN-human	GCAGAGTTGCACAATATCCTTTTG	CCAGCTTTACAGTGAATTGCTG
Ikzf1-murine	TGGACAGGCTGGCAAGCAAT	GTTGGCACTGTCATAGGGCA
Spic-murine	ATCCTCACGTCAGAGGCAAC	AAGAAGGGGGGTTGTACCAG

Table 7: Genotyping primer intergenic knockouts

KO Genotyping	Forward sequence	Reverse Sequence
KO_PtenEnh_mus	GTGGTATGCACAGTTGAGTG	GCACCAAACCCAAAGATTCA
KO_Gm10125_mus	GAGGGCTCTATGCTTGTTGA	GATCCCAATGAGTCACAGGT
KO_PTENenh_hum	TTTCTGAGTAGCTCATTTGTTTCCC	TTTTCATTATCACCCCATGTCCTC
KO_Gm11998_mus	CTCTGCAAATTACATGCCTGG	CCATGGAAGGACTGGGTATT

Table 8: CRISPR sgRNAs

sgRNA	Forward sequence	Reverse Sequence
sgLacZ	CACCGTGCGAATACGCCCACGCGAT	AAACATCGCGTGGGCGTATTCGCAC
Pten_Enh_mus_g1	CACCGAACAGCATTAGATCCACGTT	AAACAACGTGGATCTAATGCTGTTC
Pten_Enh_mus_g2	CACCGTTAGCAATCGGCCTGCTATG	AAACCATAGCAGGCCGATTGCTAAC
Pten_Enh_mus_g3	CACCGCTGCTGTGTTACTCATTAGC	AAACGCTAATGAGTAACACAGCAGC
Pten_Enh_mus_g4	CACCGCTTTTGGGCGATCCAACCCC	AAACGGGGTTGGATCGCCCAAAAGC
Pten_Enh_mus_g5	CACCGTTTTTAAGCAGGATCTCGTT	AAACAACGAGATCCTGCTTAAAAAC
Pten_Enh_mus_g6	CACCGGCTGTTCTTTAAGCAACCA	AAACTGGTTGCTTAAAGAACAGCC

Gm10125_mus_g1	CACCGTGATTTGATAGTACCACCTA	AAACTAGGTGGTACTATCAAATCAC
Gm10125_mus_g2	CACCGCCCCTTATTCTCTTACTAAC	AAACGTTAGTAAGAGAATAAGGGGC
Gm10125_mus_g3	CACCGGCCTACATGATTTGCATCA	AAACTGATGCAAATCATGTAGGCC
Gm10125_mus_g4	CACCGTGAACCTCGAGCCGCATACA	AAACTGTATGCGGCTCGAGGTTCAC
Gm10125_mus_g5	CACCGGAATTCTGACATACTCGAC	AAACGTCGAGTATGTCAGAATTCC
Gm10125_mus_g6	CACCGCTCAGCGAGCTCAGCGTTTG	AAACCAAACGCTGAGCTCGCTGAGC
PTEN_Enh_hum_g1	CACCGAGATGTGTTCCAATAGACGG	AAACCCGTCTATTGGAACACATCTC
PTEN_Enh_hum_g2	CACCGAATATTTTACCACCGTCTAT	AAACATAGACGGTGGTAAAATATTC
PTEN_Enh_hum_g3	CACCGTATTCATCAGCGGTGCTTTG	AAACCAAAGCACCGCTGATGAATAC
PTEN_Enh_hum_g4	CACCGATGCTTGGGGACAACTACAC	AAACGTGTAGTTGTCCCCAAGCATC
PTEN_Enh_hum_g5	CACCGATGATTAACAATTCTCAGTA	AAACTACTGAGAATTGTTAATCATC
PTEN_Enh_hum_g6	CACCGAGTCTTCAGTTAGTTTACAT	AAACATGTAAACTAACTGAAGACTC
Gm11998_mus_g1	CACCGCTAGCTAGAACACATCTCAC	AAACGTGAGATGTGTTCTAGCTAGC
Gm11998_mus_g2	CACCGATATTGACTGTCCCTTCCCA	AAACTGGGAAGGGACAGTCAATATC
Gm11998_mus_g3	CACCGGGAGGCACCTGTTCAGAGA	AAACTCTCTGAACAGGTGCCTCCC
Gm11998_mus_g4	CACCGCACATAAGTCAGGGGCATAT	AAACATATGCCCCTGACTTATGTGC
Gm11998_mus_g5	CACCGTTATATACCAAGATTGCAGC	AAACGCTGCAATCTTGGTATATAAC
Gm11998_mus_g6	CACCGAAAACTCTTATCAAATTAG	AAACCTAATTTGATAAGAGTTTTC

Table 9: Genotyping primer mouse lines

Genotyping primer	Forward sequence	Reverse sequence
Rosa26-LSL-Spic	ATCCCCATCAAGCTGATCC	GCGTTGCCTCTGACGTGAGG
TET-Spic	TAGGGTTAAAATCTAGATAGGCG TGTACGGTGGGAG	GCGTTGCCTCTGACGTGAGG
Vav-iCre	GGTGTTGTAGTTGTCCCCACT	CAGGTTTTGGTGCACAGTCA
Rosa26-rtTA3	GTTCGGCTTCTGGCGTGTGA	CGCTTGTTCTTCACGTGCCA
Rosa26-PB	GCTGGGGATGCGGTGGGCTC	GGCGGATCACAAGCAATAATAA CCTGTAGTTT
Rosa-26 (WT)	CTCTCCCAAAGTCGCTCTG	TACTCCGAGGCGGATCACAAGC
ATP2	CTCGTTAATCGCCGAGCTAC	GCCTTATCGCGATTTTACCA

2.1.6 Library preparation and sequencing

Table 10: Reagents for library preparation and sequencing

Reagent	Source
Agilent High Sensitivity DNA Kit	Agilent Technologies
EB Puffer	Qiagen GmbH
KAPA DNA standards	Kapa Biosystems
KAPA HiFi HotStart ReadyMix (2x)	Kapa Biosystems
KAPA SYBR Fast qPCR ABI Mix (2x)	Kapa Biosystems
MiSeq Reagent Kit v2 (300 cycle)	Illumina
NEBNext® Ultra DNA Library Prep Kit for Illumina®	New England Biolabs
NEBNext® Ultra II DNA Library Prep Kit for Illumina®	New England Biolabs
Nextera XT Kit	Illumina
Sodium hydroxide (NaOH)	Carl Roth

2.1.7 Plasmids

Table 11: Plasmids

Plasmid	Source
pX333 #64073	Addgene
pENTR1A no ccDB #17398	Addgene
pRosa26-DEST #21189	Addgene

2.1.8 Bacteria and Cell lines

Table 12: Bacteria and Cell lines

Cell line	Source
One Shot® Stbl3™ chemically competent E. coli	Thermo Fisher Scientific
EL4 ATCC® TIB-39™	ATCC
Jurkat ATCC® TIB-152™	ATCC
HEK293T ATCC® CRL-3216™	ATCC
Fibroblasts (primary)	Mouse lines

2.1.9 Mice

Table 13: Mouse lines

Mouse Strain	Source
PB	Rad et al., 2010
ATP2-S1	Rad et al., 2010
ATP2-H27	Rad et al., 2010
ATP2-H32	Rad et al., 2010
Rosa26-CAG-rtTA3 #:029627	The Jackson Laboratory
Vav-iCre #:008610	The Jackson Laboratory
Col1a1-TRE-Spic	unpublished
Rosa26-Spic	unpublished

2.1.10 Antibodies

Table 14: Antibodies

Antibody target	Conjugate	Company
	Flow cytometry	
CD8a	PE	Invitrogen
CD4a	PE Cy7	eBioscience
B220	FITC	Invitrogen
CD11b	APC Cy7	Invitrogen
TER119	PE Cy5.5	Invitrogen
CD48	Biotin	Invitrogen
Gr1	PB	eBioscience
Streptavidin	PB	eBioscience
CD150	PE	Invitrogen
Sca1	PE Cy7	eBioscience
c-Kit	APC	eBioscience
CD34	FITC	eBioscience
	Immunohistochemistry	
Primary Antibodies		
Rat anti-B220/CD45R		BD Bioscience
Rabbit anti-CD3 (Sp7)		DCS

Rabbit anti-MPO (A0398)	DAKO
Rat anti-CD138 (281-2)	BD Bioscience
Rabbit anti-Tdt (005)	Supertechs
Rat anti-CD4 (GHH4)	Dianova

Secondary Antibodies	
AffiniPure Goat Anti-Rabbit IgG (H+L) (111-005-003)	Jackson ImmunoResearch
AffiniPure Rabbit Anti-Rat IgG (312-005-045)	Jackson ImmunoResearch

2.1.11 Kits

Table 15: Kits

Kit	Source
AllPrep DNA/RNA Mini	Qiagen
Amaxa® Cell Line Nucleofector® Kit L	Lonza
Amaxa® Cell Line Nucleofector® Kit V	Lonza
DNeasy Blood & Tissue Kit	Qiagen
MinElute Reaction Cleanup Kit	Qiagen
mirVana™ miRNA Isolation Kit	Thermo Fisher Scientific
QIAprep Spin Miniprep Kit	Qiagen
QIAquick Gel Extraction Kit	Qiagen
QIAquick PCR Purification	Qiagen
Qubit® dsDNA BR Assay Kit	Thermo Fisher Scientific
Qubit® RNA BR Assay Kit	Thermo Fisher Scientific
RNeasy Plus Mini Kit	Qiagen
TaqMan™ Advanced miRNA cDNA Synthesis Kit	Thermo Fisher Scientific

2.1.12 Databases and Software

Table 16: Software, databases and programs

Software/Database/Program	Source
CIMPL	https://github.com/NKI-CCB/cimpl
dbSUPER	https://asntech.org/dbsuper/
deepTools	https://github.com/deeptools/deepTools
FlowJo Version 10.2	FlowJo, LLC
Genomic Workbench 7	Agilent Technologies
GSEA v4.0.3	Broad Institute
GWAS catalogue	https://www.ebi.ac.uk/gwas/
Immgen	https://www.immgen.org/
Inkscape	https://inkscape.org/
Office 2016	Microsoft Corporation
R Software Environment 4.0.1	The R Project, The R Foundation

Snapgene 5.0.8	GSL Biotech
StepOne v2.3	Thermo Fisher Scientific

Table 17: R packages

R package	Version
BiocManager	1.30.10
biomaRt	2.44.0
Circlize	0.4.9
Cola	2.0.0
ComplexHeatmap	2.4.2
data.table	1.12.8
DeSeq2	1.28.1
devtools	2.3.1
qdapTools	1.3.5
GenomeInfoDbData	1.24.0
GenomicRanges	1.40.0
ggforce	0.3.1
ggplot2	3.3.1
ggpubr	0.4.0
ggrepel	0.8.2
pheatmap	1.0.12
S4Vectors	0.26.1
scales	1.1.1
Survival	3.1-12
Survminer	0.4.7
tidyverse	1.3.0
tidyr	1.1.0
RColorBrewer	1.1-2
rio	0.5.16

2.1.13 Publicly available datasets

Publicly available (epi-)genomic datasets used in this study.

Table 18: Publicly available murine datasets

Dataset murine	Accession number	Publication
Dnase-Seq HSC	GSE79422	Hu et al., Immunity 2018
Dnase-Seq MPP	GSE79422	Hu et al., Immunity 2018
Dnase-Seq CLP	GSE79422	Hu et al., Immunity 2018
Dnase-Seq ETP	GSE79422	Hu et al., Immunity 2018
Dnase-Seq DN2	GSE79422	Hu et al., Immunity 2018
Dnase-Seq DN3	GSE79422	Hu et al., Immunity 2018
Dnase-Seq DN4	GSE79422	Hu et al., Immunity 2018
Dnase-Seq DP	GSE79422	Hu et al., Immunity 2018
RNA-Seq HSC	GSE79422	Hu et al., Immunity 2018

RNA-Seq MPP	GSE79422	Hu et al., Immunity 2018
RNA-Seq CLP	GSE79422	Hu et al., Immunity 2018
RNA-Seq ETP	GSE79422	Hu et al., Immunity 2018
RNA-Seq DN2	GSE79422	Hu et al., Immunity 2018
RNA-Seq DN3	GSE79422	Hu et al., Immunity 2018
RNA-Seq DN4	GSE79422	Hu et al., Immunity 2018
RNA-Seq DP	GSE79422	Hu et al., Immunity 2018
Hi-C HSC	GSE79422	Hu et al., Immunity 2018
Hi-C MPP	GSE79422	Hu et al., Immunity 2018
Hi-C CLP	GSE79422	Hu et al., Immunity 2018
Hi-C ETP	GSE79422	Hu et al., Immunity 2018
Hi-C DN2	GSE79422	Hu et al., Immunity 2018
Hi-C DN3	GSE79422	Hu et al., Immunity 2018
Hi-C DN4	GSE79422	Hu et al., Immunity 2018
Hi-C DP	GSE79422	Hu et al., Immunity 2018
H3K27ac DP	GSE61428	Ing-Simmons et al., Genome Research 2015
H3K4me1 DP	GSE20898	Wei et al., Immunity 2011
ChrAccess.increase	GSE79422	Hu et al., Immunity 2018
ChrAccess.decrease	GSE79422	Hu et al., Immunity 2018
EL4 CTCF	GSE66343	Ren et al., Molecular Cell 2017
EL4 H3K27ac	GSE125384	Sidoli et al., Scientific Reports 2019
EL4 H3K36me2	GSE125384	Sidoli et al., Scientific Reports 2019
EL4 H3K27me3	GSE125384	Sidoli et al., Scientific Reports 2019
EL4 H3K9ac	GSE125384	Sidoli et al., Scientific Reports 2019
EL4 H3K4me4	GSE125384	Sidoli et al., Scientific Reports 2019
EL4 ATAC	GSE125384	Sidoli et al., Scientific Reports 2019
EL4 RNAseq	GSE125384	Sidoli et al., Scientific Reports 2019
Thymus CTCF	GSE49847	Yue et al., Nature 2014
Thymus POL2RA	GSE49847	Yue et al., Nature 2014
Thymus H3K36me3	GSE49847	Yue et al., Nature 2014
Thymus H3K4me1	GSE49847	Yue et al., Nature 2014
Thymus H3K4me3	GSE49847	Yue et al., Nature 2014
Thymus H3K27ac	GSE49847	Yue et al., Nature 2014
Thymus H2K27me3	GSE49847	Yue et al., Nature 2014
H3k27ac_DP1	GSE79422	Hu et al., Immunity 2018
H3k27ac_DP2	GSE79422	Hu et al., Immunity 2018
H3K27ac_LT_HSC	GSE59636	Lara-Astiaso et al., Science 2014
H3K27ac_ST_HSC	GSE59636	Lara-Astiaso et al., Science 2014
H3K27ac_MPP	GSE59636	Lara-Astiaso et al., Science 2014
H3K27ac_CLP	GSE59636	Lara-Astiaso et al., Science 2014
H3K27ac_CD4	GSE59636	Lara-Astiaso et al., Science 2014
H3K27ac_CD8	GSE59636	Lara-Astiaso et al., Science 2014
ATAC-Seq-DP-1	GSE99159	Johnson et al., Immunity 2018
ATAC-Seq-DP-2	GSE99159	Johnson et al., Immunity 2018

ATAC-Seq-BM-HSC-1	GSE77695	Shih et al., Cell 2016
ATAC-Seq-BM-HSC-2	GSE77695	Shih et al., Cell 2016
ATAC-Seq-BM-MPP-1	GSE77695	Shih et al., Cell 2016
ATAC-Seq-BM-MPP-2	GSE77695	Shih et al., Cell 2016
ATAC-Seq-BM-CLP-1	GSE77695	Shih et al., Cell 2016
ATAC-Seq-BM-CLP-2	GSE77695	Shih et al., Cell 2016
ATAC-Seq-BM-B-1	GSE77695	Shih et al., Cell 2016
ATAC-Seq-BM-B-2	GSE77695	Shih et al., Cell 2016
ATAC-Seq-BM-NK-1	GSE77695	Shih et al., Cell 2016
ATAC-Seq-BM-NK-2	GSE77695	Shih et al., Cell 2016
ATAC-Seq-SP-CD4-1	GSE77695	Shih et al., Cell 2016
ATAC-Seq-SP-CD4-2	GSE77695	Shih et al., Cell 2016
ATAC-Seq-SP-CD8-1	GSE77695	Shih et al., Cell 2016
ATAC-Seq-SP-CD8-2	GSE77695	Shih et al., Cell 2016
ATAC-Seq-ETP-T-1	GSE100738	Yoshida et al., Cell 2019
ATAC-Seq-ETP-T-2	GSE100738	Yoshida et al., Cell 2019
ATAC-Seq-DN2a-T-1	GSE100738	Yoshida et al., Cell 2019
ATAC-Seq-DN2a-T-2	GSE100738	Yoshida et al., Cell 2019
ATAC-Seq-DN2b-T-1	GSE100738	Yoshida et al., Cell 2019
ATAC-Seq-DN2b-T-2	GSE100738	Yoshida et al., Cell 2019
ATAC-Seq-DN3-T-1	GSE100738	Yoshida et al., Cell 2019
ATAC-Seq-DN3-T-2	GSE100738	Yoshida et al., Cell 2019
ATAC-Seq-DN4-T-1	GSE100738	Yoshida et al., Cell 2019
ATAC-Seq-DN4-T-2	GSE100738	Yoshida et al., Cell 2019

Table 19: Publicly available human datasets

Dataset human	Accession number	Publication
GRO-Seq-HEK293T	GSE92375	Bouvy-Liivrand et al., Nucleic Acids Res. 2017
GRO-Seq-Jurkat	EGAS00001005864	Fischer et al.
GRO-Seq-T-ALL-patient1	EGAS00001005864	Fischer et al.
GRO-Seq-T-ALL-patient1	EGAS00001005864	Fischer et al.

2.1.14 Manufacturers

Table 20: Manufacturers

Manufacturer	Location
Abbott GmbH	Ludwigshafen, Germany
Addgene	Cambridge, Massachusetts, USA
Agilent Technologies, Inc.	Santa Clara, CA, USA
Analytik Jena AG	Jena, Germany
Applied Biosystems, Inc.	Carlsbad, CA, USA
ATCC	Manassas, VA, USA
B. Braun Melsungen AG	Melsungen, Germany

BD Biosciences, BD, Inc.	Franklin Lakes, NJ, USA
Beckman Coulter	Pasadena, CA, USA
Bertin Instruments	Montigny-le-Bretonneux, France
Biochrom GmbH	Berlin, Germany
Biometra GmbH	Göttingen, Germany
Bio-Rad Laboratories, Inc.	Hercules, CA, USA
Biozym Scientific GmbH	Hessisch Oldendorf, Germany
Brand GmbH	Wertheim, Germany
Carl Roth	Karlsruhe, Germany
Carl Zeiss AG	Oberkochen, Germany
Corning, Inc.	Corning, NY, USA
Covaris, Inc.	Woburn, MA, USA
DAKO, Agilent Technologies, Inc.	Santa Clara, CA, USA
DCS	Hamburg, Deutschland
Dianova	Hamburg, Deutschland
eBioscience	San Diego, CA, USA
Eppendorf AG	Hamburg, Germany
Eurofins Genomics GmbH	Ebersberg, Germany
Feather Safety Razor Co., Ltd.	Osaka, Japan
Greiner Bio-One GmbH	Kremsmünster, Austria
GSL Biotech LLC	Chicago, IL, USA
Illumina, Inc.	San Diego, CA, USA
Invitrogen	Carlsbad, CA, USA
Integra Biosciences AG	Biebertal, Germany
Jackson ImmunoResearch, Inc.	West Grove, PA, USA
Leica Biosystems	Nußloch, Germany
Lonza	Basel, Switzerland
Merck KGaA	Darmstadt, Germany
Microsoft Cooperation	Redmond, WA, USA
New England Biolabs, Inc.	Ipswich, MA, USA
NuAire	Plymouth, MN, USA
Qiagen GmbH	Hilden, Germany
R&D Systems, Inc.	Minneapolis, MN, USA
Sarstedt AG	Nümbrecht, Germany
Sartorius AG	Göttingen, Germany
Scientific Industries, Inc.	Bohemia, NY, USA
Scil animal care company Scilvet	Viernheim, Germany
Seidel medipool GmbH	Gauting-Buchendorf, Germany
Sigma-Aldrich Corporation	St. Louis, MO, USA
Simport Scientific, Inc.	Beloeil, QC, Canada
Ssniff	Soest, Germany
Supertechs	Rockville, Maryland, USA
Swann-Morton, Ltd.	Sheffield, United Kingdom
Thermo Fisher Scientific, Inc.	Waltham, MA, USA
Vectorlabs	Burlingame, CA, USA
Viagen Biotech, Inc.	Los Angeles, CA, USA
VWR International GmbH	Darmstadt, Germany
Waldeck GmbH	Münster, Germany
Worthington Biochemical Corporation	Lakewood, NJ, USA

2.2 Methods

All procedures were performed according to manufacturer's instructions unless specified otherwise.

2.2.1 Generation of mouse models

The generation of transposon mice and available lines was described earlier (Rad et al., 2010). Transposon mice were kept in the animal facilities of the Wellcome Trust Sanger Institute, Hinxton/Cambridge, UK under specific-pathogen-free conditions on a 12-h light/dark cycle, receiving food and water ad libitum. Experimental (*ATP2;Rosa26PB/+*) and control (*Rosa26PB/+* and *ATP2* single transgenic) mice were maintained on a mixed C57BL/6 x 129Sv x FVB background. Different ATP2 lines were used to generate final cohorts, which differ in their number of transposon copies and the donor locus (ATP2-S1: donor locus chr17, 15 copies; ATP2-H27: donor locus chr4, 20 copies; ATP2-H32: donor locus chr2, 25 copies). *Necropsy of transposon mice was performed by Roland Rad, Lena Rad and Alexander Strong.*

To generate Rosa26-Spic knock-in mice (*Gt(ROSA)26Sortm1(Spic)Rrad*), the *Spic* sequence (murine open reading frame) was cloned into a Gateway-compatible entry vector (Addgene #17398), which was then shuttled into a Rosa26-targeting Gateway destination vector with loxP-flanked puromycin resistance-containing stop cassette (modified after Addgene #21189). Embryonic stem (ES) cell (JM8) targeting, blastocyst injections, and subsequent breeding steps were performed using standard protocols/techniques. Crossing of Rosa26-Spic mice with a Cre recombinase leads to tissue-specific expression of *Spic*.

To generate inducible *Spic* mice, the TET system was used (TET-Spic mice). The *Spic* sequence was first cloned into the pENTR1A vector (Addgene #17398) and then shuttled to the Col1a targeting vector containing a minimal CMV promoter together with the Tet response elements (TRE) cassette (under a PGK promoter). Using recombinase mediated cassette exchange embryonic stem cells with a modified *Col1a* locus (MESKH2-VJ1, Jaenisch cells, (Wu et al., 1994)) were targeted. Crossing of TET-Spic mice with reverse transactivator mice (rtTA3) will lead to *Spic* expression only after doxycycline administration. Doxycycline containing food (625 mg/kg) was administered to TET-Spic rtTA3 double transgenic animals starting at 8 weeks of age until the termination criteria of the experiment were reached. *Cloning of Spic sequences and generation of ES cell was performed by Rupert Öllinger.*

Blastocyst injections were performed at Wellcome Trust Sanger Institute, Hinxton/Cambridge, UK (*by Mathias Friedrich and Allan Bradley*). Mice were kept in the animal facilities of the Klinikum rechts der Isar, Technical University Munich, Munich, Germany. All animal experiments were carried out in compliance with the requirements of the European guidelines for the care and use of laboratory animals and were approved by the Technical University

Munich (Regierung von Oberbayern, Munich, Germany, license number AZ ROB-55.2Vet-2532.Vet_02_17-84). *For a small subset of mice, genotyping (3/31) and necropsy (1/31) was performed by Majdaddin Rezaei. All other mice were genotyped and analyzed by myself.*

2.2.2 Necropsy and histology

All animals were monitored regularly and all signs of sickness (e.g., inactivity, pale paws, hunched posture, palpable/visible masses and poor grooming) were reported. Mice were anesthetized with Isoflurane (Abbott) before being euthanized. During necropsy, a thorough inspection of all hematological organs (thymus, spleen, lymph nodes) was carried out. For later DNA/RNA isolation, tissue samples were stored in RNAlater (Sigma). For histology, tissue samples were fixed in 4% formaldehyde, paraffin-embedded, sectioned, and stained using hematoxylin and eosin following standard protocols. Blood was drawn from the heart and stored in S-Monovettes (Sarstedt) at -20°C. Bones (femur and tibia) were isolated and stored in FACS buffer until bone marrow isolation. Ear and/or tail samples for later re-genotyping or germline control DNA was also stored in RNAlater.

2.2.3 Immunohistochemistry

Immunohistochemistry (IHC) was performed on a Bond Rxm (Leica) using a Polymer Refine detection kit without post-primary antibody. Slides were deparaffinized and pretreated with Epitope retrieval solution 1 (ER1, citrate buffer, pH = 6) or solution 2 (ER2, EDTA buffer, pH = 9) as indicated. The following primary antibodies were used: rat anti-B220/CD45R (B220, BD Bioscience, 1:50 dilution, ER1, 20 min), rat anti-CD138 (281-2, BD Bioscience, 1:50, ER2, 20 min), rat anti-MPO (A0398, DAKO, 1:100, ER2, 20 min), rabbit anti-CD3 (Sp7, DCS, 1:100, ER1, 20 min), rabbit anti-Tdt (005, Supertechs, 1:100, ER2, 20 min) and rat anti-CD4 (GHH4, Dianova DIA-404, 1:50, ER2, 40 min). Rabbit anti-rat secondary antibody (Vector, 1:400) was applied for primary rat antibodies. Slides were counterstained with hematoxylin and coverslipped after manual rehydration. Slides were scanned with a Leica AT2 scanning system. HE stainings and IHCs were evaluated by experienced mouse pathologists. *Professor Leticia Quintanilla de Fend and her team at University Tübingen histopathologically analyzed the large ATP2 cohort and provided the first diagnoses. Subsequent in detail analysis of T cell leukemias was performed by Dr.med.vet. Hsi-Yu Yen and PD Dr.med.vet. Katja Steiger (Pathology department, Klinikum rechts der Isar, TUM). All pathologists were blinded to the mouse genotypes.*

2.2.4 DNA/RNA isolation from tissue and cell lines

DNA and RNA from RNAlater stored tissue samples were isolated using the Qiagen Allprep DNA/RNA Mini Kit according to manufacturer's instructions. miRNA isolation of tissue samples

was performed using the mirVanaTM miRNA Isolation Kit (Thermo Fisher Scientific) according to manufacturer's instructions. DNA and RNA concentrations were measured using the Qubit® fluorometer.

2.2.5 Quantitative insertion site sequencing (QiSeq)

To obtain the exact location as well as the abundance of transposon insertions in the genome, we previously developed a semi-quantitative sequencing approach (QiSeq, (Friedrich et al., 2017)). Briefly, DNA samples were sheared with a Covaris AFA sonicator to a mean fragment length of 250 bp. The fragmented DNA was then end-repaired, A-tailed and a splinkerette adapter was ligated to each DNA end. For the 5' and 3' transposon end, subsequent steps (amplification and sequencing of transposon-genome junctions) were conducted separately. The specific structure of the splinkerette adapter (Y-shaped design with a template and a hairpin strand) ensures that only transposon-genome junction fragments (and not genomic fragments without transposon insert) can be amplified in the following first PCR step (which was conducted with transposon- and splinkerette-specific primers). Afterwards, a second nested PCR step was performed for further amplification, barcoding of samples and extension with Illumina flow cell binding sites P5 and P7. Each sample was then quantified with quantitative real-time PCR (using P5- and P7-specific primers). Subsequently, samples were equimolarly mixed and the library pool was again quantified. Libraries were sequenced on the Illumina MiSeq sequencer (75 bp, paired-end). Mapping of integrations to the mouse genome was performed using the SSAHA2 algorithm and sequences containing transposon-genome junctions were selected for downstream analyses. *For these analyses, the Wellcome Trust Sanger Institute Bioinformatics Pipeline was used (scripts generated by Hannes Ponstingl and Mathias Friedrich).*

2.2.6 Common insertion sites (CISs) and downstream analysis

To identify regions in the genome that are more frequently hit than observed by change, we performed CIMPL (Common Insertion Site Mapping Platform) CIS analysis, which is based on a Gaussian kernel convolution framework (de Ridder et al., 2006). Therefore, raw files resulting from QiSeq analysis were processed. Processing included the removal of insertions 3 Mb around the donor locus (chr17) and the application of a read coverage cut off. CISs were ranked according to the number or samples of contributing insertions. *Sfi1*, a known artefact frequently detected in insertional mutagenesis screens, was removed from the list of CIS genes. Additionally, *Arid1b* and *Mmp16* were excluded due to their close proximity to the donor locus on chr17 and chr4, respectively. A scale parameter of 30 k was used for CIS identification. Profile plots and profile heatmap plots for visualization of ChIP-Seq peak enrichment in CIS regions were created using deeptools (Ramírez et al., 2016). Subgroup

specific CIS analysis were performed using a scale parameter of 5 k and were ranked according to the number of contributing insertions.

2.2.7 Annotation of regulatory common insertion sites

The computational part of the ARCIS tool (data preprocessing, chromHMM run and overlapping using the GenomicRanges package) was conducted by Niklas de Andrade Krätzig. Analysis strategies, ranking scores and manual inspection rules were developed by myself.

2.2.7.1 Computational pipeline to annotate CISs

For the identification of regulatory regions using CIMPL, the scale parameter was set to 5 k to identify narrow regions with regulatory potential. The resulting CIS coordinates were overlapped with a collection of publicly available datasets using the GenomicRanges R package (Lawrence et al., 2013). The data was post-processed into a BED3 format with an additional column for name assignment. For overlap with peak-based files (ChIP-Seq, DNase-Seq), the number of overlapping peaks and the distance to the closest peak were reported. For interaction datasets (Hi-C from different stages of development and data from dbSUPER [mouse thymus]), linked target genes are annotated. Chromatin regions specifically increasing or decreasing during T cell development (change in A and B compartment scores called from Hi-C data in Hu et al. (2018)) were also overlapped with CIS regions.

Additionally, we run a chromatin Hidden Markov Model (chromHMM) (Ernst and Kellis, 2012) with a collection of ChIP-Seq data from ENCODE to define chromatin states, based on distinct combinations of histone marks. We used six thymus-specific ChIP-Seq datasets: H3K4me1, H3K27ac, H3K4me3, H3K27me3, H3K36me3 and CTCF. The observed chromatin combinations resulted in eight manually assigned chromatin states: active/weak/poised/insulated enhancer, active promotor, gene body, CTCF binding sites and quiescent. The number of overlaps as well as the distance to the closest element was reported.

ARCIS calculates a score for each CIS region. The score is based on selected elements: overlap with an annotated super-enhancer (dbSUPER) or increasing chromatin accessibility in T cell development (from Hu et al. (2018)) was scored with +1, a weak and active enhancer (chromHMM) was scored with +1 and +2, respectively, any reported Hi-C connection was scored with +2. To avoid that inactivated PC genes are reported as intronic enhancer, the overlap with more than two exons was penalized with -2. The sum represents the final ARCIS RE-score. If the ARCIS RE-score is greater than or equal to 3, the CIS is putatively harboring a regulatory region. CIS can be ranked based on the score. To report the type of the element, the GENCODE M24 annotation type is used. According to the type, the CIS gets annotated as intergenic enhancer, as non-coding transcript, which is potentially overlapping with an

enhancer or as protein-coding transcript, which might additionally harbor an intragenic or an intergenic RE in close proximity. The category ‘non-coding’ is defined by the overlap with an annotated nPC transcript but does not exclude the presence of an additional enhancer element. In contrast, the annotation as enhancer element cannot exclude the presence of so far not annotated nPC transcripts. Assessing the expression of identified nPC-transcripts represents further implication for the nPC transcript to play a functional role.

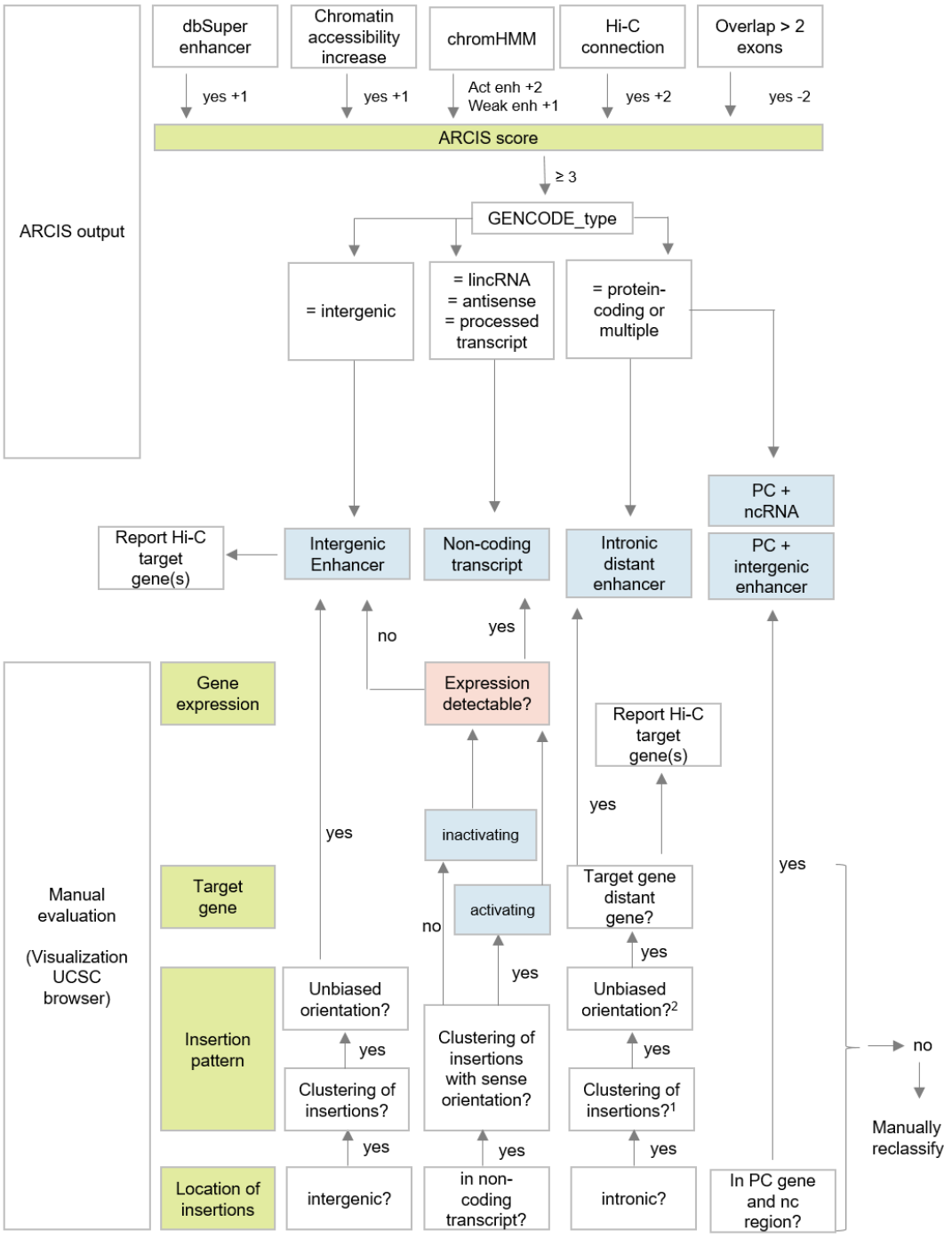

Figure 8: Workflow for the identification of regulatory CISs using ARCIS and manual evaluation.

2.2.7.2 Manual refinement and classification of regulatory CISs

The manual evaluation was based on the visualization of different datasets in the UCSC genome browser (Figure 8). Here, the exact position of the insertions, their orientation and clustering within the CIS were inspected. In a first step, it was assessed whether or not the majority of the insertions overlaps with an intergenic/intronic region, non-coding transcript or with multiple types. For intergenic enhancers, next the clustering of the insertions was investigated, a characteristic for insertion peaks in REs. To exclude that this insertion peak is activating a transcript, we assessed the orientation of the insertions. If the orientation is unbiased, it is highly unlikely that these insertions activate a transcript. If any of these steps leads to a 'no' answer meaning that either insertions are not clustered or the orientation is biased, the CIS needs to be manually reclassified to the better matching category. If all steps were answered with 'yes', the target gene of the RE is reported based on the visible Hi-C connection. An optional step includes the investigation if the reported target gene is expressed in T cell development (based on RNAseq data from Hu et al. (2018)). A target gene expressed at any stage in T cell development represents a highly interesting candidate. For intronic REs, the steps are comparable. If insertions are not clustered, the PC gene most likely gets inactivated by the transposon insertions, classifying this CIS as PC. If the insertions are clustered, but show strong orientation bias, the transposon might activate the PC gene (based on availability of a transcript with an ATG in exon 2 or higher). These CISs were classified as activating PC. If all steps were answered with 'yes', it needs to be assessed if the target gene is a distant gene (not connected to own promoter). If so, the target gene is reported. An optional step includes the investigation if the CIS gene (not the target gene) is expressed in T cell development. For genes not expressed, an inactivating function of the transposon insertions is highly unlikely, increasing the confidence to classify this CIS as intronic regulatory element. For non-coding transcripts, potentially activating or rather inactivating patterns were differentiated based on insertion clustering and orientation. An optional step includes the investigation if the non-coding transcript is expressed in T cell development. Of note, nPC transcripts are expressed at low levels compared to PC genes. Identifying a transcript with detectable expression therefore increases the confidence that the nPC transcript is relevant for T cell biology. If insertions overlap with multiple transcripts (PC and nPC) or with a PC transcript and the intergenic area, these CIS were classified as 'PC + ncRNA' or 'PC + intergenic enhancer', respectively. Ambiguity in the ARCIS output often cannot be resolved by manual annotation. Here, functional validation is necessary.

2.2.8 RNA-Seq

Library preparation for bulk 3' poly(A)-RNA sequencing was done as described previously (Parekh et al., 2016). Briefly, barcoded cDNA of each sample was generated with a Maxima

RT polymerase (Thermo Fisher) using oligo-dT primer containing barcodes, unique molecular identifiers (UMIs) and an adaptor. Ends of the cDNAs were extended by a template switch oligo (TSO) and full-length cDNA was amplified with primers binding to the TSO-site and the adaptor (Weber et al., 2019). NEB UltraII FS kit was used to fragment cDNA. After end repair and A-tailing a TruSeq adapter was ligated and 3'-end-fragments were finally amplified using primers with Illumina P5 and P7 overhangs. In comparison to Parekh et al. (2016), the P5 and P7 sites were exchanged to allow sequencing of the cDNA in read1 and barcodes and UMIs in read2 to achieve a better cluster recognition (Weber et al., 2019). The library was sequenced on a NextSeq 500 (Illumina) with 63 cycles for the cDNA in read1 and 16 cycles for the barcodes and UMIs in read2.

For data analysis, Gencode gene annotations M25 and the mouse reference genome GRCm38 were derived from the Gencode homepage (EMBL-EBI). Data was processed using the published Drop-Seq pipeline (v1.12) to generate sample- and gene-wise UMI tables (Macosko et al., 2015). The resulting UMI filtered count matrix was imported into R v3.4.4. Lowly expressed genes were filtered so that 80% of samples have at least three read counts per gene. Dispersion of the data was estimated with an intercept only model using DESeq2 v1.18.178 (Love et al., 2014). Principal Component Analysis (PCA) was conducted with the 10 percent top variable genes in the rlog transformed dataset. PAM algorithm (k parameter set to 4) was used to determine cluster membership in the PCA embedding (clusters were confirmed using the cola R package (Gu et al., 2021)). Cluster assignments were then used as explanatory variable during model fitting with DESeq2. The Wald test was used for determining differentially regulated genes between all pairwise clusters. Shrunken log2 fold changes were calculated afterwards. Rlog transformation of the data was performed for visualization and further downstream analysis.

Bioinformatic analyses were performed by Thomas Engleitner and myself.

2.2.9 aCGH for copy number analysis

Array comparative genomic hybridisation (aCGH) was carried out by the group of Kristian Unger (Helmholtz Zentrum München, Neuherberg/München, Germany). For this, Agilent 60k mouse CGH arrays were used. CGH data was pre-processed with the Agilent Genomic Workbench software v7.0.4.0. Raw log ratios were re-centered to ensure that the zero point is reflecting the most common ploidy state. Segmentation and aberration calling were performed with the implemented ADM-2 algorithm. Visualization of curated data was performed in R.

2.2.10 Gene set enrichment analysis

Gene set enrichment analysis (GSEA) was performed using the GSEA program (v4.0.3). The preranked mode with the apeglm shrunken log2 fold changes as ranking metric was used. Hallmark gene sets (h.all.v7.2.symbols.gmt) were selected for pathway analysis. Hematopoietic gene signatures were obtained from Laurenti et al. (2013) (http://www.jdstemcellresearch.ca/node/32) and Novershtern et al. (2011). A pathway was considered to be significantly associated with an experimental condition if the FWER was below 0.05.

2.2.11 cDNA synthesis and qPCR

For mRNA, cDNA synthesis was conducted using SuperScript II Reverse Transcriptase (Thermo Fisher Scientific). A total of 1 µg RNA and a mixture of oligo dT primers and random hexamers was used for cDNA synthesis, which was performed according to standard protocols. Real-time qPCR was conducted with SYBR Select Master Mix (Thermo Fisher Scientific). Murine and human GAPDH were used as housekeeping genes for normalization.

For microRNAs, expression was assessed using the TaqMan™ technology. cDNA was synthesized using the TaqMan™ Advanced miRNA cDNA Synthesis Kit (Thermo Fisher Scientific). Expression was assessed using the TaqMan™ Advanced miRNA assays hsa-miR-29a-3p and hsa-miR-29b-3p for microRNA29a and microRNA29b, respectively. Expression was normalized to microRNA16 using the hsa-miR-16-5p assay (all Thermo Fisher Scientific).

2.2.12 CRISPR/Cas9 based knockout of regulatory regions

2.2.12.1 Cloning

For functional validation of intergenic regions, CRISPR/Cas9 knockout experiments were performed using the double-guide vector pX333 (Addgene plasmid #64073). Here, two sgRNAs can be expressed from two independent U6 promoters and Cas9 is expressed from the Cbh promoter. Sequences for forward and reverse oligonucleotides were designed using the CRISPOR tool (Concordet and Haeussler, 2018) with target sequences from the mouse (mm10) and human (hg38) genome. For each knockout experiment, six guides were selected (three on each site of the knockout region). Vectors of different guide combinations were pooled before electroporation. Annealing of single-stranded oligonucleotides was performed by an initial denaturing step at 95°C for 5 minutes in a PCR cycler followed by a slow cool-down back to room temperature.

Candidate region specific guides or lacZ control guides were sequentially cloned into the pX333 vector. Annealed oligonucleotides were diluted (1:50) and Golden Gate reaction was performed (Table 21/22).

Table 21: Golden Gate protocol for cloning of sgRNAs into the pX333 vector.

Component	Volume (µl)
pX333 vector (90 ng/µl)	1
Annealed and diluted oligonucleotides	1
T4 DNA ligase buffer	2
Restriction enzyme BbsI or BsaI	1
T4 ligase	1
H_2O	14

Table 22: Thermocycler program for Golden Gate cloning protocol.

Temperature (°C)	Time (min)	Cycles
37	5	10 x
16	10	
55	5	1
80	5	1
10	∞	1

Golden Gate product was used for transformation of the newly assembled plasmid into chemically competent bacteria (homemade *Stbl3*). Bacteria were thawed on ice and 5 µl of the Golden Gate product was mixed with 10 µl 5x KCM buffer (500 mM KCl, 150 mM $CaCl_2$, 250 mM $MgCl_2$) and 35 µl H_2O. 50 µl chemically competent bacteria were added and the mixture was incubated 20 minutes on ice and additionally 10 minutes at room temperature (RT). After adding 300 µl of LB medium (without antibiotics), bacteria were incubated in a horizontal shaker (800 rpm) at 33°C for 60 minutes. After centrifugation, bacteria were plated on Agar plates containing the selection antibiotic (100 µg/mL ampicillin) at 33°C over night.

Bacterial colonies were picked and incubated in LB medium for subsequent plasmid DNA isolation using the QIAprep Spin Miniprep Kit (Qiagen) according to manufacturer's instructions. Guide sequences in the vector were confirmed using Sanger sequencing (Eurofins Genomics). Plasmid DNA containing both sgRNAs (confirmed by Sanger sequencing) were used for further electroporation of cell lines.

2.2.12.1 Cell culture

The human T-ALL cell line Jurkat (ATCC® TIB-152™) and the murine T cell lymphoma cell line EL4 (ATCC® TIB-39™) were used for knockout experiments. HEK293T cells (ATCC® CRL-3216™) were used as a control. Jurkat and EL4 suspension cell lines were maintained in uncoated flasks in RPMI1640 medium. HEK293T cells were cultured in coated flasks and

dishes in DMEM medium. All cell lines were cultured in media supplemented with fetal bovine serum (FBS, 10%) and 1% penicillin/streptomycin and maintained at 37°C with 5% CO_2. Cell counting was performed using an improved Neubauer counting chamber.

2.2.12.3 Electroporation and sorting of cells

Cell lines were electroporated using the Amaxa® Cell Line Nucleofector® Kit V (Jurkat and HEK293T cells) and Kit L (EL4 cells) (Lonza). For each knockout, the pX333 vector (2 µg mixture of three double-guide vectors) and a GFP vector (0.5 µg) were co-electroporated into 2×10^6 cells (EL4) or $1x10^6$ cells (Jurkat and HEK293T) according to manufacturer's protocol. Briefly, cells were mixed with the electroporation solution (82 µl solution and 18 µl supplement) and the plasmids. Cells were nucleofected using program X-005 (Jurkat and HEK293T) or the C-009 (EL4) of the Amaxa Nucleofector™ 2b (Lonza).

The next day (24-36 h after nucleofection), GFP positive cells were single-cell sorted in 96-well plates and cultured with conditioned medium and 20% FCS. Sorting was performed at the Cell Core Facility at TranslaTUM, Klinikum rechts der Isar, Munich by Markus Utzt. Colonies grown from single cell clones were screened for the knockout using PCR with region specific primers flanking the target sequence (using the protocols for mouse genotyping).

2.2.12.4 RNA isolation and qPCR

Positive clones (knockout was confirmed by PCR amplification and Sanger sequencing of the product) were expanded for RNA isolation. Expression of the target gene was determined by real-time quantitative PCR (qPCR) using primers specific for the target transcripts. For normalization of RNA input, Gapdh qPCR was performed. Expression of the target gene was compared to cell clones electroporated with lacZ guides.

2.2.13 Mouse genotyping

Ear punches of mice were taken 21-28 days after birth. Tissue biopsies were lysed in 50 µl DirectPCR Lysis Reagent (Viagen) with 20 µg/ml Proteinase K and overnight incubated at 55°C followed by a subsequent incubation at 95°C for 15 minutes for heat inactivation of the enzymatic activity of proteinase K. Samples were 1:10 diluted with ddH_2O and long term stored at -20°C. Standard genotyping was performed using the 2x Kapa 2G Genotyping Mastermix (Sigma) according to Table 23/24.

Table 23: PCR setup for genotyping PCR

Component	Volume (µL)
Kappa 2G Mastermix	5
Primer forw	0.5
Primer rev	0.5
H_2O	2
DNA (1:10 diluted)	2

Total	**10**

Table 24: Touchdown PCR thermocycler program.

Step	Temperature (°C)	Time (s)	Cycle
Initial Denaturation	95	180	1
Denaturation	95	20	
Annealing	65	20	13 (Δ-1°C per cycle)
Extension	72	45	
Denaturation	95	20	
Annealing	55	20	26
Extension	72	45	
Final Extension	72	120	1
Hold	10	∞	1

2.2.14 Fibroblast isolation from mouse tissue

To isolate fibroblasts from mice, the tail tip was cut and thoroughly washed with 80% Ethanol. The tail tip was transferred to a cell culture dish, cut into very small pieces and incubated with collagenase type II (200 U/ml) at 37°C overnight. The next day, the sample was centrifuged at 450 x g for 5 minutes, resuspended with RPMI medium, filtered (30 µm) and plated into a well of a 6-well plate. After 2-3 days, medium was changed. Fibroblasts grew out of tissue after a few days and were further cultured. As soon as fibroblasts grew reliably, doxycycline (1 mg/ml) was added to induce expression of *Spic* in a 1:5,000 dilution.

2.2.15 Tissue preparation and staining for FACS

During necropsy of sick or control mice, blood was drawn from the heart. Analysis of mouse blood values was performed using the scil Vet ABC ™Hematology Analyzer.

Spleen and bones (tibia and femur) were harvested for further flow cytometric analysis. Bone marrow was washed out from the bones with FACS buffer (PBS + 2% FCS) using a syringe. Cell separation of bone marrow cells was performed by repeated drawing using a needle.

Separation of spleen tissue was achieved using a 30 µm filter. After filtering, samples were centrifuged for 10 minutes at 450 x g at 4°C. Erythrocytes from blood and bone marrow samples were lysed by incubating the samples with Red Blood Lysis buffer (Thermo Fisher) for 10 minutes at 4°C. Cells were counted using the Neubauer counting chamber before antibody staining. Then, 1 to 10 $x10^6$ cells were mixed with respective antibody panel mixes. Single stains of each fluorophore were used for compensation. Antibodies were incubated for 1 h. Cells were then washed with FACS buffer, centrifuged, and pellets were resuspended in 500 µl FACS buffer with 0.2 µg propidium iodide (Thermo Fisher) and filtered before flow

cytometric analysis. FACS analysis was performed on a BeckmanCoulter CyAn, equipped with 405 nm, 488 nm, and 633 nm lasers.

FACS data analysis was performed by Michele Buck.

2.2.16 Statistical analysis

All statistical analyses were performed using R v4.0.1 with the package version indicated in the material section. Methods used for statistical hypothesis testing are directly stated in the figure legends. In general, the significance level was set to 0.05. Boxplots were generated using the default ggplot2 geom_boxplot settings (middle, median; lower hinge, 25% quantile; upper hinge, 75% quantile; upper/lower whisker, largest/smallest observation less/greater than or equal to upper/lower hinge ± 1.5 * IQR).

3. Results

Parts of the results of this chapter have been submitted in the research article "In vivo interrogation of regulatory genomes reveals extensive quasi-insufficiency in cancer evolution", Fischer et al (Cell Genomics 2023). This study comprises the characterization of a PiggyBac transposon cohort with 256 mice, which were generated by Roland Rad in the laboratory of Professor Allan Bradley at the Wellcome Trust Sanger Institute in Cambridge, UK. The analysis of this transposon screen data, validation experiments, data interpretation, and the experiments including studies on the characterization of the Spic mouse model were performed by myself.

3.1 Rosa26^{PB};ATP2 mice develop a broad spectrum of hematopoietic tumors

To enable subtype-specific analysis of hematopoietic malignancies, the previously described transposon cohort (Rad et al., 2010) was expanded and a large screening cohort with a total of 256 *Rosa26^{PB};ATP2* mice was generated (Figure 9a). Mice were aged and monitored for signs of sickness and tumor development. At the time of necropsy, most mice presented with enlarged spleen, lymph nodes and/or thymus confirming the previously observed predominance of hematopoietic tumors using ATP2 mouse lines (Rad et al., 2010). Tumors were characterized using an immunohistochemistry (IHC) panel including the markers CD3 (T cells), B220 (B cells), and MPO (myeloid cells). Murine counterparts of human T cell lymphoblastic lymphoma, acute myeloid leukemia (AML) with and without maturation, acute myeloid leukemia with erythroid or megakaryoblastic differentiation, histiocytic proliferation, and a heterogeneous group of mature B cell lymphomas (BCLs) and B cell lymphoblastic leukemia (B-LBL) were diagnosed (Figure 9b/c, Table 25).

Tumors were classified dependent on their cell-of-origin (myeloid, B or T) or to the 'mixed' category comprising mice with a combination of myeloid and lymphoid malignancies.

Table 25: Histopathological diagnoses of 256 ATP mice preselected for hematologic malignancies during necropsy. Histopathological assessment was performed by Professor Leticia Quintanilla de Fend and her team at Tübingen University. Samples classified as 'others' represent samples where no diagnosis was possible due to (i) tissue lysis, (ii) the diagnosis of a benign immunological phenomenon such as spleen hyperplasia or (iii) the diagnosis of a lymphoblastic proliferation that could not be further characterized.

Origin	Diagnosis	Differentiation/maturation	Number
Myeloid	AML	With maturation	40
		Without maturation	38
		With histiocytic proliferation	6
		With megakaryocytic differentiation	25
		With erythroid/erythrocytic differentiation	5
Mixed	Myeloid-lymphoid	AML + B cell	40
	Myeloid-lymphoid	AML + T cell	8
Lymphoid	T cell	T-LBL/T-ALL and MTL	53
	B cell	BCL, B-LBL	29
	Unknown		3
Others			9
Total			256

AMLs represent the largest subgroup in the cohort followed by T cell malignancies. Of note, survival correlated with the tumor subtype. Median tumor-related survival was lowest in mice developing T cell malignancies (203 d), followed by AML (301 d), mixed (386 d), other (418 d) and B cell (498 d) mice (Figure 9d).

Taken together, the large size of the screen allowed the collection of a diverse spectrum of hematologic malignancies and provided sufficient sample numbers for detailed subgroup analysis.

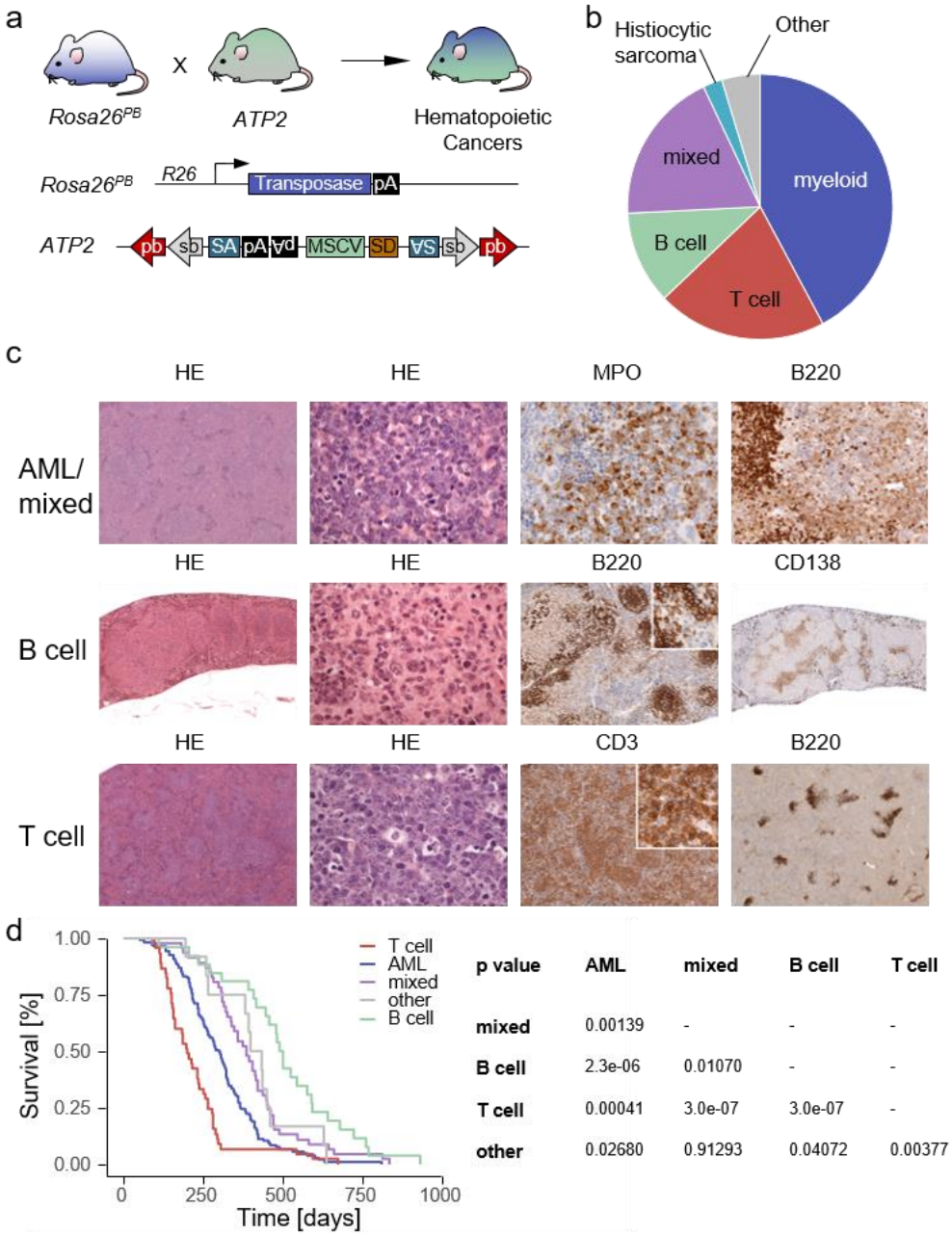

Figure 9: Histopathological characterization of the Rosa26^{PB};ATP2 screening cohort. a) *Mating scheme for PiggyBac screening using ATP2 mouse lines. In Rosa26^{PB};ATP2 double positive mice transposon mobilization is active and mice develop hematopoietic cancers.* ***b)*** *Histopathological diagnoses of hematopoietic malignancies in Rosa26^{PB};ATP2 mice. Spectrum of 256 hematopoietic tumors diagnosed in the ATP2 screening cohort. The category 'mixed' describes tumors with myeloid and lymphoid origin. Details for cases where no diagnosis was possible are given in Table 25.* ***c)*** *Representative images of tumors from all major subgroups. The first row shows an acute leukemia (AML) case with MPO expression. Additionally, the shown AML case expresses the B cell marker B220. The second row represents a diffuse large B cell lymphoma (DLBCL) characterized by B220 expression. The additional expression of CD138 led to the classification of DLBCL with plasmacytoid differentiation. The third row shows a T cell (lymphoblastic) lymphoma characterized by strong CD3 expression and the absence of B220 expression. Magnifications: first column: 25x; second: 630x; third: MPO 400x, B220 50x (insert 400x), CD3 25x (insert 400x); fourth: B220 200x, CD138 25x, B220 25x.* ***d)*** *Survival of Rosa26^{PB};ATP2 mice with different types of hematopoietic tumors (AML n = 107, mixed n = 46, B cell n = 26, T cell n = 45, other n = 12). Mice in the other group comprise 'other' and 'unknown' categories from Table 25. Mice with NA values were excluded from the analysis. All pairwise comparisons using the Log-Rank test are shown. Benjamini-Hochberg (FDR) was used to adjust for multiple testing.*

3.1.1 Mouse T cell malignancies recapitulate human T-ALL subtypes

ATP2 induced T cell malignancies (CD3 positive cases, 53/256, 21%) were selected for further in detail analysis. Based on staining with the terminal deoxynucleotidyl transferase Tdt (marker for lymphoid precursors) the majority of these cases was diagnosed as T cell lymphoblastic lymphoma (T-LBL)/T cell acute lymphoblastic leukemia (T-ALL) (Figure 10a). As according to the WHO T-LBL and T-ALL constitute the same disease entity and only differ in their clinical manifestation and the degree of bone marrow infiltration (Morse et al., 2002), hereafter Tdt positive T cell malignancies are referred to as T-ALL cases. Only two tumors were CD3 positive but lacked Tdt expression and were therefore classified as mature T cell lymphomas (MTL) (Figure 10a). To further subclassify the T-ALL cases based on their cell of origin, IHC of the mature T cell marker CD4 was performed. Expression of CD4 is associated with later stages of T cell development (compare Figure 6 introduction) and was only found expressed in a subgroup of samples (described in more detail in chapter 3.5).

Additionally, array comparative genomic hybridization (CGH) was performed to further characterize murine T-ALLs at a genomic level. DNA was isolated from all T-ALL tumors for which tissue and matching tail samples were stored in RNAlater. This analysis revealed recurrent deletions on chromosome 6qB1 (35/43 samples; 81%) and chromosome 14qC2 (41/43 samples; 95%) (Figure 10b). These deletions are caused by T cell receptor (TCR) rearrangement and are also found in 96% of human T-ALL cases (biallelic 91%) (Szczepański et al., 2000), demonstrating the validity of the model. Besides deletions in the TCR gene loci, recurrent genomic alterations were rare. The only recurrently altered region was an amplification on chromosome 10 (18/43 samples; 42%) (Figure 10c). A subgroup of mice with chr10 alterations showed whole-chromosome amplifications of chromosome 10 (5/18; 28%), whereas the majority was characterized by specific amplifications of 10qC (13/18; 72%). The 2 Mb large minimal overlap region contains 82 protein-coding genes, including *Tcf3*, a well-studied transcription factor in T cell development.

Thus, the screen revealed a cohort of T-ALLs, which recapitulate human T-ALL based on immunohistochemical and genomic analysis.

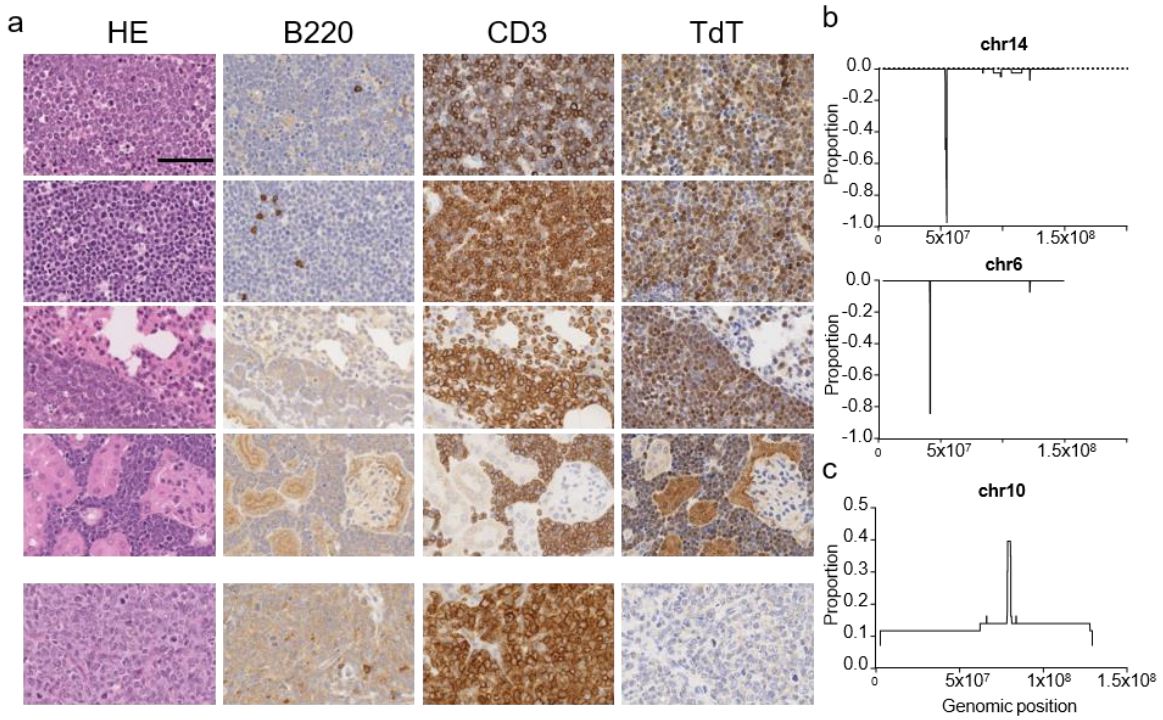

Figure 10: Histopathological and genomic characterization of T cell malignancies from Rosa26^{PB};ATP2 mice. a) *Microscopic immunohistochemical images of representative cases. The upper 4 cases were classified as T-ALL/T-LBL and the case presented in the last row was classified as mature T cell lymphoma (MTL). T-ALL/T-LBL cases were characterized by medium sized cells, round to ovoid nuclei and a scant cytoplasm. The nuclei were characterized by fine or dispersed (salt and pepper) chromatin and usually one central nucleolus. MTLs were negative for Tdt and showed either small cell size and inconspicuous nuclei or medium cell size with irregular nuclei. Tumors are negative for the B cell marker B220 and show strong positivity for the T cell marker CD3. CD3 expression was membrane associated, but was also found in the cytoplasm. The early T cell marker Tdt was found in the cytoplasm and was used to differentiate between T-ALL/T-LBL and T cell lymphomas developing from mature T cells. Infiltration in lung (third panel) and kidney (fourth panel) was observed for T-ALL cases. Tumors were classified according to the Bethesda proposals for classification of lymphoid neoplasms in mice (Morse et al., 2002) by PD Katja Steiger. Scale bar, 50 µm.* ***b)*** *Array CGH data from 43 mice showing TCR rearrangement deletions on chromosome 6 and 14 at the locus of the TCR beta and alpha chain, respectively. The proportion of samples with the deletion is shown on the y axis. Copy number changes were normalized to CGH profiles from the matching tail sample.* ***c)*** *Array CGH data showing an amplified region on chromosome 10. The minimal overlap region includes 2 Mb (chr10:79,117,736-81,076,707). The proportion of samples with the alteration is shown on the y axis. a, adapted from Fischer et al.*

3.2 Screen analysis reveals candidate T-ALL driver genes

Next, insertion landscapes were analyzed in the T-ALL cohort. To determine the exact location and the associated read abundance of each insertion, semi-quantitative insertion site sequencing (QiSeq) (Friedrich et al., 2017) was performed (Figure 11a). QiSeq of all T cell tumors (n = 51) revealed 170,075 non-redundant transposon integration sites. Genome-wide insertion profiles from all tumors showed that genes harboring the highest insertion peaks are known T-ALL drivers including *Ikzf1, Notch1, Pten* or *Myb* confirming the power of the screen (Figure 11b).

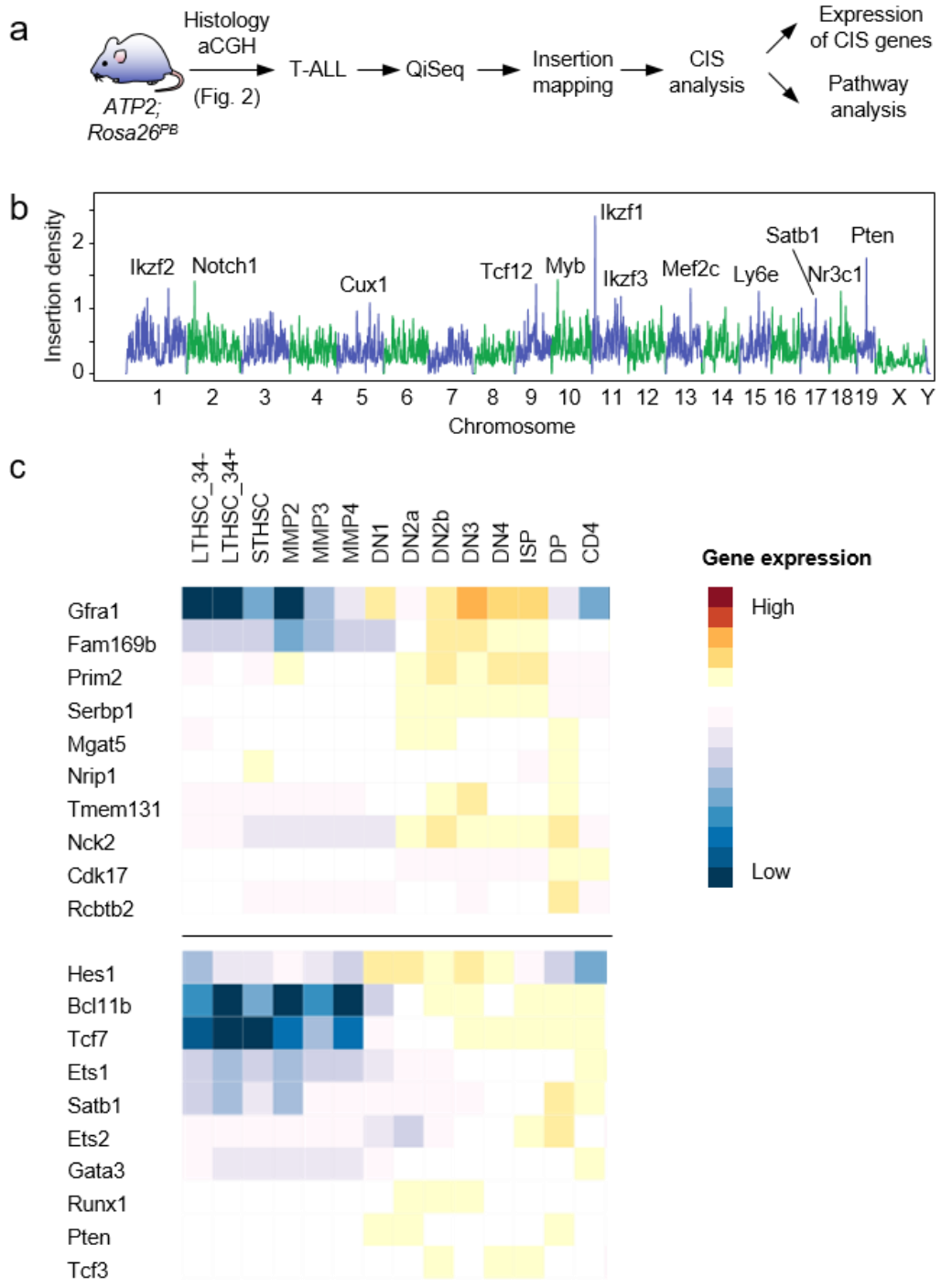

***Figure 11: Insertion data analysis identifies known and novel T-ALL genes**. **a)** Scheme showing workflow for transposon screen analysis. DNA was isolated from all samples diagnosed as T-ALL based on histology and aCGH and subsequently QiSeq was performed. After insertion mapping, insertions are submitted to statistical analysis to identify common insertion sites (CIS). Next, the expression and pathway contribution of identified CIS genes is assessed. **b)** Insertion density of 179,075 insertions from 51 samples. Highest peaks are labeled with the annotated gene in this region. The donor locus on chromosome 17 was excluded. **c)** Expression analysis of putative T-ALL candidate genes (upper panel) in T cell development from murine hematopoietic stem cells (HSCs) to mature CD4 positive T cells. Lower panel shows well-described T cell genes as a comparison. Mean-normalized heatmaps were created using the immunological genome project website (http://rstats immgen.org/MyGeneSet_New/index.html) b, c, adapted from Fischer et al.*

To identify genomic regions affected by transposon insertions more significantly than expected by chance, statistical analyses based on Gaussian Kernel Convolution (GKC) was performed (de Ridder et al., 2006). The CIMPL algorithm identified 1,062 common insertion sites (CIS). Of these CISs, 994 were found in at least 10% of samples. The top 50 CIS genes are listed in Table 26 according to their predicted or suggested molecular function and putative role in the development of hematopoietic malignancies. This list includes well-described T-ALL genes (*Bcl11b, Rasgrp1, Ezh2, Stat3/5b*) additionally to the ones mentioned above (Figure 11b). In contrast, several genes that have not been linked to T-ALL development so far, but are known to trigger leukemogenesis of other lineages were identified. Especially, an accumulation of genes linked to AML (including *Cux1, Mecom, Crebbp*) was observed. This myeloid enrichment will be discussed in more detail in chapter 3.6.4. Notably, of the top 50 CIS genes 18 have not yet been linked to hematopoietic malignancies. The majority of these genes are poorly studied, however, some have been proposed to be involved in general signaling (*Sh3kbp1, Sipa1l1*) or immune function (*Slamf6, Ly6e, Mgat5*). Moreover, some of these genes were found to be strongly regulated during T cell development (*Gfra1, Nck2, Prim2, Serbp1, Fam169b*) (Figure 11c) comparable to genes with well-known roles in T cell leukemogenesis. This indicates a potential function of these novel genes in T cell maturation.

Table 26: Top 50 CIS genes classified according to their (predicted) molecular function and relevance in leukemogenesis. The categories for the molecular function were adapted from Liu et al. Molecular function was assigned using literature research.

Molecular function	Known T-ALL gene	Novel in T-ALL	
		Associated with non-T hematopoietic malignancies	Novel in other hematopoietic malignancies
Transcriptional regulation	*Ikzf1, Myb, Tcf12, Mef2c, Ets1, Bcl11b, Satb1*	*Ikzf2, Cux1, Foxp1, Ikzf3, Bach2, Ncoa2*	
Cell cycle			*Cdk17*
Epigenomic	*Kmt2a, Ezh2*	*Chd2, Mecom, Hmgb2, Crebbp*	
Signaling	*Notch1, Pten, Stat3/Stat5b, Rasgrp1, Nf1*	*Pip4k2, Akap13*	*Sh3kbp1, Gfra1, Rapgef2, Sipa1l1, Nck2, Srgap2*
Ribosomal	*Rpl5*		
RNA processing		*Mbnl1*	*Serbp1*
Immune system	*Tnfsf11, Cxcr4*	*Icos, Cd74/Camk2a*	*Slamf6, Ly6e*
Other/unknown function	*Nr3c1, Lncpint*	*Nrip1*	*Prim2, Rcbtb2, Fam169b/Igf1r, Tmem131, Mgat5*

To globally inspect the function of the identified genes, pathway enrichment analysis was performed using the top 50 CIS genes as an input. The analysis confirmed their involvement in cancer, more specifically in immune system cancer and T-ALL (Table 27). Moreover, the importance of the JAK/STAT and the estrogen signaling pathway was highlighted. Of note, the top 50 gene list shows an accumulation of genes involved in gene regulation as indicated by

the ontology terms ‘regulation of transcription by RNA polymerase II’, ‘Cis-regulatory region sequence-specific DNA binding’ and ‘nucleus’ as the cellular compartment (Table 27; the accumulation of transcriptional regulators among CIS genes is discussed in chapter 3.7).

Thus, the screen identified well-known and novel genes in T cell leukemogenesis and provides lists of putative candidate genes for further experimental validation.

Table 27: Pathway enrichment analysis using the top50 CIS genes. Analysis was performed using Enrichr and top hits and the corresponding adjusted p values of different pathways and ontology databases is listed.

Database	Name	Adjusted p value
Pathways		
KEGG 2021	Pathways in cancer	0.00001424
Elsevier Pathway Collection	Proteins involved in T-ALL	6.289e-7
Panther 2016	JAK/STAT signaling pathway	0.005939
MSigDB Hallmark 2020	Estrogen Response Early	0.0004176
Ontologies		
Jensen DISEASE	Immune System cancer	0.00001143
GO Biological Process 2021	Regulation of transcription by RNA polymerase II	6.820e-10
GO Molecular Function 2021	Cis-regulatory region sequence-specific DNA binding	5.096e-8
GO Cellular Component 2021	Nucleus	0.00001545

3.3 Using *PiggyBac* for systematic interrogation of screening the non-protein-coding genome

3.3.1 The *PiggyBac* screening system is suitable to interrogate the regulatory genome.

To examine the suitability of the *PiggyBac* screening system to interrogate the nPC genome, general characteristics of *PiggyBac* integration bias were assessed. *PiggyBac* recurrently targets active genes (de Jong et al., 2014; Yoshida et al., 2017). However, the genomic features and major determinants of *PiggyBac* integration were so far not examined using an *in vivo* dataset and until now, analyses were focused on the protein-coding genome.

The initial analysis of the global distribution of insertions in the T-ALL cohort found that nearly half of all insertions (48.2%) are located in intergenic regions defined as overlapping neither with exons nor with introns of protein-coding genes (Figure 13a). Next, the sequencing read coverages derived from protein-coding and non-protein coding insertions were compared. Of note, no significant difference was detected suggesting a comparable functional relevance (Wilcoxon test, p = 0.45, Figure 12b). Subsequently, *PiggyBac* insertion profiles were compared to epigenomic features to investigate the integration preference. First, the

epigenomic features of protein-coding genes overlapping with a common insertion site (gCIS) were compared to PC genes without PB insertions. Here, a substantial accumulation of active chromatin marks and depletion of repressive marks at CIS-overlapping genes was detected (Figure 12c). Higher levels of H3K27ac, H3K4me1, H3K4me3 and Pol2R ChIP-Seq signals were detected at the transcriptional start site (TSS) of CIS-overlapping genes, while H3K36me3 was enriched throughout the gene body of CIS-overlapping genes. In contrast, the repressive mark H3K27me3 was lower at CIS-overlapping genes compared to non-overlapping genes (Figure 12c).

Additionally, the gene expression level of genes overlapping a CIS and not-overlapping a CIS also varied substantially. Genes overlapping a CIS and therefore more frequently hit by the transposon, were generally expressed at higher levels (Figure 12d). This even became more drastic when looking at the top100 genes of the screen.

To explain the correlation of insertions in genes expressed at a high level, the distance to the next super-enhancer (SE) was assessed assuming that highly expressed genes are regulated by an active SE. Murine thymus-specific SEs were obtained from dbSUPER and overlapped with CIS genes. Genes targeted by the *PiggyBac* system more often had a SE in close proximity (defined as either overlapping or harboring a SE within 5 kb upstream of the TSS; Figure 12e). Whilst only 3.8% of genes non-overlapping CISs had a SE in close proximity, genes overlapping a CIS harbored a proximal SE in 27.5%. Genes ranked in the top100 of the screen even had a SE in close proximity in 46% (Figure 12e, outer ring). Thus, genes targeted by PB are transcriptionally active and proximal to super-enhancers.

Finally, all CIS regions – independent of their genomic location and their overlap with protein-coding genes - were overlapped with H3K27ac enhancer histone marks in healthy (double positive T cells) and malignant T cells (EL4 cell line). This revealed a general enrichment of active chromatin in CISs and showed that also non-coding insertions are close to enhancer marks (Figure 12f).

Thus, beyond its preference for transcribed genes, *PiggyBac* has a general propensity for active chromatin, supporting its application to perturb cancer-relevant regulatory elements in the nPC genome.

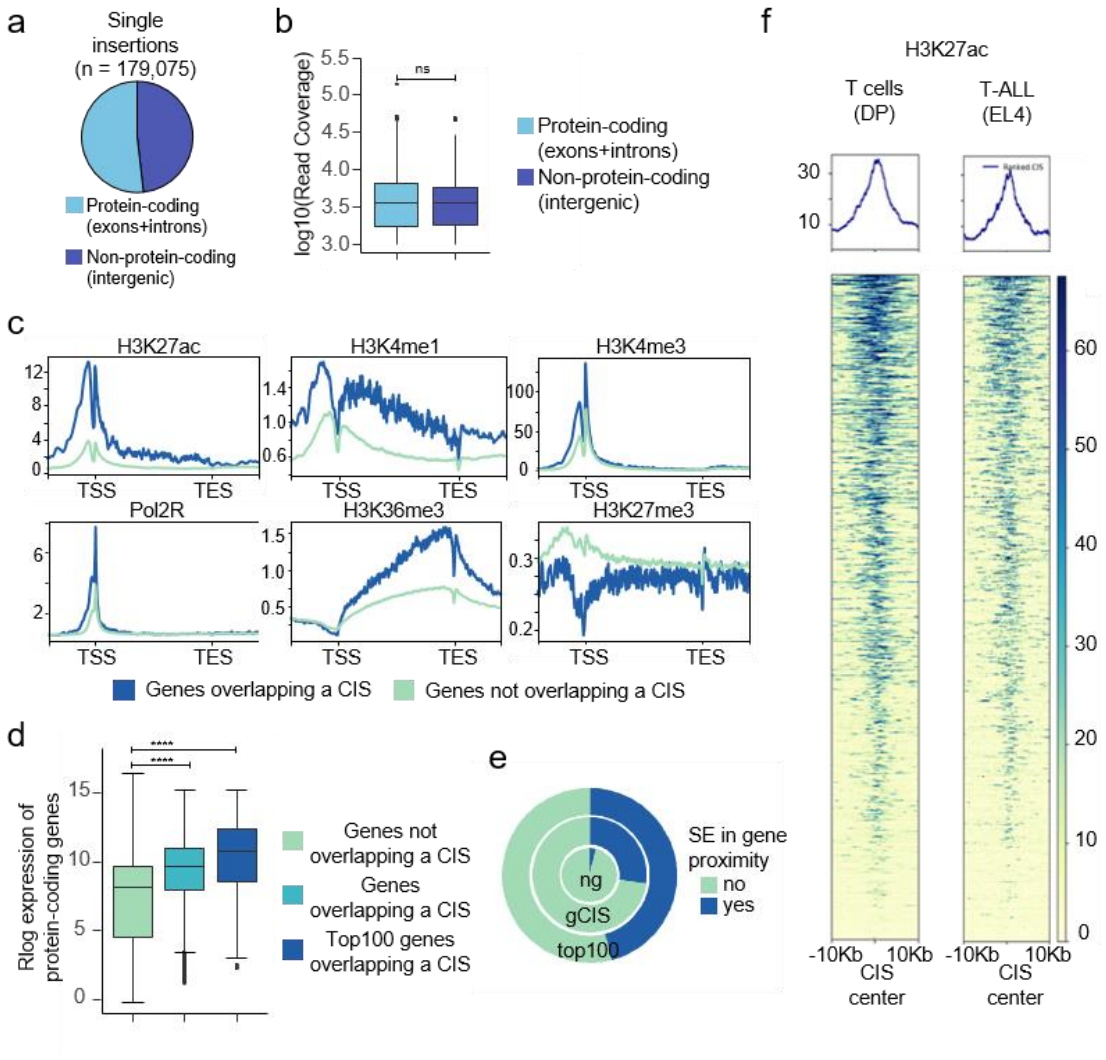

Figure 12: PiggyBac's suitability to screen for non-coding regulatory elements. a) *Proportion of protein-coding insertions (n = 92,758; 52.8%) and non-protein-coding insertions (n = 86,317; 48.2%). Protein-coding is defined as overlapping with exons or introns of PC genes.* ***b)*** *Read coverage similarity between protein-coding and non-protein-coding insertions. High-coverage insertions were selected (> 1,000 reads). P value = 0.445, Wilcoxon test.* ***c)*** *Epigenomic features compared between genes overlapping a CIS (n = 859) with genes not overlapping a CIS (n = 12,465). Murine thymus ChIP-Seq datasets were downloaded from ENCODE. A region of 2 kb around the transcriptional start site (TSS) and transcriptional end site (TES) is shown.* ***d)*** *Genes targeted by PiggyBac (genes overlapping a CIS) are expressed at higher levels in T cell development. RNA-Seq data from DN2 T cells was used (Hu et al. 2018). Rlog expression of (the top100) protein-coding genes overlapping or not overlapping a CIS are shown (****P < 0.0001, ***P < 0.001, Wilcoxon test). 'gCIS' defines a protein-coding gene overlapping a CISs. 'ng' defines genes non-overlapping a CISs.* ***e)*** *Proportion of genes with a SE in close proximity. Only 3.8% of genes not overlapping a CIS have a proximal SE (inner circle). Genes and top100 ranked genes overlapping a CIS have a proximal SE in 27.5% and 46% (middle and outer circle), respectively.* ***f)*** *Profile heatmap plot showing enrichment of H3K27ac ChIP-Seq data from healthy double positive T cells and the T cell lymphoblastic lymphoma cell line EL4 in CIS regions (n = 1062). A region of 10 kb around the CIS center is shown. a, b, c, f adapted from Fischer et al.*

3.3.2 Development of a computational pipeline to annotate non-coding CISs

Statistic algorithms used so far to identify cancer genes based on insertional mutagenesis data, focused on protein-coding genes. Some approaches are even gene or region-centric and do not allow the identification of non-annotated intergenic regions (Brett et al., 2011). As these algorithms rely on genomic regions, the size differences of genomic elements might be responsible for these constricted analyses. The average size of protein-coding genes is around 8 kb, while regulatory elements are characterized by a maximal length of 1.5 kb. For CIMPL analysis, usually scale parameters ranging between 30 and 240 k were used. As the relevant scale for regulatory elements might be smaller, different CISs scale parameters (window sizes) were compared. A scale parameter of only 5 k substantially increased the number of regulatory CISs that can be identified using CIMPL (Figure 13a/b). However, the CIMPL output only includes the information that a region is located 'intergenic' without any further details or annotation.

For an improved understanding of intergenic CISs, the annotation of overlapping regulatory regions is a central aspect. To accelerate the future analyses of the regulatory potential of a CIS (as commonly used tools to identify CISs lack this ability), a novel computational framework named ARCIS (**A**nnotation pipeline for **R**egulatory **C**ommon **I**nsertion **S**ites) was developed (Figure 13c). ARCIS integrates CIS regions with any type of (epi)genetic information. As input files CIS coordinates (from CIMPL) as well as publicly available epigenetic data characterizing different developmental stages of the T cell lineage or T-ALL were used. Chromatin accessibility, histone modifications, annotated super-enhancers and information on 3D organization was included (technical details on ARCIS performance and datasets can be found in the Methods chapter 2.2.7). Moreover, a Hidden Markov Model on a collection of thymus ChIP-Seq data (chromHMM) was run to define chromatin states, which were additionally used as input information in ARCIS.

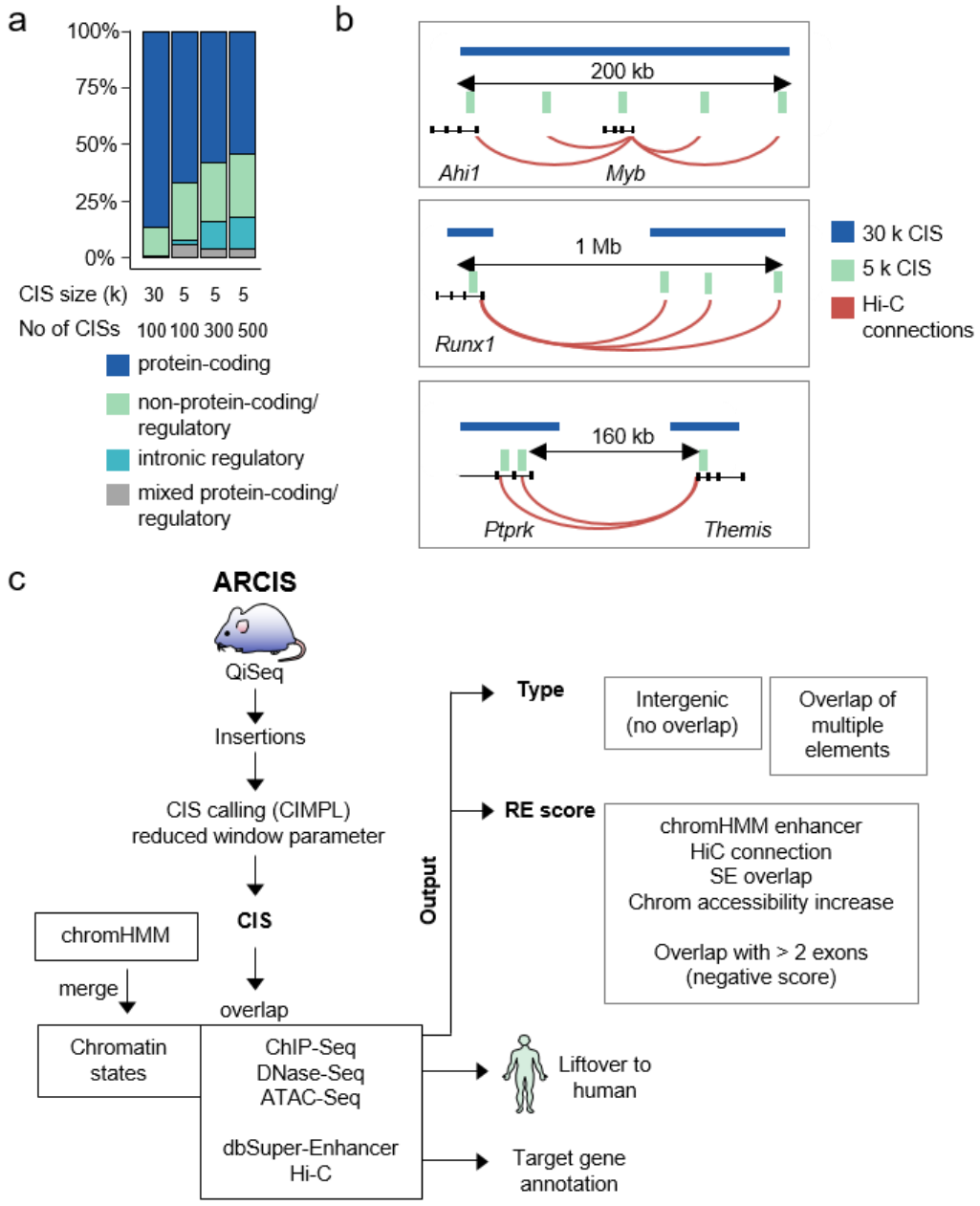

Figure 13: A computational tool to annotate regulatory common insertion sites. a) *Proportion of protein-coding, regulatory (non-coding) and intronic regulatory CISs dependent on the analysis method. Decreasing the CIMPL scale parameter from 30 k to 5 k increases the percentage of regulatory CISs (first vs second bar). The top100, top300 and top500 CIS genes were analyzed. When using the 5 k scale parameter, more (intronic) regulatory CISs can be identified when analyzing more CISs. For some CIS, classification was ambiguous and the category 'mixed' was assigned.* ***b)*** *Effect on CIS size when changing the scale parameter in the CIMPL analysis. The result of the reduction of the scale parameter from 30 k to 5k is shown for three genomic loci. 30 k CISs are depicted in blue, 5 k CISs in green. HiC connections are schematically shown as red arcs. Approximate distance between putative regulatory element and the promoter of the target gene is indicated.* ***c)*** *Scheme of the computational ARCIS tool. CIMPL-called CISs are overlapped with listed epigenomic datasets and chromatin states from a separately run chromHMM analysis. The output includes the annotation of the target gene (based on HiC connections), the putative type of the regulatory element (multiple in case of overlaps) and a RE-score (based on listed characteristics). Additionally, putative regulatory elements can be liftovered to the human genome for cross-species comparison. Adapted from Fischer et al.*

ARCIS computes for each CIS, the overlap with all input datasets and reports numbers of peaks/intersections with the dataset or the distance to the closest element/peak. Additionally,

ARCIS reports the connected target genes and information on chromatin access change during T cell development. To rank the CISs according to the probability of having regulatory potential, ARCIS calculates a 'RE-score' based on a combination of features taking into account if the region overlaps with chromHMM-predicted enhancers or dbSUPER SEs, if the region gains chromatin accessibility during T cell development and if there is a HiC connection starting from this region. Additionally, the score is penalized if the region is overlapping with more than two exons as this indicates that rather the gene is the actual transposon hit. The ARCIS output includes the type of genomic element found in the CIS region (PC gene, nPC transcript or intergenic region) and the score indicating a putative regulatory activity. However, the ARCIS output file is not only useful for fast sorting and ranking of potential regulatory CIS using the score, but also allows the search for a target gene of interest which is not directly hit by the transposon itself but regulated by identified REs.

Thus, this novel computational tool offers the possibility to pre-select CIS with potential regulatory activity for further in detail analysis.

3.3.3 Individual inspection and verification of identified regulatory elements

Although the preselection of CISs with regulatory regions represents a crucial advance, the final classification of regulatory CISs remains challenging. As functional genomic elements often overlap or are found in close proximity to each other, differentiation between these elements required individual inspection (Figure 14).

To understand and judge the result of the ARCIS pipeline in detail, all CISs above a specific threshold (insertions in ≥ 7 tumors; 537 CISs) were evaluated manually/individually following the rules described in the Methods chapter 2.2.7.2 (Figure 8). Briefly, main criteria for discrimination were the position of insertions within the CIS region and their overlap with functional genomic elements, their orientation and their insertion pattern across samples. Additionally, epigenetic marks and HiC connections were considered to evaluate a putative RE.

In case classification is still ambiguous, the integration of lncRNA and mRNA expression levels from developing T cells can further aid the discrimination. Examples include the discrimination between intronic and PC CISs using mRNA expression data (if no expression is detected during the course of T cell development, the intronic element is most likely the hit of the transposon) and the discrimination between lncRNAs and enhancers (if the lncRNA is not expressed, the intergenic enhancer is most likely the hit of the transposon). However, in some cases no definitive classification is possible. Here, downstream experimental interrogation is necessary. Figure 14 lists the main RE categories and target genes or transcripts of the highest scoring REs.

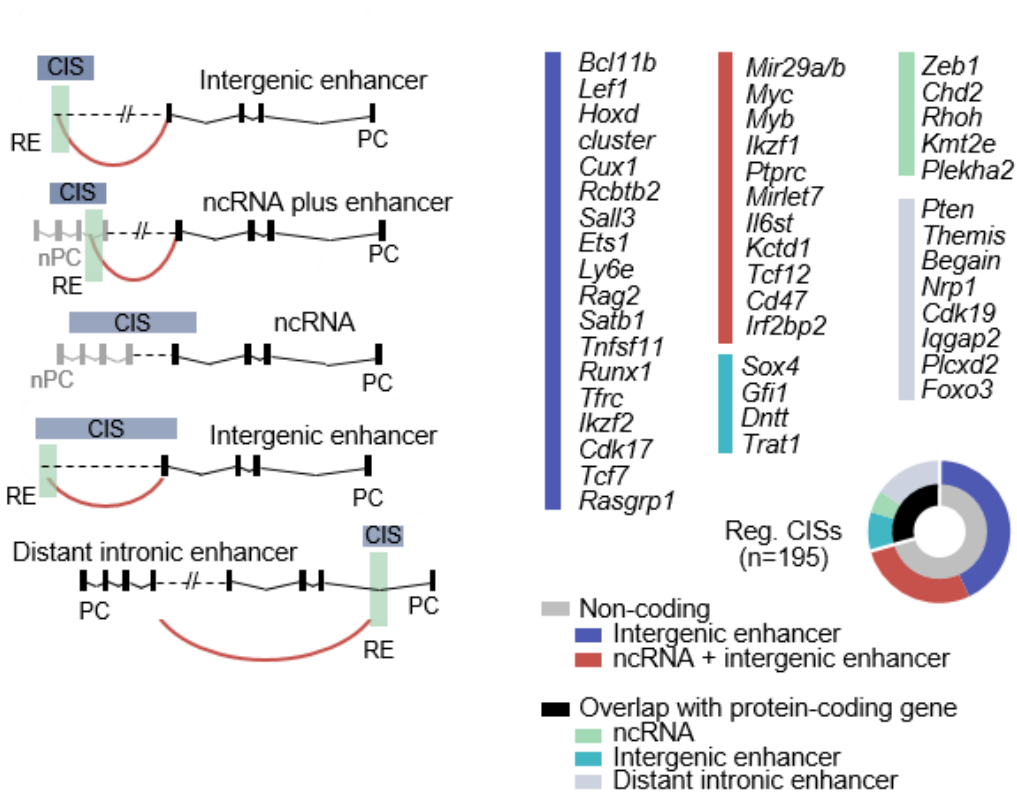

Figure 14: Resulting categories of regulatory CISs using ARCIS and individual inspection. *The ARCIS output including the type and RE-score of each potential RE further needs manual inspection to classify CISs to one of the five listed categories. During manual inspection the location and orientation of insertions in relation to protein-coding genes and non-coding transcripts is assessed. CISs (n = 537) were individually inspected and classified to one of the listed categories. Additionally, the target gene or transcript is identified based on 3D data (HiC, depicted as red arcs). Selected examples of each category with a high RE-score are listed on the right. Adapted from Fischer et al.*

Thus, these analyses showed the suitability of the PB system to screen for regulatory cancer drivers and generated large catalogues of cancer-relevant REs in T-ALL.

3.3.4 Human relevance of intergenic CISs

Before experimental targeting of selected regions, human relevance of the target genes was evaluated by comparing identified regions to cancer-risk variants in publicly available genome-wide association studies (GWAS). First, to assess if identified REs and their putative target genes are relevant in human leukemogenesis, genetic variants associated with cancer in GWAS were analyzed. Intergenic CISs most likely lead to a more subtle gene dysregulation than insertions in protein-coding genes. Equally, GWAS variants often affect the non-coding sequence, indicating a similar subtle gene regulatory effect. Genes associated with human GWAS variants were intersected with the target genes of the regulatory CIS lists. Indeed, regulatory CIS target genes (n=149, a selection is listed in Figure 14) were significantly enriched for GWAS associated cancer variants in humans (P=0.0001 pan-cancer variants; p = 3*10-5 hematopoietic cancer variants; Fisher's exact test). In contrast to the analysis using the reported gene of the GWAS variants, all CISs (n = 1056) were also lift-overed to the human

genome and overlapped with unique cancer-associated GWAS variants (n = 6,221). A significant enrichment (Chi squared test, p = 0.0001) was detected when comparing the number of CISs found in all CIS regions (3,7 variants/Gb) to the complete genome (2 variants/Gb).

Second, identified regulatory regions were cross-compared to regulatory activity in T-ALL patients. GRO-Seq data showing nascent RNA transcription including enhancer RNA (eRNA) of two T-ALL patients and the T-ALL cell line Jurkat was used for cross-species analysis. Identified CIS regions (n = 50, ranked according to ARCIS score) were manually inspected and size-reduced to only comprise the regulatory regions instead of the complete CIS region (reduction from ~16 kb to ~4 kb). Liftover to the human genome was performed for these smaller regions, which were subsequently analyzed for GRO-Seq expression peaks. In 41/50 regions (82%) clear regulatory activity was detected in human T-ALL.

These results demonstrate the human relevance of identified regulatory regions and their importance in human cancer evolution.

3.4 Functional characterization of identified REs

3.4.1 An intronic RE of the tumor-suppressor *Pten*

The integration of 3D data to the transposon analysis workflow opened the new possibility to identify regulatory elements within introns and assign their putative distant target gene. So far, CIS algorithms assigned CISs to a gene based on the location ignoring the possibility of intronic REs with long-range effects. The combination of ARCIS and manual inspection led to the identification of 30 CISs likely to harbor an intronic RE. The most relevant characteristics of intronic regulatory CIS include their clustered insertion peak with unbiased orientation (to exclude activation peaks) and the often negligible gene expression of the CIS gene (where the insertion is located but not the target gene) in murine T cell development. Several genes, which are potentially regulated by a cancer-relevant intronic RE were identified (Figure 14). Examples include the well-known tumor-suppressor *Pten*, as well as genes not described in T-ALL so far, but either known in T cell biology (*Themis*, *Nrp1)* or generally associated with tumor suppression *(Txn1*, *Iqgap2)*.

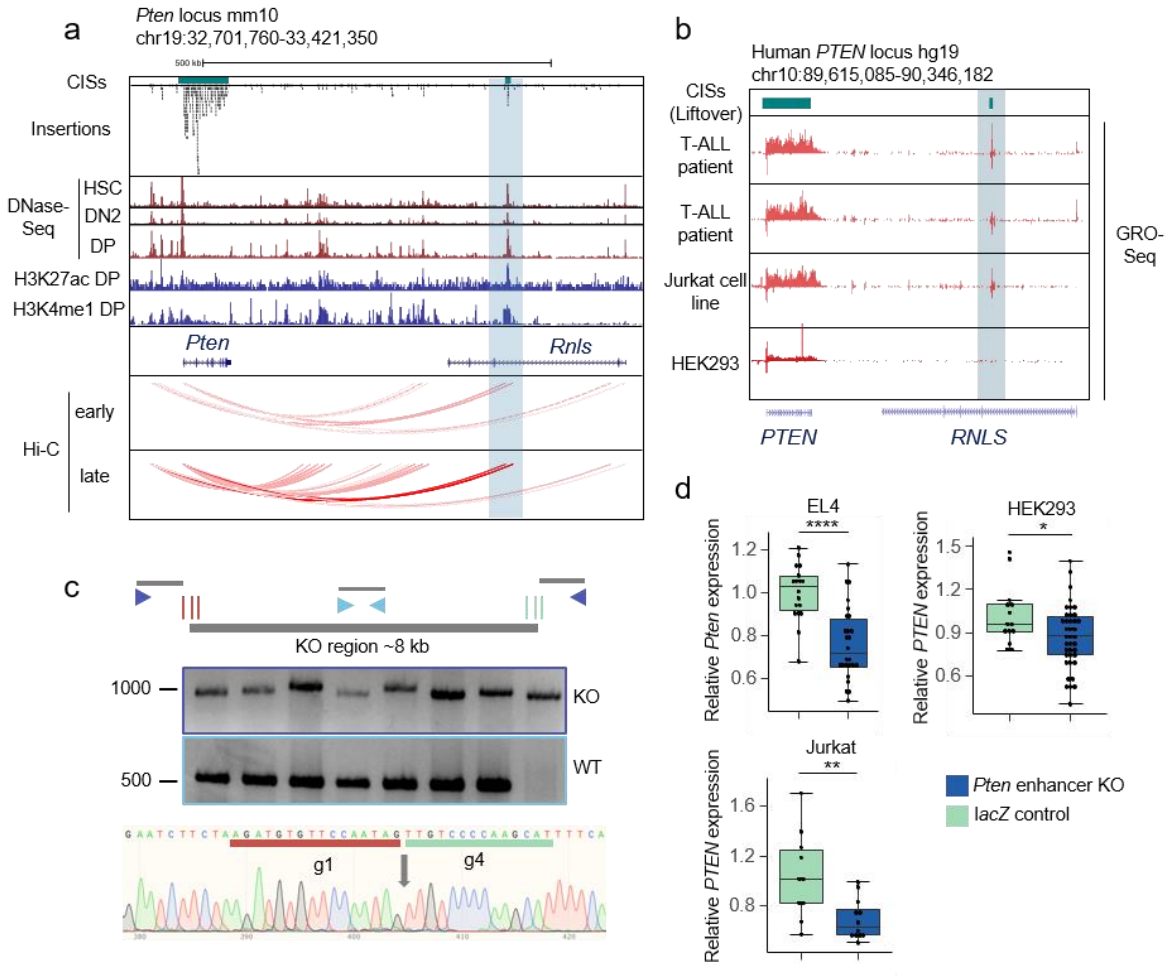

Figure 15: Targeting an regulatory element of Pten located in an intron of Rnls. a) *The murine Pten locus showing CISs (green boxes) and insertions (black) from the transposon screen together with publicly available epigenomic datasets including DNase-Seq, ChIP-Seq and Hi-C data. DNase-Seq data is shown for hematopoietic stem cells (HSC), double negative stage 2 (DN2) and double positive (DP) T cells. ChIP-Seq datasets are shown for DP cells. Hi-C data is separated into data from early (HSC-DN2) and late T cell development (DN3-DP). The putative RE is located in an intron of Rnls and highlighted in blue.* ***b)*** *The human PTEN locus showing CIS regions from the transposon screen liftovered to the human genome and four tracks of GRO-Seq data from two T-ALL patients, the T-ALL cell line Jurkat and HEK293 cells. The region of the syntenic regulatory element is highlighted in blue.* ***c)*** *Scheme showing the human knockout (KO) region targeting the putative PTEN enhancer element. Three CRISPR/Cas9 guides flanking the KO region were designed for each site (guides depicted in red and green). To confirm the knockout, PCR was performed with KO primers flanking the KO region (dark blue) and wild type primers within the KO region (light blue). The result of the PCR is exemplary shown for 8 positive clones in HEK293 cells. A knockout band (~900 bp) can be detected in all samples, the wild type band (~500 bp) is missing in the last clone suggesting a homozygous KO. The Sanger sequencing trace is shown for one clone and sequences of the guide1 and guide4 are indicated.* ***d)*** *Knockout of the putative RE using CRISPR/Cas9 in different cell lines. Knockout was performed in the T cell lymphoblastic EL4 cell line (KO n = 26, ctrl n = 18), the human T-ALL cell line Jurkat (KO n = 12, ctrl n = 10) and human HEK293 cells (KO n = 38, ctrl n= 14). A 7-8 kb intronic region was knocked out and targeted cells were single cell sorted based on GFP expression. Expression of the target gene was assessed using qRT-PCR. Each dot represents expression in a single cell-derived clone. Pten expression was normalized to Gapdh expression. *P < 0.05, **P < 0.01, ****P < 0.0001, Wilcoxon test. Adapted from Fischer et al.*

To functionally characterize these intronic elements, a potential intronic RE in the *Rnls* gene was selected. This RE showed a Hi-C connection to the ~400 kb distant *Pten* promoter (Figure 15a) and was recently described (Tottone et al., 2021). To confirm one of the main

characteristics, *Rnls* expression levels were checked first. Indeed, *Rnls* was not expressed during T cell development confirming the hypothesis that this CIS is not influencing *Rnls* but rather regulating *Pten* expression (data not shown). Next, cross-species functionality was assessed by analyzing GRO-Seq data of two T-ALL patients. A double peak of short-lived enhancer RNA, a characteristic of enhancer activity, was detected in the syntenic human region. Of note, there was additionally evidence for a cell-type specificity as GRO-Seq peaks were not present in HEK293T cells (Figure 15b). To finally prove functional relevance, the 7-8 kb intronic region in *Rnls* harboring the putative RE was knocked out using CRISPR/Cas9 (Figure 15c). A significant decrease of PTEN expression was observed in human and murine T-cells, while in HEK293 cells the knockout resulted in only a slight decrease of PTEN levels (Figure 15d).

3.4.2 T-ALL-relevant non-coding RNAs identified by the screen

Next, the expression level of the 54 nPC transcripts identified by ARCIS was assessed. Therefore, publicly available RNA-Seq data from T cells were used (Hu et al., 2018). Expression was detectable for more than 70% of this non-coding transcripts in developing T cells (Figure 16a). Many of these transcripts were found in close proximity to well-known T-ALL genes, including *Myb*, *Myc*, and Ptprc (compare Figure 14). However, also transcripts in the neighborhood of genes so far not linked to T-ALL were identified, including *Fam126a, Il6st* and *Kctd1*.

First, an antisense RNA at the *Zeb1* locus was selected for further in detail analysis. This CIS was classified as “PC transcript plus ncRNA” (compare Figure 14). The manual inspection of the locus revealed two insertion peaks. One insertion peak was predicted to activate *Zeb1* expression, while the insertions in the other peak showed an opposite orientation more likely to activate the *Zeb1* antisense transcript *Gm10125* (Figure 16b). In the human system, *ZEB1-AS* RNA was described to activate *ZEB1* expression by recruiting the H3K4 methyltransferases to its promoter (Su et al., 2017). To examine whether the mouse antisense transcript also has a similar functional role, two *Zeb1-AS* exons were knocked out in EL4 cells using CRISPR/Cas9. Importantly, an appropriate distance of 8 kb to the *Zeb1* promoter was considered in order to avoid direct interference with *Zeb1* transcription. *Zeb1* expression level were compared between knockout and control cells. Cells harboring the deletion in the antisense transcript showed decreased Zeb1 levels indicating that the murine transcript (*Gm10125*) has a similar function as *ZEB-AS1* in the human system (Figure 16c).

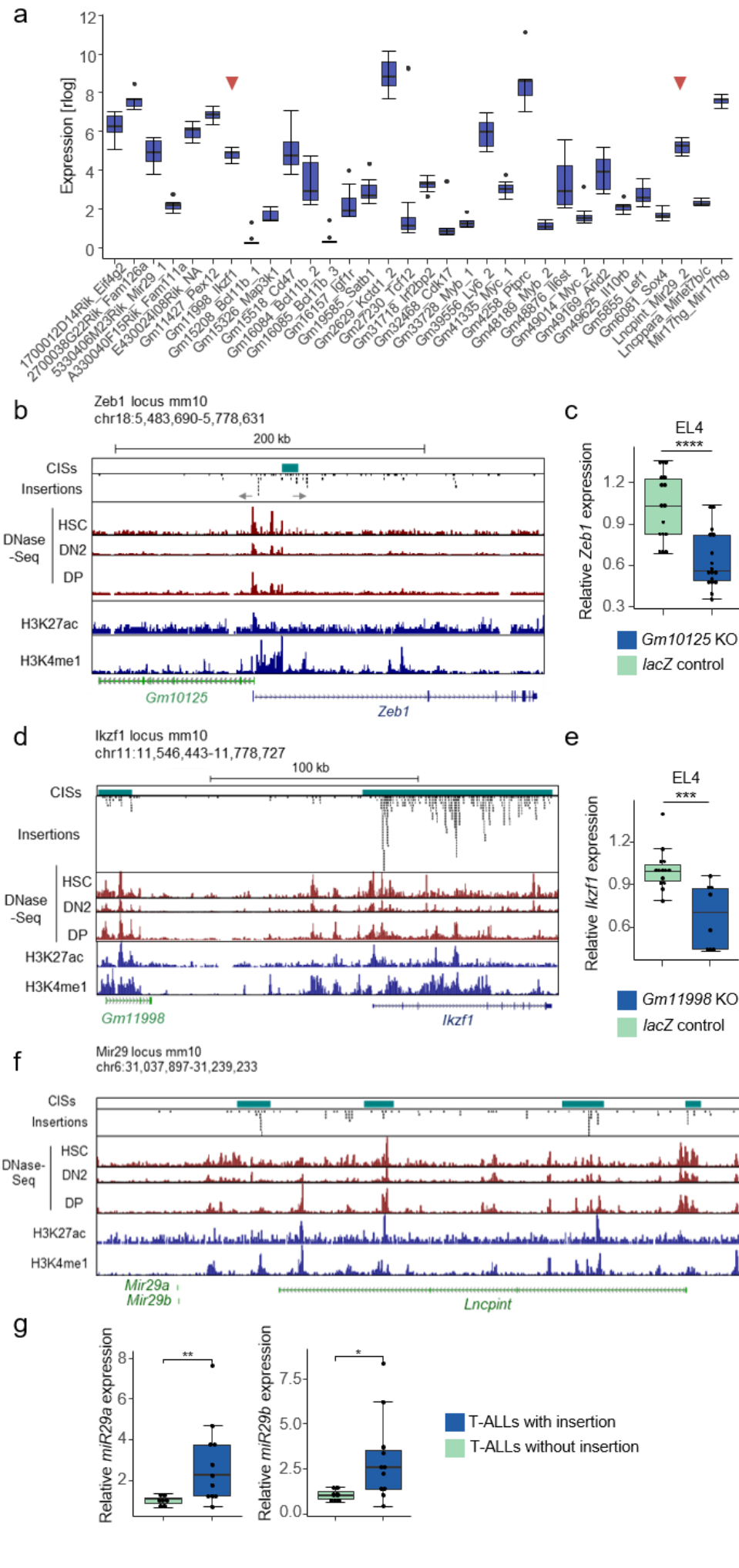
a
Expression [rlog]
1700012D14Rik_Eif4g2
2700038G22Rik_Fam126a
5330406M23Rik_Mir29_1
A330040F15Rik_Fam111a
E430024I08Rik_NA
Gm11427_Pex12
Gm11998_Ikzf1
Gm15208_Bcl11b_1
Gm15326_Map3k1
Gm15518_Cd47
Gm16084_Bcl11b_2
Gm16085_Bcl11b_3
Gm16157_Igf1r
Gm19585_Satb1
Gm26629_Kctd1_2
Gm27230_Tcf12
Gm31718_Irf2bp2
Gm32468_Cdk17
Gm33728_Myb_1
Gm39556_Ly6_2
Gm41335_Myc_1
Gm4258_Ptprc
Gm48189_Myb_2
Gm48876_Il6st
Gm49014_Myc_2
Gm49169_Arid2
Gm49625_Il10rb
Gm5855_Lef1
Gm6061_Sox4
Lncpint_Mir29_2
Lncppara_Mir17hg_Mir17hg
b
Zeb1 locus mm10
chr18:5,483,690-5,778,631
200 kb
CISs
Insertions
DNase -Seq
HSC
DN2
DP
H3K27ac
H3K4me1
Gm10125
Zeb1
c
EL4
****
Relative Zeb1 expression
Gm10125 KO
lacZ control
d
Ikzf1 locus mm10
chr11:11,546,443-11,778,727
100 kb
Gm11998
Ikzf1
e
EL4
***
Relative Ikzf1 expression
Gm11998 KO
lacZ control
f
Mir29 locus mm10
chr6:31,037,897-31,239,233
DNase-Seq
Mir29a
Mir29b
Lncpint
g
Relative miR29a expression
Relative miR29b expression
T-ALLs with insertion
T-ALLs without insertion

Figure 16: Functional relevance of identified ncRNAs. a) *rlog transformed expression values of non-coding transcript in different stages of T cell development (HSC-DP, Hu et al.(2018)). The CIS non-coding transcript identified by ARCIS and the PC gene located in close proximity are annotated. ARCIS identified 54 non-coding RNAs, 44 were annotated in the dataset and 32 were found expressed (73%). In total, the dataset includes 12,170 lncRNAs and miRNAs, of which only 1,925 are expressed (16%) with a median rlog expression level of 2,8. For in detail analysis selected transcripts (Gm11998-Ikzf1 and Lincpint-miR29) are highlighted (red arrow). The antisense transcript of Zeb1 (Gm10125) was not found expressed. b, d, f) The indicated genomic locus showing CISs (green boxes) and insertions (black) from the transposon screen together with publicly available epigenomic datasets including DNase-Seq and ChIP-Seq data.* ***b)*** *The Zeb1 locus is characterized by a CIS with a double peak with opposite transposon orientations (indicated by arrows). The Zeb1 antisense transcript Gm10125 is shown. c+e) CRIPSR/Cas9 mediated knockout of a genomic region compared to cells targeted by lacZ control guides. Nucleofected cells were single cell sorted and each dot represents RNA expression level of the target gene in one clone. Target gene expression levels were normalized to Gapdh.* ***c)*** *Zeb1 expression in clones with CRISPR/Cas9-based knockout of exon 2 and 3 of the antisense transcript Gm10125 (18 kb) in EL4 cells (KO n = 17, ctrl n = 16).* ***d)*** *The Ikzf1 locus with the distant lincRNA Gm11998.* ***e)*** *Ikzf1 expression in EL4 cells (KO n = 8, ctrl n = 14).* ***f)*** *The Lincpint/miR29 locus with multiple CIS regions. The vast majority of insertions in CIS peaks are oriented in sense with the Lincpint transcript with the promoter oriented into the direction of the miRNA29a/b.* ***g)*** *MicroRNA was isolated from tumor tissue with and without insertions in the miR29/Lincpint locus. Specific Taqman assays were used for quantification by qPCR. Expression was normalized to miR16. *P < 0.05, **P < 0.01, ***P < 0.001, ****P < 0.0001, Wilcoxon test. Adapted from Fischer et al.*

Second, a lincRNA close to the well-known leukemia-associated transcription factor *Ikzf1* was analyzed. Recently, a critical enhancer of the *Ikzf1* gene was found using laborious high-throughput enhancer assays and 4C-Seq (Alomairi et al., 2020). Here, CISs in the *Ikzf1* locus were not only identified in the coding gene (one of the top hits of the screen) but also in a lincRNA (*Gm11998*) more than 100 kb upstream of *Ikzf1* (Figure 16d). Additionally, this lincRNA was found expressed during T cell development indicating a functional role of not only the enhancer element but also the non-coding transcript (Figure 16a). CRISPR/Cas9-mediated knockout of *Gm11998* led to a reduction of *Ikzf1* expression suggesting a positive regulatory effect of the lincRNA on *Ikzf1* expression (Figure 16e).

Finally, CISs with sense-oriented insertions were found located in the lincRNA *Lncpint,* more than 150kb upstream of the *miRNA29* (Figure 16f). As *Lncpint* is described as a tumor-suppressor in T-ALL cells (Garitano-Trojaola et al., 2018), it was hypothesized that the transposon is activating *miR29* expression through activation of the pri-miRNA transcript overlapping *Lincpint* (Bouvy-Liivrand et al., 2017). Performing a microRNA specific qPCR, *microRNA29* expression levels were found increased in samples with *PiggyBac* insertions compared to samples without insertions (Figure 16g). Thus, the *PiggyBac* is able to activate miRNA expression from a far distance.

These results provide evidence that (i) transposon insertions mark functionally important non-coding transcripts, (ii) the identified transcripts alone have an effect on target gene expression upon knockout and (iii) dosage of target genes might be important for leukemia development and insertions in these regulatory transcripts might fine-tune gene expression by not inserting in protein-coding genes but rather in regulatory transcripts.

3.5 Phenotypic diversification of T-ALL subtypes

As mentioned in chapter 3.1.1, a subgroup of transposon-induced T-ALL samples showed expression of the mature T cell marker CD4 in histopathological analysis. This sporadic CD4 expression let to the hypothesis that the identified T-ALL cohort consists of multiple subtypes. To prove this, RNA sequencing of tissue samples was performed and expression patterns were analyzed. RNA from the thymus of five healthy wild type mice was used as a control. A first analysis of the data revealed that two samples behaved differently than the others. The Tdt negative samples, which were diagnosed as mature T cell lymphomas (MTLs) earlier, cluster apart from the T-ALL samples in a principal component analysis (PCA) (Figure 17a) confirming their different origin. After removing these MTL samples from the cohort, a heterogeneity especially along the second principal component was observed (Figure 17b). To detect subgroups within the T-ALL cohort, PAM clustering was conducted revealing three major subtypes (Figure 17c). To understand the underlying biology of the subtypes and to assign subtypes to the human counterparts, gene expression profiles (GEPs) were analyzed in detail. First, the three main clusters were compared using gene set enrichment analysis (GSEA). Hallmark gene sets from the Broad Institute and hematopoietic gene signatures were used (Figure 17d+e, respectively). The two biggest subgroups showed differentially enriched pathways in line with the human ETP-ALL versus classical T-ALL comparison. Therefore, these subgroups are referred to as ‘ETP-like’ and ‘classical’. The ETP-like subtype showed an enrichment for interleukin, Jak/Stat signaling as well as Kras and inflammatory pathways (Figure 17d). In contrast, the classical group was characterized by an enrichment of cell cycle related pathways including G2M and E2F associated genes (Figure 17d). Using the hematopoietic signatures for the GSEA confirmed the subtype assignment. The GEP of the ETP-like group seems to be comparable to that of early T cells (ETP, DN1), hematopoietic stem cells (HSC) and precursors of other lineages (GMP, pro-B cells), while the classical subtype is characterized by a GEP similar to double positive (mature) T cells (Figure 17e). Due to the lack of a murine genetic classifier, a 20-gene classifier was built to differentiate between the ‘ETP-like’ and the classical subtype. Genes enriched in ETP were linked to early T cell development (*Mef2c*, *Il7r*, *Il2ra*, *Lmo2*), the B cell lineage (*Syk*, *Lyn*, *Bcl3*), HSCs (*Spi1*, *Cd34*, *Cebpa*, *Id2*) and the innate immune system (*Lyz2*), whilst classical T-ALL showed enrichment for genes associated with T cell commitment (*Tcf7*, *Bcl11b*, *Satb1*, *Cd4*), TCR rearrangement/signaling (*Rag1*, *Themis*) or specific oncogenes (*Rasgrp1*, *Myb*) (Figure 17f).

The characteristics of the third subgroup were less clear from the initial GSEA analysis as most comparisons performed were not leading to a significant result (Figure 17c+e). Therefore, a thorough inspection of the insertion data was performed and revealed high-coverage activating *Mef2c* insertions (Figure 17g). The significantly increased *Mef2c* expression in this group

compared to the ETP-like and classical group confirmed this observation (Figure 17h). For further analyses, this group was referred to as ‘Mef2c-driven’.

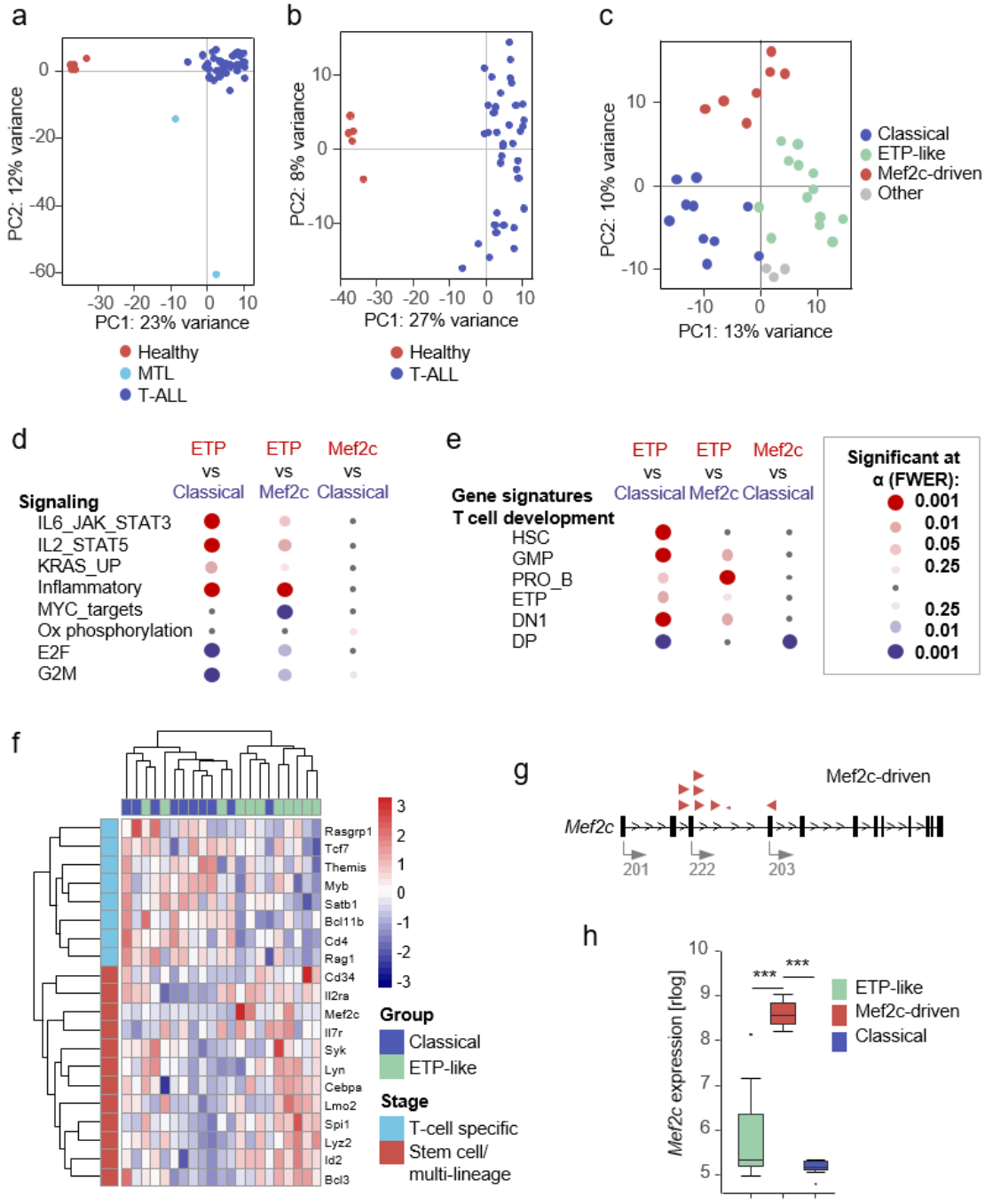

Figure 17: Phenotypic diversification of T-ALL. a-c) *Gene expression analysis of thymus samples of healthy mice and mice with T cell malignancies. RNA was isolated from 42 T-ALLs (34 from thymus tissue), 2 MTLs and 5 healthy thymi and RNA-Sequencing was performed.* ***a)*** *Principal component analysis (PCA) showing that healthy samples cluster together (red) and apart from T-ALL samples (dark blue), while the two MTL samples cluster apart from the T-ALLs.* ***b)*** *PCA showing only healthy and T-ALL samples. T-ALL samples show a variance along the PC2.* ***c)*** *PCA only showing T-ALL samples. PAM clustering was performed using k = 4 and cluster assignment is indicated by color. As one cluster only had three samples (grey) this cluster was excluded from further analysis. d+e) Gene set enrichment analysis (GSEA) to compare the three major subtypes. The size of the circle indicates the FWER value (big circles reflect high significance). Color indicates the group where the gene set is enriched (positive vs negative normalized enrichment score).* ***d)*** *Hallmark gene sets from the Broad institute (MSigDB) were used for three pairwise comparisons.* ***e)*** *Hematopoietic gene signatures obtained from Laurenti et*

al. (2013), and Novershtern et al. (2011), were used for three pairwise comparisons ***f)*** *A murine classifier gene set (n = 20) was generated to differentiate classical and ETP-ALLs. The heatmap shows z-transformed expression values.* ***g)*** *Location and orientation of insertions in the Mef2c gene in samples from the Mef2c-driven subgroup. Mef2c possesses several different isoforms starting in different exons. Arrows show direction of insertion and arrow size indicates sequencing read coverage supporting individual insertions.* ***h)*** *Mef2c expression level is increased in Mef2c-driven subtype. Rlog expression from all samples of each group is shown (***P < 0.001, Wilcoxon test). Adapted from Fischer et al.*

Expression-based analysis confirmed the previous histology-based observation that multiple subtypes are present in the T-ALL cohort. Indeed, CD4 expression was only found in samples of the classical subtypes but not in samples developing from earlier T cells including the ETP-like and Mef2c-driven subtype (Figure 18, compare Figure 6).

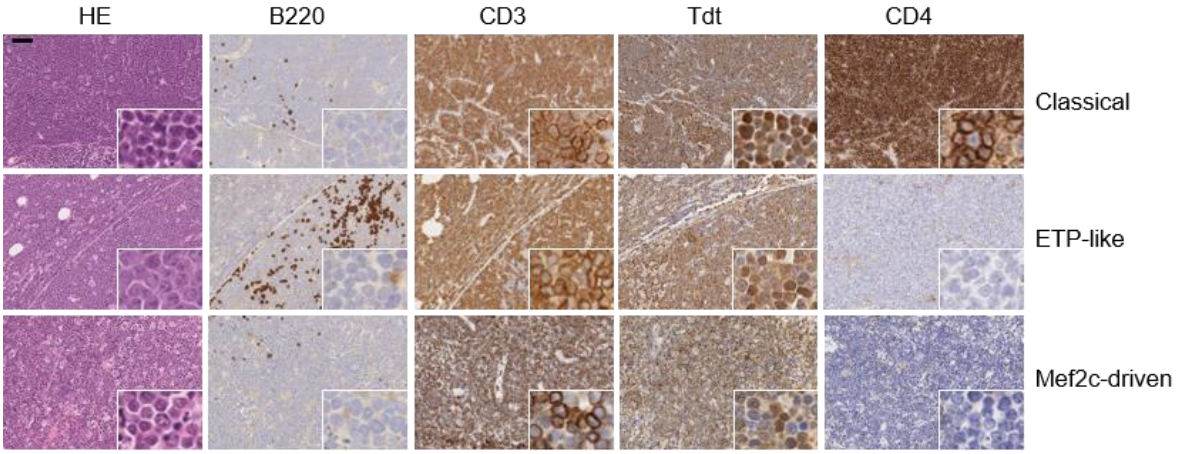

Figure 18: CD4 expression correlates with molecular subtype. *Microscopic immunohistochemical images of one representative case of each subtype. All cases were classified as T-ALL/T-LBL. Tumors are negative for the B cell marker B220 and show strong positivity for the T cell marker CD3 and the early T cell marker Tdt. CD4 positive T-ALL/T-LBL reflect the classical human T-ALL subtype. CD4 negative T-ALLs are immature T-ALLs. Scale bar, 50 µm. Adapted from Fischer et al.*

To summarize, *PiggyBac* induced T-ALLs show heterogeneous gene expression profiles and recapitulate human T-ALL subtypes.

3.6 Differential evolution of T-ALL subtypes

Next, it was explored whether the *PiggyBac*-induced T-ALLs can be used to study subentity-specific biology. Different subtypes triggered in the same experimental system allow for direct side-by-side comparisons. To investigate whether transposon insertion data can be exploited to understand evolutionary principles, several aspects of evolution were analyzed.

3.6.1 Cell of origin

Although it is well known that T-ALL can arise from different developmental precursor cells, it remains difficult to infer the cell-of-origin with standard sequencing approaches. As shown in Figure 12, *PiggyBac* has a preference for open chromatin. It was aimed to investigate whether transposon insertions profiles represent an approximation of global chromatin conformation at

the stage of genome integration. To analyze whether these profiles thereby can give an indication on a tumor´s cell of origin, CISs identified specifically in one subtype were overlapped with regions of accessible chromatin in different cell types along the T cell developmental lineage (Stage-specific ATAC-Seq from Johnson et al. (2018)). ETP-specific CISs (79%) predominantly overlapped with open chromatin regions detected in progenitor, natural killer, B or early T cells. In contrast, the majority of CISs from classical T-ALL (68%) overlapped, as expected, with ATAC-peaks specific for intermediate and late stages of T cell development (Figure 19a). These results suggest that although transposition is ongoing during the course of tumor development, insertions contain a specific level of "historical information". Using this integrated information from open chromatin in evolution and *PiggyBac* insertion data, it is possible to infer the developmental origin of the tumors.

3.6.2 Sequentiality

The molecular drivers of individual stages of T-ALL subtype evolution are poorly understood barring few examples (Albertí-Servera et al., 2021; De Bie et al., 2018). To extract information on evolutionary principles from the *PiggyBac* dataset, clonal deconvolution of transposon insertions is required. However, standard CIS calling algorithms such as CIMPL ignore the information on supporting read counts for single insertions and only consider insertions as non-quantitative events. This means that these algorithms used to search for genomic "insertion hotspots" in a cohort of mice are not able to differentiate between early and late events. In this study, a second type of analysis was established to overcome this problem. The quantitative data for each of the 170,000 insertions based on QiSeq (Friedrich et al., 2017) was integrated into the analysis as for each sample these read counts (ranging between 2 to 10,000) reflect their clonal distribution.

These analyses identified distinct clonal variances between the top CIS genes depending on the subtype. Although these top genes were identified as CISs in all T-ALL subtypes indicating a comparable insertion number, their supporting read counts and therefore the clonal distribution showed striking differences, indicating unique evolutionary hierarchies (Figure 19b). For instance, in essentially all ETP-like tumors *Ikzf1* was identified as a dominant hit based on read counts (Figure 19c). In classical T-ALL, however, insertions in *Pten* showed the highest dominance in read counts. In contrast, *Pten* showed only subclonal insertions in ETP-like tumors (Figure 19c), indicating differential sequentiality of tumor driving events in T-ALL subtypes.

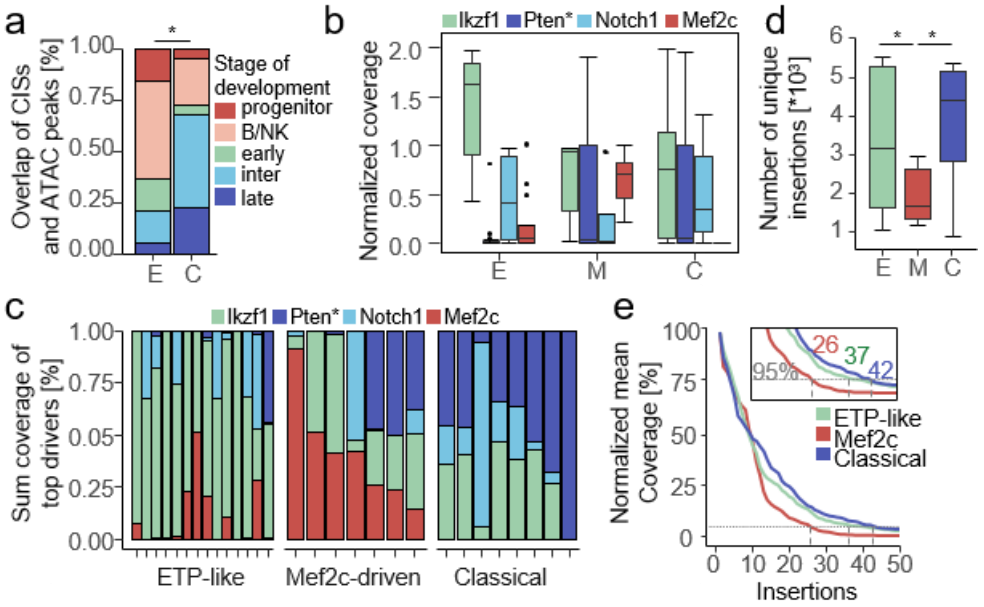

Figure 19: Exploiting insertional landscapes to understand evolution of different T-ALL subtypes. a) *Insertional landscapes predict cell of origin. Subtype specific and mutual exclusive common insertions sites (CISs) from ETP-like (n = 30) and classical (n = 117) samples were overlapped with stage-specific ATAC-Seq peaks from five different precursor stages (obtained from Johnson et al., 2018). Proportional overlap with stages is shown. (*P = 0.046, Fisher's exact test).* ***b)*** *Top CIS genes show different read coverage support in the three indicated subtypes. Normalized coverage of top CIS genes (Ikzf1, Pten, Notch1, Mef2c) is shown. The PI3K signaling and proliferation genes Rasgrp1 and Rpl11 were assigned to 'Pten'.* ***c)*** *Sum of read coverages supporting the top CIS genes for each individual sample. Read coverage of all 4 genes (or Rasgrp1/Rpl1 instead of Pten) were added together and proportions are shown for each sample.* ***d)*** *Number of unique insertions for each sample from the three indicated subtypes (*P < 0.05, Student's t test).* ***e)*** *Number of insertions required to describe 95% of reads in the tumors. The mean normalized coverage was calculated for each ranked insertion within one group. Differences in the number of insertions necessary to cover 95% of sequencing reads are shown in more detail in the inlet. a)-e) CIS and insertions used for all the described analyses were obtained from 14 ETP-like samples, 7 Mef2c-driven samples and 8 classical samples. Adapted from Fischer et al.*

3.6.3 Clonality

To additionally analyze differences in intratumor heterogeneity and the insertional burden characterizing each subtype, the number of CISs and insertions was assessed as a proxy for these characteristics. Therefore, transposon data was exploited to infer clonal architecture. First, this approach revealed that samples of the Mef2c-driven subgroup showed reduced numbers of total insertions per tumor (Figure 19d). Additionally, the clonal distribution shown by the normalized mean coverage revealed that the number of insertions in Mef2c-driven samples constituting 95% of the tumor is less than for the other subgroups (Figure 19e). Second, the Mef2c-driven group differs from the other subtypes as it was characterized by fewer CISs (Figure 20a). This observation also held true in sample-size matched permutation analyses (Figure 20b). The lower number of insertions and CISs are a measure for clonality as tumors with a very strong driver gene or combination of strong drivers (such as *Mef2c*) are

less prone for additional passenger mutations in other genes, which would lead to additional insertions.

3.6.4 Regulatory principles of subtype-specific driver genes

The subgroup-specific CISs used for clonality analyses, were examined in more detail to identify genes characteristic for the ETP or classical T-ALL subtype (Figure 20c).

The classical subgroup was characterized by two prominent features: First, CISs were found in genes linked to late thymocyte development (*Tcf12, Rpl5*), This finding is consistent with the hypothesis that the cell of origin of these classical T-ALLs is the post-commitment DP cell (Figure 20c). Second, an enrichment of CISs from the classical group were found in intergenic regulatory elements. This effect was especially pronounced among the top CISs (Figure 20d). This indicates that subtle gene regulation is especially important in classical T-ALL. The fact that these intergenic insertions were found in T cell commitment genes (such as *Bcl11b, Satb1, Ptprc* or *Runx1*) (Figure 20c) further suggests that classical T-ALLs depend on fine-tuned expression levels. In contrast to this, complete gene inactivation might lead to another phenotype or even can be deleterious at this specific stage.

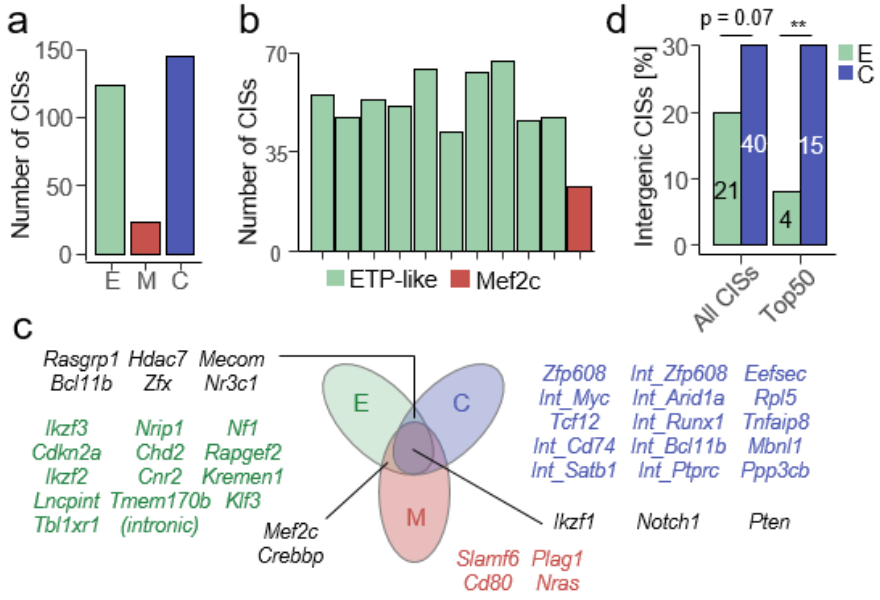

Figure 20: T-ALL subtypes differ in their clonal architecture and regulatory principles. a) *Number of common insertion sites in samples from the ETP-like subgroup (n = 14), the Mef2c-driven subgroup (n = 7) and the classical subgroup (n = 8). Only CISs with at least two insertions were considered.* ***b)*** *Permutation test to assess number of CISs in ETP-like tumor cohorts with the same sample size. Number of CISs in seven randomly selected samples from the ETP-like group. Ten different random selections of 7 ETP samples were generated and submitted to CIMPL analysis. Number of CISs was compared to the Mef2c group (n = 7).* ***c)*** *Comparison of subtype specific CISs. Shared and unique CISs are shown. Regulatory CISs are labelled with 'Int' for intergenic and 'intronic' for intronic CISs. For regulatory CISs the potential target gene is reported.* ***d)*** *Percentage of intergenic CISs amongst all and the top 50 CIS regions in the classical and ETP-like subgroup. Only intergenic CISs were considered (not overlapping with protein-coding genes). The number of intergenic CISs in each group is indicated. **P =0.009, Fisher's exact test. Adapted from Fischer et al.*

CISs specific for ETP-ALL affected mature T cell genes (inactivation of *Ikzf2, Ikzf3*), Ras pathway components (*Rapgef2, Nf1*) and potential negative regulators of Wnt signaling (*Kremen1, Tmem170b*; not linked to T-ALL so far). Moreover, several genes linked to stemness or the myeloid lineage were among the top hits in this group (*Cnr2, Chd2, Crebbp, Mecom*) (Figure 20c).

Here, the transposon insertion landscapes were exploited to map the evolutionary history of different subentities and to identify molecular forces in evolution, including cell of origin, clonality (intertumoral heterogeneity), temporal sequence (hierarchy of events) and key drivers (genes and their regulatory principles).

3.7 ETP-ALL induction by the transcription factor Spic

3.7.1 Transcription factors identified in the T-ALL cohort

As outlined in the introduction, transcription factors represent common hits of *PiggyBac* integration and were found enriched within top common insertion sites. To investigate if this holds true for the T-ALL cohort, the number of genes involved in transcriptional regulation was assessed (Figure 21a). Among the top 50 CISs, 15 genes (30%) were involved in transcriptional regulation. Additionally, TFs were quantified across the complete CIS list. Whilst only 8% of all protein-coding genes are annotated TFs in the murine genome, more than 50% of the top 10 CISs across all T-ALL samples (and 25% of the top 100 CISs) were found to be TFs (Figure 21b). These results show that *PiggyBac* screens are a suitable tool to discover TFs.

3.7.2 Subtype-specific transcription factors in T-ALL with overlap to AML

To gain deeper insight into the transcription factor landscape of ETP and T-ALL subtypes and to examine whether perturbation of transcription factors differs depending on the cell-of-origin, subtype-specific TFs were assessed (Figure 21c). For the ETP-like subgroup, inactivation of the T cell specific factors *Gata3* and *Ikzf3* and activation of the HSC/AML-related factor *Erg* was observed. Additionally, novel ETP-ALL candidate TFs including *Spic, Foxb1* and *Zfp217* were found. For the classical T-ALL subgroup well-described TFs like *Lef1, Ets1* and *Tcf12* but also novel candidates like *Zfp608, Jazf1* and *Baz2b* were detected (Figure 21c).

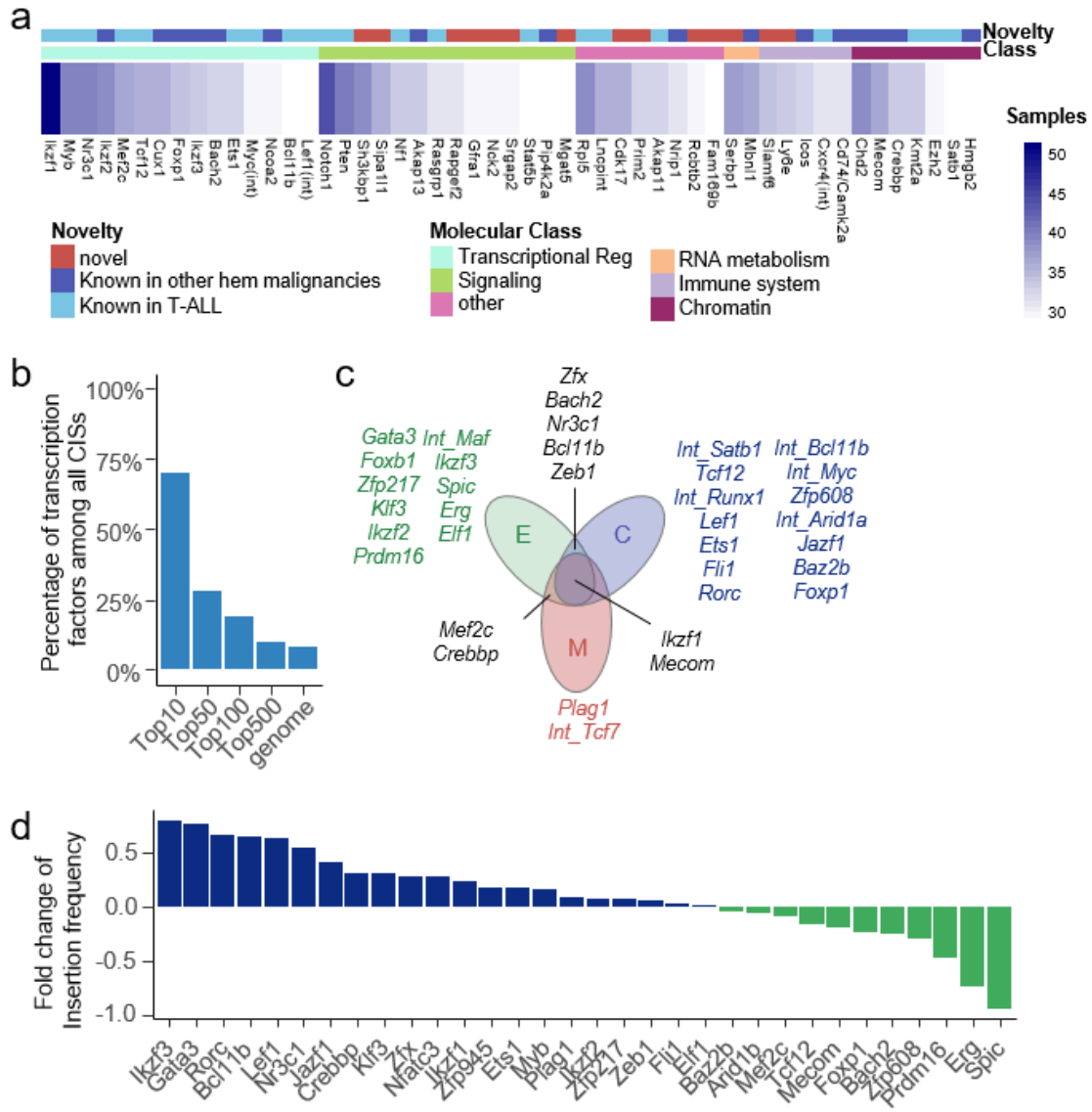

Figure 21: Transcription factors identified in the transposon screen. a) *Heatmap showing top 50 T-ALL genes and their molecular class (compare Table 26). Genes belonging to the class of transcriptional regulation are shown at the left. The heatmap shows the number of samples in which the genes are hit. The annotation of the novelty in T-ALL is based on literature research.* ***b)*** *Percentage of transcription factors identified in PiggyBac-induced T-ALLs. Percentage is shown for the top 10, top 50, top 100 and top 500 CISs. Additionally, percentage of TFs amongst all genes in the genome is indicated (8%).* ***c)*** *Shared and unique transcription factors are shown for the different subtypes. Intergenic CISs are labelled with 'Int' and the putative target gene is listed.* ***d)*** *Insertion frequency of the PiggyBac transposon in selected transcription factors. Fold change of insertion frequency in T-ALL (n = 51) and AML (n =107) samples is shown. Genes with more insertions in the T-ALL cohort show a positive fold change and are colored in blue, whilst genes enriched in AML samples show a negative fold change and are colored in green. a) adapted from Fischer et al.*

As shown in chapter 3.6.4, genes driving ETP-ALL are often involved in pathways of other hematopoietic lineages. Especially the association of ETP-driver genes with the myeloid lineage was described previously (Zhang et al., 2012). To investigate these characteristics, transcription factors were overlapped with hits found in the AML subgroup (subtypes described in chapter 3.1). The insertion frequency in selected transcription factors were compared

between AML and T-ALL samples. All T-ALL samples were combined (and not analyzed separated into subtypes) to increase the cohort size. T-ALL specific transcription factors, including *Gata3*, *Rorc*, *Bcl11b* and *Lef1,* showed an enrichment of insertions in T-ALL samples, whilst myeloid specific TFs such as *Mecom*, *Erg* and *Prdm16* were enriched in AML samples. *Ikzf3* was found as top enriched transcription factor in T-ALL samples and seems to be exclusively inactivated in T cells. Of note, *Spic* was detected as top enriched transcription factor in AML samples and was also found as CIS in ETP-like T-ALLs.

Taken together, this result exemplifies the power of the *PiggyBac* screen to uncover lineage-specific TFs as critical cancer drivers.

3.7.3 *Spic* as a candidate ETP-ALL and AML transcription factor

The insertional mutagenesis screen revealed the transcription factor *Spic* as a CIS and therefore putative candidate gene for ETP-ALL and AML development.

In both entities, transposon-based activation of *Spic* was found. The transposon insertion pattern in *Spic* suggests an oncogenic function in AML (Rad et al., 2010) and also in the T-ALL cohort (Figure 22a). Of note, AML samples showed a higher frequency of *Spic* insertions, especially for insertions with high coverage (Figure 22b). For AML, a significant increase of *Spic* expression level was detected in samples with *Spic* insertions (Rad et al., 2010). To investigate whether *Spic* expression is also increased in ETP-like ALL cases, expression of *Spic* and its family members and regulators was analyzed in the transposon cohort. While most samples of the classical subgroup express *Bach1*, a known negative regulator of *Spic*, *Spic* expression was only detectable in a subgroup of ETP-ALL samples (Figure 22c). To get more detailed insights into the role of *Spic* in T cells, open chromatin at the *Spic* locus was investigated in T cell development. As *Spic* is highly expressed in red pulp macrophages (RPMs), open chromatin in the course of T cell development was compared to RPMs. The *Spic* promoter seems only to be active during early progenitor stages (until ETP stage). In contrast, the upstream enhancer is active until T cells undergo commitment in stage DN2b (Figure 22d). As chromatin seems to be still open in early steps of T cell development, a role of *Spic* in these pre-commitment T cells is reasonable.

To study the role of *Spic* overexpression in hematopoietic development, a series of *Spic* knockin mice were engineered. To compare germline and adult onset of *Spic* overexpression, an inducible and conditional mouse model was generated (Figure 22e).

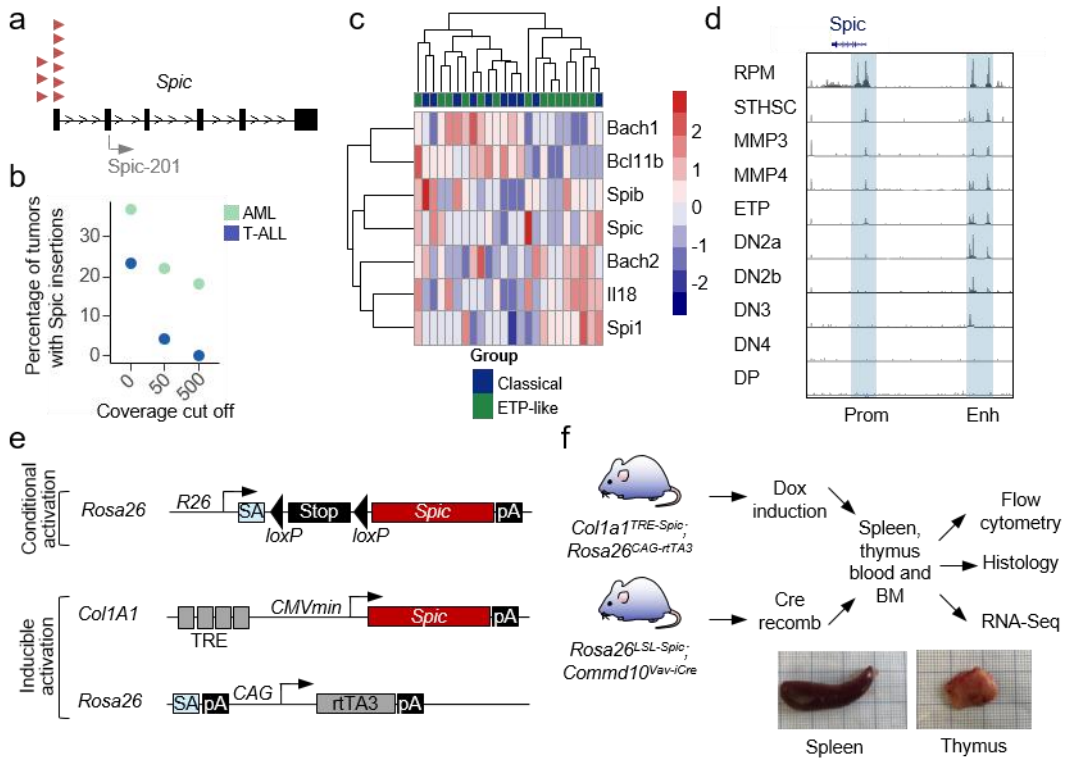

Figure 22: Spic as a candidate AML and ETP-ALL transcription factor. a) *Schematic transposon insertion pattern in the Spic gene. Murine Spic consists of 6 exons and the ATG is located in exon 2 enabling splice-based activation by the transposon. All insertions are located in sense-orientation in close proximity to exon 1.* ***b)*** *Percentage of AML and T-ALL tumors in the PiggyBac screen with insertions in the Spic gene. Percentage is shown for different coverage cut offs ranging from no cutoff (0) to a cutoff of 50 and 500 read counts.* ***c)*** *Heatmap showing expression of Spi family members (Spi1, Spib, Spic) as well as upstream (Bach1) and downstream (Il18) regulated genes in ETP-like and classical tumors. The heatmap shows z-transformed expression values.* ***d)*** *Murine Spic locus showing open chromatin in T cell development. ATAC-Seq data for different stages of hematopoietic and T cell development were obtained from IMMGEN. Red pulp macrophages (RPM) with high Spic expression are shown as a control. The Spic promoter region (Prom) and a putative enhancer region upstream of the gene (Enh) are highlighted.* ***e)*** *Alleles of mouse models generated to overexpress Spic. In the conditional model, the Spic sequence was inserted in the Rosa26 locus after a loxP flanked stop cassette. Spic expression is activated upon Cre recombination. For inducible activation, Spic sequence was inserted into the Col1a1 locus after tet responsible elements (TRE) and a minimal CMV promoter. Spic expression is induced by doxycycline administration, which mediates expression through binding the rtTA protein and the operator. Here, Spic is expressed in a whole-body approach.* ***f)*** *Schematic workflow for the inducible and conditional mouse model. After induction of Spic expression through either Doxycycline (Dox) administration or Cre recombination in the hematopoietic system, animals are monitored for signs of sickness. Sick animals were euthanized and hematopoietic organs were isolated during necropsy. For further analysis, flow cytometry, histology and RNA-Seq was performed.*

3.7.4 *Spic* mice develop hematologic malignancies

To systematically investigate the role of *Spic* in the development of hematopoietic malignancies, an inducible and conditional mouse model was generated (Figure 22e/f). The inducible model *Rosa26*$^{CAG\text{-}rtTA}$*;Col1a1*$^{TRE\text{-}Spic}$ (**rt**TA **Sp**ic, rtSP) allows the doxycycline dependent induction of *Spic* in a whole-body approach. The conditional *Rosa26*$^{LSL\text{-}Spic}$*; Commd10*$^{Vav\text{-}iCre}$ (**Va**v-iCre **Sp**ic, VaSP) model, however, restricted Spic expression to the hematopoietic system using the Vav-Cre line.

To prove that *Spic* expression can be induced by treatment with doxycycline, fibroblasts from rtSP mice were isolated and treated with doxycycline *in vitro*. *Spic* expression was measured with RT-qPCR after 3 days of doxycycline treatment. A more than 1000-fold increase of *Spic* expression was detected (Figure 23a) confirming the high sensitivity of the system. Next, *Spic* expression was induced in mice. High-dose doxycycline (625 mg/kg) was applied via food. While mice without doxycycline food did not show a phenotype over a period of one year (not shown), treated rtSP mice rapidly succumbed 4-6 days after doxycycline administration (Figure 23b). Histopathological analysis showed a degenerative architecture of the spleen accompanied with a sinus histiocytosis characterized by an activated macrophage system in the red pulp (Figure 23c). Flow cytometric analysis revealed an increase of granulocytes in the blood and monocytes in the bone marrow (Figure 23d). To understand this severe phenotype on a molecular level, RNA-Seq was performed. rtSP mice treated with doxycycline (n = 4) were compared to wild type age-matched control mice (n = 4) also treated with doxycycline (Figure 23e). GSEA revealed an increased expression of signaling pathways connected to inflammatory reactions (IL/Jak/Stat, Complement, Coagulation) in rtSP mice, while heme metabolism was found to be downregulated compared to control mice (Figure 23f). As *Spic* controls the production of RPMs, which are known to be involved in the recycling of red blood cells and heme metabolism, the drastic overexpression of *Spic* in all cells of the body might interfere with this process leading to a sudden and severe phenotype with detectable alterations in the hematopoietic system.

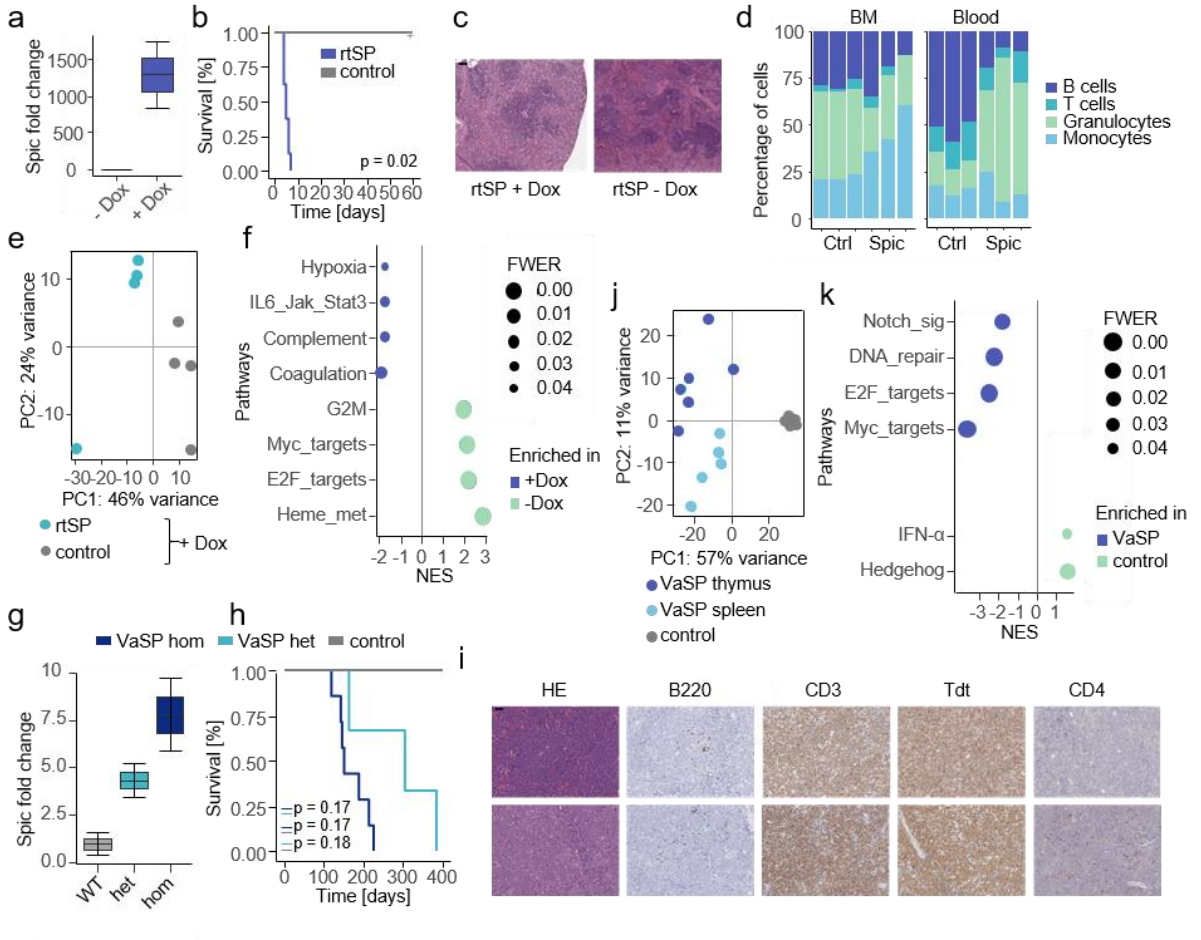

Figure 23: Spic-induced histiocytosis and ETP-like T cell leukemia. a) *Spic fold change in doxycycline treated fibroblasts isolated from Col1a1TRE-Spic;Rosa26CAG-rtTA (rtSP) mice. Spic expression level was assessed by qRT-PCR and normalized to Gapdh.* ***b)*** *Kaplan-Meier plot showing survival of rtSP mice after start of doxycycline induction via food. rtSP mice (n = 8) were compared to control mice also treated with doxycycline. (p = 0.02, log-rank test).* ***c)*** *Histopathological analysis of spleens in mice with and without doxycycline-mediated Spic induction. HE stainings of one exemplary case is shown for each group. Spleens with Spic overexpression are degenerated and show sinus histiocytosis and activation of the mononuclear phagocyte system in the red pulp. Necrotic tissue partly with hemorrhages.* ***d)*** *Flow cytometric analysis of blood and bone marrow from mice with (n = 3) and without (n = 3) Spic induction. Proportion of B cells, T cells, granulocytes and macrophages are shown.* ***e)*** *Transcriptomic analysis of rtSP mice. PCA showing clustering of rtSP (n = 4) and control (n = 4) mice.* ***f)*** *GSEA comparing rtSP and control mice. Hallmark pathways from MSigDB are compared. FWER values are depicted as circles relative to significance. Enriched pathways in the control group are displayed with a positive normalized enrichment score (NES), pathways identified in the rtSP group are displayed with a negative NES.* ***g)*** *Spic fold change in spleens of Rosa26$^{LSL\text{-}Spic}$;Commd10$^{Vav\text{-}iCre}$ mice (hetero- and homozygous VaSP) compared to wildtype mice. Spic expression level was assessed by qRT-PCR and normalized to Gapdh.* ***h)*** *Kaplan-Meier plot showing survival of hetero- and homozygous VaSP mice compared to wildtype mice (het, n = 7; hom, n = 3). (p value indicated, log-rank test).* ***i)*** *Immunohistochemical characterization and sub-classification of T cell malignancies. Samples show no or only partial CD4 expression and therefore were diagnosed as immature/ETP-like T-ALLs.* ***j)*** *Transcriptomic analysis of VaSP mice. PCA showing clustering of VaSP mice (thymus n = 6, spleen n = 5) and control (n = 5) mice.* ***k)*** *GSEA comparing VaSP and control mice. Hallmark pathways from MSigDB are compared. FWER values are depicted as circles relative to significance. Enriched pathways in the control group are displayed with a positive normalized enrichment score (NES), pathways identified in the VaSP group are displayed with a negative NES.*

To restrict *Spic* expression to the hematopoietic system, the Vav-Cre line was used, which is active in all hematopoietic cells. To confirm *Spic* expression, spleens of VaSP mice were isolated. *Spic* expression was 5 to 10-fold increased in young hetero- and homogenous VaSp mice, respectively (Figure 23g). Next, mice were monitored for tumor development. Leukemia development started after 5-6 months of age (Figure 23h). Of note, the majority of mice (5/6) developed immature T-ALLs ($B220^-$, $CD3^+$, Tdt^+, $CD4^-$) (Figure 23i) with transcriptomic differences in the spleen and thymus (Figure 23j). A minority of tumors (1/6) also showed partial CD4 expression. Transcriptomic analysis showed upregulation of *Myc, Notch1* and *Kras* signaling in T-ALLs compared to healthy thymi and a downregulation of interferon alpha and gamma response (Figure 23k).

These results show that *Spic* overexpression can indeed lead to the development of T cell malignancies and overall indicate a dose dependent role of *Spic* in the hematopoietic system. The detailed mechanism of *Spic*-induced pathogenicity will be analyzed in future studies.

4. Discussion

This thesis focused on the question if the *PiggyBac* system is suitable to screen for non-protein-coding and subtype-specific hematopoietic cancer drivers. Although transposon screens have been performed for many organs, cancer entities and genetic contexts, this is the first study demonstrating that this system can be used to interrogate the cancer's regulome. So far, technical and methodological constraints hindered us to study the cancer's regulome during evolution and to compare cancer subtypes in living organisms. Here, numerous intergenic insertions were identified, their regulatory potential was annotated and selected regions were functionally validated providing evidence that there is a widespread role for subtle gene regulation in cancer.

A second aspect of this thesis was the study of tumor evolution. A major challenge in cancer evolution research represents the distinction between driver and passenger mutations. This is a crucial task in order to understand the chronological order of accumulated genetic mutations during tumorigenesis and the biological and clinical implications. Despite novel techniques, most studies on human cancer evolution rely on retrospective (endpoint) analyses that are limited because of selection and clonal sweep during early cancer evolution. In this thesis, an approach was developed to interrogate evolution *in vivo* (in a living organism) and in a prospective manner overcoming several bottlenecks in the field. The presented advances in data analysis made it possible for the first time to reveal information on evolutionary principles such as the cell of origin, clonality and sequentiality of driver genes (Figure 24a).

Together, by combining regulatory and evolutionary data, this study provides insights into the importance of subtle gene regulation in cancer evolution. Moreover, as the first T-ALL *PiggyBac* screen is presented here, the newly established technology was used to discover biological principles driving different subtypes of T-ALL. This screen assembled catalogues of T-ALL-relevant regulatory elements and non-coding transcripts. Additionally, the phenotypic plasticity of T cells and their different subgroups were directly studied in mice allowing for a side-by-side comparison of evolutionary characteristics. The three different subtypes showed very distinct features regarding their driver genes and the temporal acquisition of hits. The more differentiated subtype of classical T-ALL was more reliable on intergenic insertions than the ETP-like group confirming a context-specific role of subtle gene regulation. In contrast, ETP-ALL samples were characterized by hits in multi-lineage genes including *Spic* (Figure 24b). *Spic* was further validated as an ETP-specific oncogenic transcription factor in mouse models. The development of a mouse model overexpressing a rather myeloid specific transcription factor leading to T-ALL induction highlights the close proximity of different hematopoietic lineages.

Taken together, this study provides methodology and tools to understand regulatory effects in T-ALL subtype evolution.

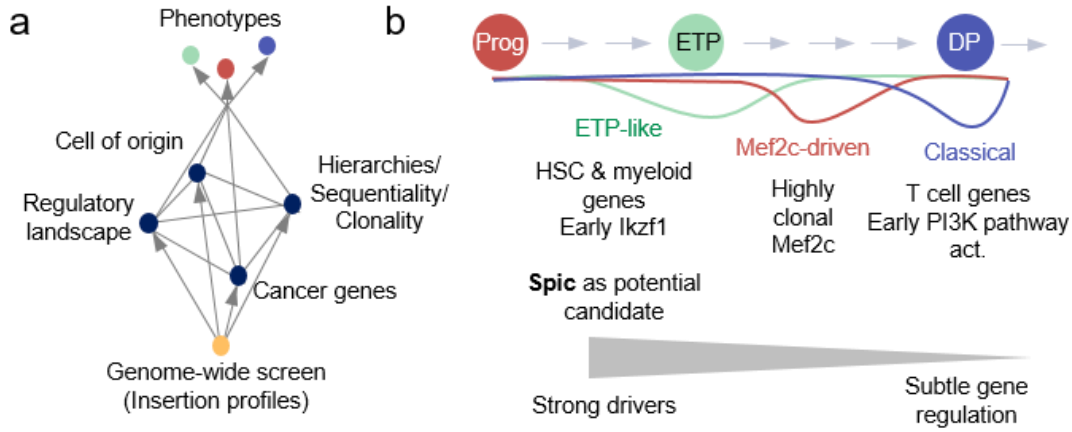

Figure 24: Methodological and biological conclusions drawn from this study. a) *Insertion profiles in a cancer induced by insertional mutagenesis can be used to analyze multiple different parameters such as cancer genes, regulatory landscapes, the cell of origin and evolution (Hierarchy, Sequentiality, Clonality). Of note, all these parameters influence the phenotypic evolution. In one in vivo experimental system, all parameters can be compared side-by-side.* ***b)*** *Biological characteristics of the three identified T-ALL subtypes. Adapted from Fischer et al.*

4.1 T-ALL subtypes induced in a pan-hematopoietic screen

In 2010, the *PiggyBac* transposon mice were first used for cancer gene discovery. By designing a transposon with a promoter especially active in the hematopoietic system (ATP2 transposon with MSCV promoter), more than 90% of mice developed aggressive leukemias and lymphomas without using specific Cre lines. Here, all hematopoietic tumors (n = 63) were analyzed together (Rad et al., 2010). To allow subtype-specific screening, this initial ATP2 transposon screening cohort was expanded. With a cohort of more than 250 mice, this study represents the largest cohort of *PiggyBac* induced hematopoietic tumors. Different subtypes of hematopoietic malignancies are covered with a sufficient sample size. Although the sample size for T-ALL (n = 53) was still low compared to AML cases (n = 107), this size enabled the study of genetic subtypes of T-ALL. For some very specific questions regarding the differential evolution of the T-ALL subtypes, however, the sample size still restricted detailed analyses (e.g. n = 7 for Mef2c-driven T-ALL).

Here, T-ALL was chosen as a model system to establish *PiggyBac* screening of the non-protein-coding genome. In T-ALL, enhancers have been described as potential oncogenic drivers. In detail, the mechanisms include either chromosomal translocations where oncogenes are driven by juxtaposing a strong T cell specific enhancer element or the specific genomic amplification of oncogenic enhancers (reviewed in Bhagwat et al. (2018)).

T-ALL subtype screening was already performed exploiting the *Sleeping Beauty* transposon system (Berquam-Vrieze et al., 2011). Using different Cre lines activated in different potential cells of origins showed that the cellular origin strongly influenced lymphomagenesis. Interestingly, the Cd4 induced model showed similarities to human ETP, while Vav- and Lck-Cre induced tumors did not show ETP characteristics. These results were unexpected as Cd4 is expressed in later stages of T cell development whilst Vav-Cre is active already in hematopoietic stem cells. Therefore, these results were questioning whether ETP cells are the cell of origin of ETP-ALL as described initially by Coustan-Smith et al. (2009). Some years later, Booth et al. (2018) confirmed ETP cells as the cell of origin of ETP-like ALL in the murine system by the generation of an ETP mouse model. The study presented here, was different to the published ones in many ways: (i) here, no Cre line was used and the transposon system was active in all cells of the body, including all cells of the hematopoietic system, (ii) the *PiggyBac* instead of the *Sleeping Beauty* system was used, and (iii) the insertion site sequencing method used in this study (QiSeq) is much more sensitive allowing for a more detailed analysis (Friedrich et al., 2017). In this study, all subtypes were induced without specific Cre lines, but only the transposon insertions were driving subtype evolution. Therefore, it was possible to compare subtype evolution directly side-by-side in the same genetic mouse model.

4.2 *PiggyBac* can be used to screen active chromatin

As previously published and described in the introduction, transposons display insertion biases regarding sequence context and functional genomic elements (de Jong et al., 2014; Yoshida et al., 2017). *PiggyBa*c was described to be biased for transcribed units and open chromatin (Li et al., 2013; Wang et al., 2008). Therefore, not all regions of the genome show an equal probability of transposon integration. This characteristic can be an advantage if screening of active chromatin is intended. Regulatory elements in open chromatin are more likely to be hit by the transposon, therefore insertions label regulatory regions. This feature of *PiggyBac* transposition was exploited in this study for the successful establishment of the *PiggyBac* system for non-protein-coding genome screening. However, there are also disadvantages of this characteristic. Unfortunately, existing statistical approaches to identify CISs do not sufficiently correct for these insertion preferences. It is assumed that all TTAA sites have an equal likelihood to be a target of insertion. de Jong et al. (2014) already discussed the problem of discerning between real CISs arising through tumorigenic selection and ‘spurious’ CISs resulting from the a priori integration bias of the transposon and described most intergenic CISs as ‘spurious’. However, at this time sensitive sequencing methods and functional data was not yet available. In this study, it was clearly shown that identified intergenic regions were not just hit by chance but label important functional elements with impact on cancer development – rather than spurious side effects.

Furthermore, during the course of this study, we developed a novel computational method (Transmicron) modelling neutral insertion probabilities based on chromatin state, transcriptional activity and sequence context (Bredthauer et al., 2023). Thereby, an appropriate background distribution of insertions was generated. Although Transmicron was superior in estimating oncogenic selection for each genomic region using Poisson regression, the background distribution was not significantly influencing resulting CISs (Bredthauer et al., 2023).

To conclude, experimental validation combined with novel computational tools showed that intergenic CISs are not artefacts of insertion probabilities but carry an important role in transposon-induced tumorigenesis.

Understanding the integration biases of transposons in more detail is not only important for insertional mutagenesis screens. Integration patterns and their influence on epigenomics also affect other research areas. The ability of the *PiggyBac* and *Sleeping Beauty* transposons to insert into DNA makes them a widely used tool in cancer research but also an interesting opportunity for gene therapy (Sandoval-Villegas et al., 2021). Transposon systems are continuously modified to show less genotoxic risk in therapeutic applications (Miskey et al., 2022). Additionally, retroviral integration sites in humans after gene therapy (Wünsche et al., 2018) as well as HIV integration sites (Lucic et al., 2019) show similar characteristics. In both studies, viral integration sites clustered with enhancers and were used to identify regulatory elements.

4.3 *In vivo* screening to study non-protein-coding genome

Oncogenesis is driven by altered cellular signaling states, resulting from genetic or regulatory changes (Bradner et al., 2017). Within a single cell, thousands of genes are dysregulated at the epigenetic, transcriptional or epitranscriptional level (Miano et al., 2021) and identifying the critical players among them is difficult. Over the past two decades, much effort in cancer genetics has focused on identifying coding mutations, a process that had transformative impact in cancer biology. However, beyond the “mutanome” there is a vast layer of molecular dysregulations that is not understood at a functional level. For example, chromatin/regulatory landscapes undergo extensive changes during oncogenesis, but their global functional interrogation has been hampered due to a lack of methods.

There is an increasing number of high-throughput CRISPR screening approaches targeting the non-protein-coding genome using catalytically inactive Cas9 proteins fused to different effector domains such as the Krüppel-associated box (KRAB) spreading repressive histone modifications (Montalbano et al., 2017; Thakore et al., 2015). However, the limitations of library-based approaches were discussed in the introduction.

This study aimed to use insertional mutagenesis to interrogate the non-coding genome. Until now, examples of transposon hits in the regulatory genome represented punctuated analyses (Weber et al., 2020), such as the discovery of miRNAs at the *Rian* locus in hepatocellular carcinoma (Dupuy et al., 2009) or a regulatory element downstream of *Cdkn2a* in pancreatic cancer (Rad et al., 2015). The low number of identified non-coding CISs in previous screens, can also be explained by the fact that there are two classes of methods for CIS detection. Here, we used a locus-centric approach where all regions in the genome are considered. However, so called gene-centric approaches focusing on regions with genetic annotation (mostly protein-coding genes) are easier to interpret and require less post processing steps (Newberg et al., 2018). Therefore, gene-centric approaches are popular but exclude the identification of non-coding drivers from the very beginning.

This study presents the first systematic, genome-wide approach to functionalize the non-protein-coding genome using transposon screening data. Beyond the results shown for T-ALL in this thesis, the abundance of these intergenic CISs could be shown in many different tumor types such as colorectal, liver, pancreatic and bile duct cancer, AML and B cell lymphomas (Fischer et al., 2023).

Although the role of non-coding sequence variants became more and more important and the overwhelming majority of cancer variants are found in the non-coding genome, the interpretation of these variants still remains challenging (Rheinbay et al., 2020). The differentiation between drivers and passengers, the identification of the linked and most likely only slightly affected target gene and the difficulty of functional validation are main reasons why non-coding sequence variants received less attention compared to protein coding ones in the last decade (Khurana et al., 2016). The recent analysis of WGS data from more than 2,500 tumors by the Pan-Cancer Analysis of Whole Genomes (PCAWG) consortium shed light on the reasons and difficulties in the detection of positive selection in ncDNA (2020; Elliott and Larsson, 2021). Association-based population studies also represent a popular method to identify non-coding cancer drivers. However, for T-ALL GWA-studies have only identified one significant intergenic T-ALL risk locus which affects *USP7* (Qian et al., 2019) highlighting that GWA-studies are not the adequate method to systematically interrogate the effect of subtle gene dysregulation in oncogenesis – and of course exploring context-dependencies poses yet another level of complexity.

This study focused on enhancer and non-coding-transcripts within the non-coding genome. However, there are other non-coding elements with an important role in cancer development: Silencers represent regions with a negative impact on gene expression (opposite to enhancer), which are, however, difficult to identify explaining why a high-throughput annotation was realized only recently (Doni Jayavelu et al., 2020). Insulators represent ‘boundaries’ in the genome and separate so called topologically associated domains (TADs) (Dixon et al., 2012).

The disruption of such boundaries can lead to oncogene induction and was described in T cell leukemogenesis (Hnisz et al., 2016). Additionally, the human non-protein coding genome comprises largely interspersed repeats (Burns, 2017) with almost half our DNA representing repeated sequences from mobile DNA (Jurka et al., 2005). In future analyses these additional regulatory elements should be also included into the ARCIS annotation pipeline.

Concerning regulatory activity, this study focused on open chromatin (ATAC-Seq, DNase-Seq) data as well as histone marks. However, DNA methylation at promoters and enhancers can also affect gene expression and was recently shown to be associated with T-ALL subtypes (Roels et al., 2020). Adding DNA methylation data was, however, beyond the scope of this study.

4.4 Experimental validation of enhancers and non-coding transcripts

In order to validate identified regulatory regions in experimental settings, the effect of intergenic insertions on target gene expression was first estimated dependent on the type of the target gene. The effect of intergenic transposon insertions on gene expression is relatively easy to trace for tumor suppressor genes. By the disruption of 'positive' regulatory elements such as enhancers, the target gene expression is diminished. For oncogenes, however, the effect can also be explained. The possibilities include several mechanisms such as (i) disruption of silencing elements, (ii) inactivation of activating non-coding transcripts or (iii) the disruption of chromosome neighborhoods as recently described in T-ALL (Hnisz et al., 2016). Moreover, the transposon itself contains strong promoter and enhancer elements, which might also influence gene expression of genes in close proximity. The detailed mechanism how intergenic insertions affect gene expression needs to be investigated in future studies.

In addition to intergenic insertions in enhancers or non-coding transcripts, we found several regulatory elements in introns. Introns were shown to be involved in many steps of mRNA processing (Chorev and Carmel, 2012). However, the identification of functional introns remained challenging (Chorev et al., 2017). With the approach described in this study it became possible to identify intronic regions which are marked by an accumulation of insertions and most likely be specifically relevant in regulatory networks. Thereby, an intronic enhancer in the *Rnls* gene was identified regulating *Pten* expression. In contrast to the *Pten* example for distant enhancers in introns of other genes, we also identified intronic enhancers connected to the gene they are located in. Although these 'own intronic enhancers' were not investigated in detail in this study, they could be of further interest: it was shown that while extragenic enhancers positively contribute to transcriptional output, intragenic enhancers play an unanticipated role in attenuating host gene expression (Cinghu et al., 2017). The details on how insertions influence gene expression by inserting in intronic enhancers need further investigation.

Discerning functional noncoding transcripts from a vast transcriptome represents a priority for the lncRNA field (Kopp and Mendell, 2018). The transposon screening system described in this study identified many intergenic common insertion sites. We established a computational approach to annotate their function and manually and individually inspected all intergenic CISs to classify them to different categories such as intergenic enhancers, intronic enhancers or non-coding transcripts. However, many CISs remained ambiguous due to the overlap of multiple functional elements in the genome. Insertions in introns can overlap with enhancers and non-coding transcripts (antisense RNA or sense-intronic RNA) and intergenic lincRNAs may additionally harbor enhancers in their introns. As the annotation of non-coding transcripts is far from being complete, a nPC transcript can never be excluded even if there is no transcript annotated so far in the respective region. To resolve this ambiguity, gene expression data was used. High-coverage RNA-Seq or GRO-Seq data was used to detect polyA containing lncRNAs and other native transcripts, respectively. Some cases, however, were still difficult to classify.

To prove enhancer activity, a selection of plasmid-based enhancer activity assays can be used where the candidate DNA fragment is placed downstream of a reporter gene (Muerdter et al., 2018). However, many widely used assays were shown to be unreliable due to the specific characteristics of bacterial DNA and activation of an immune response (Muerdter et al., 2018).

Here, CRISPR/Cas9-based knockout was used to functionally validate selected regulatory elements. Therefore, the complete regulatory region (6-18 kb) was deleted from the genomic DNA. Although this method worked with a surprisingly high efficiency and revealed a link between regulatory elements and target gene expression, unfortunately, it was unable to differentiate between the effect of enhancers and non-coding transcripts. To directly target non-coding transcripts in future studies, the RNA targeting CRISPR enzyme Cas13 (Abudayyeh et al., 2017), which is efficiently binding and cleaving RNA instead of DNA, should be used. Targeted RNA knockdown will help to differentiate between the knockout effect of the regulatory DNA region (which might also contain enhancer regions) and the effect of the RNA transcript itself. During the course of this study, however, this method was not yet suitable for high-throughput applications.

An ambiguous example for the challenge described above is the *Ikzf1* locus. In this study, a potential lincRNA was identified regulating *Ikzf1*, a crucial transcription factor in the hematopoietic system. However, recently an *Ikzf1* enhancer was published overlapping the lincRNA (Alomairi et al., 2020). Here, we show that this lincRNA is expressed supporting the hypothesis that the lincRNA is involved in leukemogenesis. However, due to the mentioned ambiguity, the mechanism was not solved and needs further investigation through direct targeting of the non-coding transcript with e.g. Cas13.

Another example is the *Lncpint* locus where multiple CISs overlap with the lncRNA *Lncpint* but also the pri-miRNA transcript of miR29a/b. There is evidence that members of the microRNA29 family play a critical role in human cancer (Volinia et al., 2006), however, there are contrasting findings reported for miR29 levels in different hematopoietic malignancies. Whilst lower levels of miR29 members were described in aggressive subtypes of chronic lymphocytic leukemia and mantle cell lymphoma (Calin et al., 2005; Zhao et al., 2010), the ectopic expression of miR29a in murine HSCs led to AML (Han et al., 2010). In T-ALL, low miR29a levels were associated with an altered epigenetic status (Oliveira et al., 2015). Here, we found increased expression levels of miR29a/b driven by transposon insertions arguing for a tumor-promoting role of miR29 in this context. To clarify the role of miR29 in leukemogenesis, directly targeting the microRNAs and the *Lincpint* transcript in cell lines could reveal the function.

4.5 *PiggyBac* screening exploited to study tumor evolution

A disadvantage of studying tumor evolution in humans is the retrospective analysis only capturing the last events of the evolutionary tree. Here, a forward genetic screening approach was described where the timing of the lesion (insertion) and the connected effect can be assessed. Although insertion site sequencing was adapted to Illumina sequencing a long time ago (Brett et al., 2011), the sequencing depth has not yet been exploited for studying evolution. Now, for the first time, read coverage supporting single insertions were systematically analyzed to understand evolutionary hierarchies. Using this approach, *Ikzf1* and *Pten* were found as major drivers of the ETP-like and classical T-ALL subgroup, respectively. Of note, these genes were found as top CISs in both subtypes and usual CISs methods were not able to identify any differences. Applying this novel approach, biologically relevant associations were identified: Strong positive selection for *Ikzf1* insertions in virtually all ETP-ALL establishes a critical role of *Ikzf1* in the initiation of this T-ALL subtype. In line with this, *IKZF1* alterations were found to be enriched in human ETP-ALL as compared to classic T-ALL (Zhang et al., 2012). Here, two new dimensions are added to this association: a functional (*in vivo* relevance) and a temporal (stage). In contrast, the dominance of *Pten* inactivation in classical mouse T-ALLs reflects specific evolutionary constraints and rationalizes the enrichment of PI3K pathway alterations in mature forms of human T-ALL.

Furthermore, a biologically striking observation was the fact that many genes typical for T-ALL were not hit by the transposon. T-ALL is a disease characterized by oncogenic transcription factors (Liu et al., 2017). However, the typical T-ALL transcription factors such as TAL1/2, LYL1, TLX1/3, NKX2-1/2-2/2-5, LMO1/2 were not hit by the transposon. This can be explained by the restriction of the *PiggyBac* transposon technology used in this study, which is only able to activate protein-coding genes that harbor an (alternative) ATG in the second or following exon. As *PiggyBac* activation is dependent on splicing and the first exon does not harbor a

splice acceptor, proteins containing their only ATG in exon 1 cannot be activated by the transposons (however, exceptions of cryptic activation are described) (Weber et al., 2020). To circumvent this problem in future studies, novel mouse models are being established which will enable the evolutionary tracking of all protein-coding genes (unpublished data: Roland Rad, Technical University Munich).

In this study, tumor samples were analyzed presenting only a snapshot and not capturing dynamics of tumor evolution. To get more valuable insights into tumor evolution, not only the final tumor samples but also blood samples at specific time points can be isolated and analyzed with QiSeq. In an independent project, a cohort of mice developing B cell lymphomas was analyzed during the complete course of lymphomagenesis. This analysis gave interesting insights into the detailed sequentiality and remobilization of transposon insertions (unpublished data: Roland Rad, Technical University Munich, Munich, Germany and Ursula Zimber-Strobl, Helmholtz Center Munich).

The combination of longitudinal sampling, sequential blood collection, high-resolution sequencing methods and newly developed analytical tools will give much more detailed insights into blood cancer evolution in future studies. The described concept that insertions carry historical information can in future studies be expanded to other cancer types.

4.6 Biological characteristics of induced T-ALL subtypes

4.6.1 Human counterparts of ETP-like and classical T-ALL

The differentiation of the ETP-like and classical T-ALL subtypes was based on immunohistochemistry and gene expression profiling. Both methods clearly differentiated mature and immature T cell malignancies. Whereas all T-ALLs showed positivity for Tdt, classical T-ALL samples showed (partial) positivity for the mature marker CD4 in immunohistochemistry. In gene expression profiles, hallmark gene sets for ETP-ALL such as Jak/Stat and Kras signaling were identified in ETP-like murine leukemias. Additionally, the comparison of gene expression to human gene signatures of hematopoietic cells confirmed the mature origin of classical T-ALLs (double positive T cells) whilst ETP-like leukemias showed similarity to multiple immature lineages including hematopoietic stem cells, myeloid and B cell progenitors as well as early T cells. Moreover, this study established a murine classifier gene set to distinguish between ETP-like and classical T-ALLs. Therefore, genes were chosen which are described to be enriched in one of the two subtypes. For ETP, genes linked to early T cell development, the B cell lineage, HSCs and the innate immune system were selected. For classical T-ALLs, genes associated with T cell specific processes such as T cell commitment and TCR rearrangement/signaling were included. The established 20-gene

classifier has reliably distinguished between both subtypes and will facilitate future studies on T-ALL subtype evolution.

In the copy number analysis using aCGH, however, deletions at the T cell receptor loci were found in almost all samples. According to literature, immature T-ALLs show lower frequencies of TCR rearrangements compared to more mature leukemias (Neumann et al., 2013). In this study, no difference in TCR deletion frequencies was observed between the subtypes. A possible explanation for the fact that comparable numbers of TCR deletions were found in the ETP and the classical subtype include the early activation of the Rag enzyme in T cell development (Welner et al., 2009). In the aCGH data, a deletion as a consequence of Rag activity rather than a fully rearranged locus is detected potentially explaining why the deletions can be found in all samples. T cell precursors initiate TCR rearrangements even before reaching the thymus but the rearrangement is not successful and not leading to TCR gene expression (Rothenberg, 2019). Most likely, aCGH only detects precursor rearrangements as deletions but is not able to differentiate between these precursor deletions and a fully rearranged T cell receptor locus. T cell receptor sequencing approaches would be necessary to clarify the difference of TCR rearrangements between subtypes.

4.6.2. Mef2c-driven T-ALL as a single disease entity?

In addition to the well-described T-ALL subtypes ‘ETP-like’ and ‘classical’, this study identified a subgroup of T-ALLs driven by *Mef2c*. The myocyte enhancer factor 2C was originally identified in muscle development but is also associated with an oncogenic function in different leukemias (Canté-Barrett et al., 2014). In hematopoietic development, *Mef2c* is highly expressed in HSCs, myeloid progenitors and B cells before commitment. However, *Mef2c* is absent in the T cell lineage (Canté-Barrett et al., 2014).

In T-ALL, *MEF2C* is a downstream target of the cardiac homeobox gene *NKX2-5* (Nagel et al., 2008) and was itself found involved in chromosomal rearrangements (Homminga et al., 2011). However, in T-ALL patients there is controversy as to whether MEF2C-dysregulated and ETP-ALL feature a single disease entity. On the one hand, it was shown that ETP-ALLs and immature MEF2C-dysregulated T-ALL exclusively overlap (Zuurbier et al., 2014). In contrast, it was reported that MEF2C-dysregulated T-ALLs were only partially associated with the immunophenotype of ETP-ALL (Colomer-Lahiguera et al., 2017). Although all studies confirm an immature origin of both groups, the question arises whether these groups differ at the biological or genetic level.

This study supports the view that Mef2c-dyregulated T-ALL should be considered as a separate disease entity from ETP-ALL. Differences between both groups were shown at the transcriptional level (clustering and enriched pathways), insertional level (high coverage,

activating *Mef2c* insertions as distinguishing factor between the groups) and based on the clonal architecture. The Mef2c subgroup was characterized by a lower number of insertions and common insertion sites what can be seen as a measure for clonality. Therefore, Mef2c-driven tumors were highly clonal compared to the other subtypes confirming that *Mef2c* is a strong oncogene in hematopoietic malignancies.

Investigating the molecular differences between both immature subtypes in more detail in the future might also have clinical relevance. It was shown that there are differences in treatment response and resistance between ETP-ALL and MEF2C T-ALL, with the latter responding poorly to glucocorticoids (Colomer-Lahiguera et al., 2017). Furthermore, the MEF2C status might be important as MEF2C phosphorylation (S222) was identified as biomarker for primary chemoresistance in AML, which can be circumvented by a selective inhibitor (Brown et al., 2018).

4.7 *PiggyBac* screening reveals extensive quasi-insufficiency in cancer evolution

In this study, the relevance of intergenic transposon insertions disrupting regulatory elements was highlighted. The phenotypic impact of enhancer alterations was validated *in vitro* and in novel mouse models (described below) and differs substantially from coding sequence insertions. Alterations in regulatory elements can have additive effects and the modularity of enhancers allows fine-tuning of gene expression (Gordon and Lyonnet, 2014). It was speculated that many cancer genes rely on a very specific gene dosage to exert their tumor-promoting effect (Berger et al., 2011). As transposons most likely disturb regulatory elements and abolish their activating effect on gene expression, a subtle decrease of gene expression is probably the most common effect of the transposon. Dynamic and subtle dosage changes of tumor suppressor genes were described earlier with the term ‘quasi-insufficiency’ (Berger et al., 2011). While haplo-insufficient tumor suppressors rely on the loss of one allele, a continuum model was described for quasi-insufficient genes (Berger et al., 2011). However, it is notable that since the definition of this term in 2011, this phenomenon was not described again. In the last decade, we looked at mutations in a binary way (heterozygous vs homozygous), but the role of quasi-insufficiency remained largely unexplored. A possible explanation is the lack of suitable tools enabling scalable interrogation of quasi-insufficiency, which requires genome-wide subtle perturbations and an experimental system that can capture the relevant readout: cancer development in an organism. The transposon screening approach fulfills these requirements and shows that subtle gene regulation is a major contributor to oncogenesis. A large part of screening hits are heterozygous hits in regulatory elements, leading to subtle gene dysregulations. It was shown that such subtle gene dysregulation can indeed be highly oncogenic, and we refer to this as quasi-insufficiency. This

study led to the creation of catalogues of quasi-insufficient cancer genes and showed their enrichment with human GWAS hotspots (which typically also have only subtle effects). Additionally, the relevance of intergenic CISs and subtle gene regulation shown for T-ALL in this thesis could be validated across multiple cohorts (Fischer et al., 2023).

By comparing the induced T-ALL subtypes, context-dependencies of quasi-insufficiency were investigated. It was shown that there is an association of the cell-of-origin with the number of regulatory CISs identified in the respective subtype. In detail, classical T-ALL was characterized by an increased number of regulatory CISs. A similar dependency was found in AML samples where immature *Erg*-driven leukemias showed less intergenic CISs compared to more mature leukemias driven by *Mecom* or *Prdm16* (Fischer et al., 2023). A possible explanation of this context-dependencies includes the decrease of multilineage potential during differentiation. While immature cells might need strong, genic hits to avoid differentiation, mature cells might 'only' need to fine tune gene expression of lineage-specific tumor genes. The cell of origin of classical T-ALLs lost multilienage potential and only needs to adjust the level of T cell tumor genes to proliferate.

As it is extremely unlikely that the transposon inserts on both alleles of a regulatory region, it was assumed that interference with the function of the regulatory elements is generally mono-allelic. This suggests that malignant transformation can be promoted by very subtle interference with gene regulation. To address this, we engineered mice with kilo- to megabase scale germline deletions in the regulatory region downstream of *Bcl11b* ((Fischer et al., 2023), doctoral thesis of Robert Lersch). We found that deletions in the regulatory region of a well-described tumor suppressor gene on an otherwise wild-type background are sufficient to induce prominent cancer phenotypes. Intriguingly, the exact position and size of the deletion and the knockout dosage (hetero- or homozygosity) were profoundly affecting the tumor type and frequency. We observed a higher penetrance of T-ALL in mice with a larger deletion suggesting additive effects of enhancers and confirming a context-dependent role of quasi-insufficiency.

Altogether, this data indicated widespread roles of quasi-insufficiency in tumor evolution and suggests a dependency on the cell of origin in oncogenic transformation.

4.8 *Spic* in T cell leukemogenesis

The transcription factor *Spic* was found as oncogenic CISs in the AML cohort of the first *PiggyBac in vivo* screen (Rad et al., 2010) and also in the T-ALL cohort of this study. Data from CRISPR/Cas9 drop-out screening in cancer cell lines showed a selective *SPIC* dependency of cell lines connected to the hematopoietic system (Source: Depmap, data not shown). This underlines a potential oncogenic function of *Spic* in hematopoietic malignancies.

In the AML cohort, the coverage of *Spic* insertions was high and the effect on *Spic* expression was clearly activating (Rad et al., 2010). However, *Spic* insertions were found to be of low coverage in T-ALL and *Spic* is not expressed in human and murine T cells (Immgen data, not shown). Therefore, the observation that conditional *Spic* overexpressing mice show a T-ALL phenotype was intriguing. The hypothesis to explain this phenomenon includes the possibility that *Spic* mimics *Spi1* function in T cells. SPI1 is described as important transcription factor in early T cell development and SPI1 fusions were described in human ETP-ALL patients associated with poor outcome (Seki et al., 2017). *Spi1* is expressed in prethymic progenitors and ETP cells and replaced by the expression of *Tcf7* during T cell commitment (Rothenberg et al., 2019; Ungerbäck et al., 2018). The increased *SPI1* expression keeps immature T cells in their pre-commitment state. An increased expression of *Spic* might have a similar effect on the differentiation block of T cells. In line with this, it was shown recently that PU.1 and Spic share a 5'-GGAA-3' motif indicating that *SPIC* can compete with PU.1 binding (Laramée et al., 2020). The fact that *Spic* overexpression leads to ETP-ALL rather than classical T-ALL in the model presented in this study might indicate that *Spic* imitates *Spi1* function associated with early myeloid progenitors, but usually switched off in T cell maturation.

A second hypothesis how Spic induction could lead to ETP-ALL development includes lineage plasticity of myeloid and early T cells. Although the development of T cell leukemias was intriguing as the screening data suggested a more prominent role of *Spic* in the myeloid system, there is extensive literature on the concept of lineage fate and plasticity between T and myeloid cells. Precursor T cells loose B cell potential early during differentiation but keep myeloid, dendritic cell and natural killer cell potential in early stages within the thymus (Rothenberg, 2019). This can be explained by a mixture of T cell- and stem cell transcription factors in early T cell progenitors (Laiosa et al., 2006). The resulting instability is the basis for alternative lineage fates of these uncommitted progenitor T cells. The reason to retain this plasticity is not yet fully clear. A possible explanation includes that the proliferation of early T cells is controlled by the same mechanism as the proliferation of multipotent progenitors (Rothenberg, 2019). However, contradictory studies are showing that T cell progenitors lack the potential for myeloid lineages in the thymus and that an early separation into lymphoid and myeloid branches is the correct *in vivo* scenario (Schlenner and Rodewald, 2010). Although not finally decided yet, it seems highly likely that there exists an overlap or lineage promiscuity between T cells and myeloid cells. This is further supported by the existence of mixed phenotype acute leukemias (MPALs) or acute leukemias of ambiguous lineage (ALAL). This disease entity comprises a collection of leukemias with features from both, ALL and AML. It was recently shown that the priming of these leukemia cells depends on the cell of origin and the founding deletion (Alexander et al., 2018). The cell of origin of theses MPALs is most likely

a multipotent progenitor cell (Granja et al., 2019). Defining the precise mechanism of *Spic* in this lineage ambiguity and induction of ETP-ALL needs further experimental investigation.

Although the induction of ETP-ALL by *Spic* overexpression could not finally be resolved mechanistically, the fact that *Spic* overexpression led to an ETP-ALL phenotype in a mouse model might be an important information for future studies. Since transgenic mice overexpressing *Spi1* develop erythroleukemias (Moreau-Gachelin et al., 1996), these animals cannot be used for ETP-ALL studies. So far, only complex genotypes combining several tumor suppressor and oncogenes (e.g*. Ezh2, Runx1, Flt3*) were able to induce ETP-like ALL in the mouse (Booth et al., 2018). The *Spic* mouse model might be of use for further studies on the *Spi* family of transcription factors and their role in ETP-ALL development. Finding alternative treatment options for the subsets of patient with increased *SPI1* expression is of high importance and the proposed mouse model could be used to identify and test targets.

In contrast to the ETP-ALL phenotype induced by the conditional overexpression of *Spic* in the hematopoietic system, the phenotype of the inducible whole-body overexpression of *Spic* was less clear. Mice succumbed very suddenly already 5 days after administration of doxycycline. Flow cytometry as well as histological examination indicated alterations in the hematopoietic system and expression analysis confirmed a dysregulation of the complement pathway and heme metabolism. Although heme metabolism is connected to leukemogenesis and therapy resistance (Lin et al., 2019), the cause of the very sudden death remained unclear. To investigate if the observed effect is only due to changes in the hematopoietic system, the inducible allele was combined with the Vav-Cre mouse to restrict the inducible expression of *Spic* to hematopoietic cells. Future experiments will clarify if these mice also show this very sudden and severe phenotype.

4.9 Outlook

This study can be seen as a natural evolution of previous *PiggyBac* screening efforts, which primarily interrogated the protein-coding genome (Rad et al., 2010; Rad et al., 2015; Weber et al., 2019). Here, novel tools and methods are described opening up two entirely new application areas for transposon-based *in vivo* screening: the cancer's regulome and evolution.

The developed methods, analytical approaches and computational pipelines are universally applicable for any cancer type and can also be used for retrospective data analysis of existing data. Additionally, the biological discoveries are of broad relevance, even beyond T-ALL. This was supported by the analysis of extensive data sets from a large number of genome-wide *in vivo* screens (~1450 solid and hematopoietic cancers from 15 *in vivo* screens from our

laboratory, the largest part yet unpublished; not shown), which reinforced, for example, the broad relevance of quasi-insufficiency across cancer types.

In this study, a novel approach was developed to annotate the putative effect of regulatory common insertion sites. Therefore, a combination of a computational annotation part based on a large set of publicly available epigenomic data was implemented. Additionally, selected CISs were inspected manually and individually as often the overlap of transcripts and regulatory elements in the genome made a definitive conclusion difficult. Future approaches for CIS identification and annotation might use machine learning tools with the aim to better identify CISs and their potential effect.

The study covers two of the newly defined hallmarks of cancer: non-mutational epigenetic reprogramming and phenotypic plasticity (Hanahan, 2022). This underlines the importance of the newly established analytical approach as both hallmarks are at the moment discussed as important reasons for therapy failure and resistance. The *PiggyBac* system is able to investigate mutation-driven and epigenetic regulation in the evolution of tumors in the same experimental approach. Additionally, it was shown that the *PiggyBac* system identifies genes important in multiple hematopoietic lineages (*Spic*) probably involved in dedifferentiation, block of differentiation and/or trans-differentiation, mechanisms crucial to understand cancer phenotypes in the future.

5. Bibliography

(2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57-74.

(2020). Pan-cancer analysis of whole genomes. Nature *578*, 82-93.

Abudayyeh, O. O., Gootenberg, J. S., Essletzbichler, P., Han, S., Joung, J., Belanto, J. J., Verdine, V., Cox, D. B. T., Kellner, M. J., Regev, A.*, et al.* (2017). RNA targeting with CRISPR-Cas13. Nature *550*, 280-284.

Alaggio, R., Amador, C., Anagnostopoulos, I., Attygalle, A. D., Araujo, I. B. O., Berti, E., Bhagat, G., Borges, A. M., Boyer, D., Calaminici, M.*, et al.* (2022). The 5th edition of the World Health Organization Classification of Haematolymphoid Tumours: Lymphoid Neoplasms. Leukemia *36*, 1720-1748.

Albertí-Servera, L., Demeyer, S., Govaerts, I., Swings, T., De Bie, J., Gielen, O., Brociner, M., Michaux, L., Maertens, J., Uyttebroeck, A.*, et al.* (2021). Single-cell DNA amplicon sequencing reveals clonal heterogeneity and evolution in T-cell acute lymphoblastic leukemia. Blood *137*, 801-811.

Alexander, T. B., Gu, Z., Iacobucci, I., Dickerson, K., Choi, J. K., Xu, B., Payne-Turner, D., Yoshihara, H., Loh, M. L., Horan, J.*, et al.* (2018). The genetic basis and cell of origin of mixed phenotype acute leukaemia. Nature *562*, 373-379.

Alexandrov, L. B., Kim, J., Haradhvala, N. J., Huang, M. N., Tian Ng, A. W., Wu, Y., Boot, A., Covington, K. R., Gordenin, D. A., Bergstrom, E. N.*, et al.* (2020). The repertoire of mutational signatures in human cancer. Nature *578*, 94-101.

Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Aparicio, S. A., Behjati, S., Biankin, A. V., Bignell, G. R., Bolli, N., Borg, A., Børresen-Dale, A. L.*, et al.* (2013). Signatures of mutational processes in human cancer. Nature *500*, 415-421.

Alimonti, A., Carracedo, A., Clohessy, J. G., Trotman, L. C., Nardella, C., Egia, A., Salmena, L., Sampieri, K., Haveman, W. J., Brogi, E.*, et al.* (2010). Subtle variations in Pten dose determine cancer susceptibility. Nat Genet *42*, 454-458.

Alomairi, J., Molitor, A. M., Sadouni, N., Hussain, S., Torres, M., Saadi, W., Dao, L. T. M., Charbonnier, G., Santiago-Algarra, D., Andrau, J. C.*, et al.* (2020). Integration of high-throughput reporter assays identify a critical enhancer of the Ikzf1 gene. PLoS One *15*, e0233191.

Ameres, S. L., and Zamore, P. D. (2013). Diversifying microRNA sequence and function. Nat Rev Mol Cell Biol *14*, 475-488.

Belver, L., and Ferrando, A. (2016). The genetics and mechanisms of T cell acute lymphoblastic leukaemia. Nat Rev Cancer *16*, 494-507.

Bergemann, T. L., Starr, T. K., Yu, H., Steinbach, M., Erdmann, J., Chen, Y., Cormier, R. T., Largaespada, D. A., and Silverstein, K. A. (2012). New methods for finding common insertion sites and co-occurring common insertion sites in transposon- and virus-based genetic screens. Nucleic Acids Res *40*, 3822-3833.

Berger, A. H., Knudson, A. G., and Pandolfi, P. P. (2011). A continuum model for tumour suppression. Nature *476*, 163-169.

Berquam-Vrieze, K. E., Nannapaneni, K., Brett, B. T., Holmfeldt, L., Ma, J., Zagorodna, O., Jenkins, N. A., Copeland, N. G., Meyerholz, D. K., Knudson, C. M.*, et al.* (2011). Cell of origin strongly influences genetic selection in a mouse model of T-ALL. Blood *118*, 4646-4656.

Bhagwat, A. S., Lu, B., and Vakoc, C. R. (2018). Enhancer dysfunction in leukemia. Blood *131*, 1795-1804.

Bonev, B., and Cavalli, G. (2016). Organization and function of the 3D genome. Nat Rev Genet *17*, 661-678.

Booth, C. A. G., Barkas, N., Neo, W. H., Boukarabila, H., Soilleux, E. J., Giotopoulos, G., Farnoud, N., Giustacchini, A., Ashley, N., Carrelha, J.*, et al.* (2018). Ezh2 and Runx1 Mutations Collaborate to Initiate Lympho-Myeloid Leukemia in Early Thymic Progenitors. Cancer Cell *33*, 274-291.e278.

Bouvy-Liivrand, M., Hernandez de Sande, A., Polonen, P., Mehtonen, J., Vuorenmaa, T., Niskanen, H., Sinkkonen, L., Kaikkonen, M. U., and Heinaniemi, M. (2017). Analysis of primary microRNA loci from nascent transcriptomes reveals regulatory domains governed by chromatin architecture. Nucleic Acids Res *45*, 9837-9849.

Bradner, J. E., Hnisz, D., and Young, R. A. (2017). Transcriptional Addiction in Cancer. Cell *168*, 629-643.

Bray, F., Laversanne, M., Weiderpass, E., and Soerjomataram, I. (2021). The ever-increasing importance of cancer as a leading cause of premature death worldwide. Cancer *127*, 3029-3030.

Bredthauer, C., Fischer, A., Ahari, A. J., Cao, X., Weber, J., Rad, L., Rad, R., Wachutka, L., and Gagneur, J. (2023). Transmicron: accurate prediction of insertion probabilities improves detection of cancer driver genes from transposon mutagenesis screens. Nucleic Acids Res *51*, e21.

Brett, B. T., Berquam-Vrieze, K. E., Nannapaneni, K., Huang, J., Scheetz, T. E., and Dupuy, A. J. (2011). Novel molecular and computational methods improve the accuracy of insertion site analysis in Sleeping Beauty-induced tumors. PLoS One *6*, e24668.

Brown, F. C., Still, E., Koche, R. P., Yim, C. Y., Takao, S., Cifani, P., Reed, C., Gunasekera, S., Ficarro, S. B., Romanienko, P.*, et al.* (2018). MEF2C Phosphorylation Is Required for Chemotherapy Resistance in Acute Myeloid Leukemia. Cancer Discov *8*, 478-497.

Brown, L., Cheng, J. T., Chen, Q., Siciliano, M. J., Crist, W., Buchanan, G., and Baer, R. (1990). Site-specific recombination of the tal-1 gene is a common occurrence in human T cell leukemia. Embo j *9*, 3343-3351.

Burns, K. H. (2017). Transposable elements in cancer. Nat Rev Cancer *17*, 415-424.

Calin, G. A., Ferracin, M., Cimmino, A., Di Leva, G., Shimizu, M., Wojcik, S. E., Iorio, M. V., Visone, R., Sever, N. I., Fabbri, M.*, et al.* (2005). A MicroRNA signature associated with prognosis and progression in chronic lymphocytic leukemia. N Engl J Med *353*, 1793-1801.

Canté-Barrett, K., Pieters, R., and Meijerink, J. P. (2014). Myocyte enhancer factor 2C in hematopoiesis and leukemia. Oncogene *33*, 403-410.

Chorev, M., and Carmel, L. (2012). The function of introns. Front Genet *3*, 55.

Chorev, M., Joseph Bekker, A., Goldberger, J., and Carmel, L. (2017). Identification of introns harboring functional sequence elements through positional conservation. Sci Rep *7*, 4201.

Cinghu, S., Yang, P., Kosak, J. P., Conway, A. E., Kumar, D., Oldfield, A. J., Adelman, K., and Jothi, R. (2017). Intragenic Enhancers Attenuate Host Gene Expression. Mol Cell *68*, 104-117.e106.

Collier, L. S., Carlson, C. M., Ravimohan, S., Dupuy, A. J., and Largaespada, D. A. (2005). Cancer gene discovery in solid tumours using transposon-based somatic mutagenesis in the mouse. Nature *436*, 272-276.

Colomer-Lahiguera, S., Pisecker, M., König, M., Nebral, K., Pickl, W. F., Kauer, M. O., Haas, O. A., Ullmann, R., Attarbaschi, A., Dworzak, M. N., and Strehl, S. (2017). MEF2C-dysregulated pediatric T-cell acute lymphoblastic leukemia is associated with CDKN1B deletions and a poor response to glucocorticoid therapy. Leuk Lymphoma *58*, 2895-2904.

Concordet, J. P., and Haeussler, M. (2018). CRISPOR: intuitive guide selection for CRISPR/Cas9 genome editing experiments and screens. Nucleic Acids Res *46*, W242-w245.

Core, L. J., Waterfall, J. J., and Lis, J. T. (2008). Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science *322*, 1845-1848.

Coustan-Smith, E., Mullighan, C. G., Onciu, M., Behm, F. G., Raimondi, S. C., Pei, D., Cheng, C., Su, X., Rubnitz, J. E., Basso, G.*, et al.* (2009). Early T-cell precursor leukaemia: a subtype of very high-risk acute lymphoblastic leukaemia. Lancet Oncol *10*, 147-156.

Davies, J. O., Oudelaar, A. M., Higgs, D. R., and Hughes, J. R. (2017). How best to identify chromosomal interactions: a comparison of approaches. Nat Methods *14*, 125-134.

De Bie, J., Demeyer, S., Alberti-Servera, L., Geerdens, E., Segers, H., Broux, M., De Keersmaecker, K., Michaux, L., Vandenberghe, P., Voet, T.*, et al.* (2018). Single-cell sequencing reveals the origin and the order of mutation acquisition in T-cell acute lymphoblastic leukemia. Leukemia *32*, 1358-1369.

de Jong, J., Akhtar, W., Badhai, J., Rust, A. G., Rad, R., Hilkens, J., Berns, A., van Lohuizen, M., Wessels, L. F., and de Ridder, J. (2014). Chromatin landscapes of retroviral and transposon integration profiles. PLoS Genet *10*, e1004250.

de Ridder, J., Uren, A., Kool, J., Reinders, M., and Wessels, L. (2006). Detecting statistically significant common insertion sites in retroviral insertional mutagenesis screens. PLoS Comput Biol *2*, e166.

DeKoter, R. P., Geadah, M., Khoosal, S., Xu, L. S., Thillainadesan, G., Torchia, J., Chin, S. S., and Garrett-Sinha, L. A. (2010). Regulation of follicular B cell differentiation by the related E26 transformation-specific transcription factors PU.1, Spi-B, and Spi-C. J Immunol *185*, 7374-7384.

Ding, S., Wu, X., Li, G., Han, M., Zhuang, Y., and Xu, T. (2005). Efficient transposition of the piggyBac (PB) transposon in mammalian cells and mice. Cell *122*, 473-483.

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.

Doni Jayavelu, N., Jajodia, A., Mishra, A., and Hawkins, R. D. (2020). Candidate silencer elements for the human and mouse genomes. Nat Commun *11*, 1061.

Dupuy, A. J., Akagi, K., Largaespada, D. A., Copeland, N. G., and Jenkins, N. A. (2005). Mammalian mutagenesis using a highly mobile somatic Sleeping Beauty transposon system. Nature *436*, 221-226.

Dupuy, A. J., Fritz, S., and Largaespada, D. A. (2001). Transposition and gene disruption in the male germline of the mouse. Genesis *30*, 82-88.

Dupuy, A. J., Rogers, L. M., Kim, J., Nannapaneni, K., Starr, T. K., Liu, P., Largaespada, D. A., Scheetz, T. E., Jenkins, N. A., and Copeland, N. G. (2009). A modified sleeping beauty transposon system that can be used to model a wide variety of human cancers in mice. Cancer Res *69*, 8150-8156.

Elkon, R., and Agami, R. (2017). Characterization of noncoding regulatory DNA in the human genome. Nat Biotechnol *35*, 732-746.

Elliott, K., and Larsson, E. (2021). Non-coding driver mutations in human cancer. Nat Rev Cancer.

Ernst, J., and Kellis, M. (2012). ChromHMM: automating chromatin-state discovery and characterization. Nat Methods *9*, 215-216.

Fischer, A., Lersch, R., de Andrade Krätzig, N., Strong, A., Friedrich, M. J., Weber, J., Engleitner, T., Öllinger, R., Yen, H. Y., Kohlhofer, U.*, et al.* (2023). In vivo interrogation of

regulatory genomes reveals extensive quasi-insufficiency in cancer evolution. Cell Genom *3*, 100276.

Friedrich, M. J., Rad, L., Bronner, I. F., Strong, A., Wang, W., Weber, J., Mayho, M., Ponstingl, H., Engleitner, T., Grove, C.*, et al.* (2017). Genome-wide transposon screening and quantitative insertion site sequencing for cancer gene discovery in mice. Nat Protoc *12*, 289-309.

Gallagher, M. D., and Chen-Plotkin, A. S. (2018). The Post-GWAS Era: From Association to Function. Am J Hum Genet *102*, 717-730.

Garitano-Trojaola, A., José-Enériz, E. S., Ezponda, T., Unfried, J. P., Carrasco-León, A., Razquin, N., Barriocanal, M., Vilas-Zornoza, A., Sangro, B., Segura, V.*, et al.* (2018). Deregulation of linc-PINT in acute lymphoblastic leukemia is implicated in abnormal proliferation of leukemic cells. Oncotarget *9*, 12842-12852.

Gökbuget, N., Stanze, D., Beck, J., Diedrich, H., Horst, H. A., Hüttmann, A., Kobbe, G., Kreuzer, K. A., Leimer, L., Reichle, A.*, et al.* (2012). Outcome of relapsed adult lymphoblastic leukemia depends on response to salvage chemotherapy, prognostic factors, and performance of stem cell transplantation. Blood *120*, 2032-2041.

Gordon, C. T., and Lyonnet, S. (2014). Enhancer mutations and phenotype modularity. Nat Genet *46*, 3-4.

Granja, J. M., Klemm, S., McGinnis, L. M., Kathiria, A. S., Mezger, A., Corces, M. R., Parks, B., Gars, E., Liedtke, M., Zheng, G. X. Y.*, et al.* (2019). Single-cell multiomic analysis identifies regulatory programs in mixed-phenotype acute leukemia. Nat Biotechnol *37*, 1458-1465.

Graur, D. (2017). An Upper Limit on the Functional Fraction of the Human Genome. Genome Biol Evol *9*, 1880-1885.

Greaves, M., and Maley, C. C. (2012). Clonal evolution in cancer. Nature *481*, 306-313.

Gu, Z., Schlesner, M., and Hübschmann, D. (2021). cola: an R/Bioconductor package for consensus partitioning through a general framework. Nucleic Acids Res *49*, e15.

Haldar, M., Kohyama, M., So, A. Y., Kc, W., Wu, X., Briseño, C. G., Satpathy, A. T., Kretzer, N. M., Arase, H., Rajasekaran, N. S.*, et al.* (2014). Heme-mediated SPI-C induction promotes monocyte differentiation into iron-recycling macrophages. Cell *156*, 1223-1234.

Han, Y. C., Park, C. Y., Bhagat, G., Zhang, J., Wang, Y., Fan, J. B., Liu, M., Zou, Y., Weissman, I. L., and Gu, H. (2010). microRNA-29a induces aberrant self-renewal capacity in hematopoietic progenitors, biased myeloid development, and acute myeloid leukemia. J Exp Med *207*, 475-489.

Hanahan, D. (2022). Hallmarks of Cancer: New Dimensions. Cancer Discov *12*, 31-46.

Hanahan, D., and Weinberg, R. A. (2000). The hallmarks of cancer. Cell *100*, 57-70.

Hanahan, D., and Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. Cell *144*, 646-674.

Haydu, J. E., and Ferrando, A. A. (2013). Early T-cell precursor acute lymphoblastic leukaemia. Curr Opin Hematol *20*, 369-373.

Hnisz, D., Weintraub, A. S., Day, D. S., Valton, A. L., Bak, R. O., Li, C. H., Goldmann, J., Lajoie, B. R., Fan, Z. P., Sigova, A. A.*, et al.* (2016). Activation of proto-oncogenes by disruption of chromosome neighborhoods. Science *351*, 1454-1458.

Ho, L., and Crabtree, G. R. (2010). Chromatin remodelling during development. Nature *463*, 474-484.

Homminga, I., Pieters, R., Langerak, A. W., de Rooi, J. J., Stubbs, A., Verstegen, M., Vuerhard, M., Buijs-Gladdines, J., Kooi, C., Klous, P.*, et al.* (2011). Integrated transcript and genome

analyses reveal NKX2-1 and MEF2C as potential oncogenes in T cell acute lymphoblastic leukemia. Cancer Cell *19*, 484-497.

Hosokawa, H., and Rothenberg, E. V. (2021). How transcription factors drive choice of the T cell fate. Nat Rev Immunol *21*, 162-176.

Hu, G., Cui, K., Fang, D., Hirose, S., Wang, X., Wangsa, D., Jin, W., Ried, T., Liu, P., Zhu, J.*, et al.* (2018). Transformation of Accessible Chromatin and 3D Nucleome Underlies Lineage Commitment of Early T Cells. Immunity *48*, 227-242 e228.

Hu, S., Qian, M., Zhang, H., Guo, Y., Yang, J., Zhao, X., He, H., Lu, J., Pan, J., Chang, M.*, et al.* (2017). Whole-genome noncoding sequence analysis in T-cell acute lymphoblastic leukemia identifies oncogene enhancer mutations. Blood *129*, 3264-3268.

Inaba, H., Greaves, M., and Mullighan, C. G. (2013). Acute lymphoblastic leukaemia. Lancet *381*, 1943-1955.

Ing-Simmons, E., Seitan, V. C., Faure, A. J., Flicek, P., Carroll, T., Dekker, J., Fisher, A. G., Lenhard, B., and Merkenschlager, M. (2015). Spatial enhancer clustering and regulation of enhancer-proximal genes by cohesin. Genome Res *25*, 504-513.

Ivics, Z., Hackett, P. B., Plasterk, R. H., and Izsvák, Z. (1997). Molecular reconstruction of Sleeping Beauty, a Tc1-like transposon from fish, and its transposition in human cells. Cell *91*, 501-510.

Jain, N., Lamb, A. V., O'Brien, S., Ravandi, F., Konopleva, M., Jabbour, E., Zuo, Z., Jorgensen, J., Lin, P., Pierce, S.*, et al.* (2016). Early T-cell precursor acute lymphoblastic leukemia/lymphoma (ETP-ALL/LBL) in adolescents and adults: a high-risk subtype. Blood *127*, 1863-1869.

Johnson, J. L., Georgakilas, G., Petrovic, J., Kurachi, M., Cai, S., Harly, C., Pear, W. S., Bhandoola, A., Wherry, E. J., and Vahedi, G. (2018). Lineage-Determining Transcription Factor TCF-1 Initiates the Epigenetic Identity of T Cells. Immunity *48*, 243-257.e210.

Jurka, J., Kapitonov, V. V., Pavlicek, A., Klonowski, P., Kohany, O., and Walichiewicz, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res *110*, 462-467.

Kaikkonen, M. U., Spann, N. J., Heinz, S., Romanoski, C. E., Allison, K. A., Stender, J. D., Chun, H. B., Tough, D. F., Prinjha, R. K., Benner, C., and Glass, C. K. (2013). Remodeling of the enhancer landscape during macrophage activation is coupled to enhancer transcription. Mol Cell *51*, 310-325.

Kapranov, P., Cheng, J., Dike, S., Nix, D. A., Duttagupta, R., Willingham, A. T., Stadler, P. F., Hertel, J., Hackermüller, J., Hofacker, I. L.*, et al.* (2007). RNA maps reveal new RNA classes and a possible function for pervasive transcription. Science *316*, 1484-1488.

Khurana, E., Fu, Y., Chakravarty, D., Demichelis, F., Rubin, M. A., and Gerstein, M. (2016). Role of non-coding sequence variants in cancer. Nat Rev Genet *17*, 93-108.

Kimura, H. (2013). Histone modifications for human epigenome analysis. J Hum Genet *58*, 439-445.

Klijn, C., Koudijs, M. J., Kool, J., ten Hoeve, J., Boer, M., de Moes, J., Akhtar, W., van Miltenburg, M., Vendel-Zwaagstra, A., Reinders, M. J.*, et al.* (2013). Analysis of tumor heterogeneity and cancer gene networks using deep sequencing of MMTV-induced mouse mammary tumors. PLoS One *8*, e62113.

Kloetgen, A., Thandapani, P., Ntziachristos, P., Ghebrechristos, Y., Nomikou, S., Lazaris, C., Chen, X., Hu, H., Bakogianni, S., Wang, J.*, et al.* (2020). Three-dimensional chromatin landscapes in T cell acute lymphoblastic leukemia. Nat Genet *52*, 388-400.

Kohyama, M., Ise, W., Edelson, B. T., Wilker, P. R., Hildner, K., Mejia, C., Frazier, W. A., Murphy, T. L., and Murphy, K. M. (2009). Role for Spi-C in the development of red pulp macrophages and splenic iron homeostasis. Nature *457*, 318-321.

Kopp, F., and Mendell, J. T. (2018). Functional Classification and Experimental Dissection of Long Noncoding RNAs. Cell *172*, 393-407.

Koudijs, M. J., Klijn, C., van der Weyden, L., Kool, J., ten Hoeve, J., Sie, D., Prasetyanti, P. R., Schut, E., Kas, S., Whipp, T.*, et al.* (2011). High-throughput semiquantitative analysis of insertional mutations in heterogeneous tumors. Genome Res *21*, 2181-2189.

Laiosa, C. V., Stadtfeld, M., and Graf, T. (2006). Determinants of lymphoid-myeloid lineage diversification. Annu Rev Immunol *24*, 705-738.

Lambert, S. A., Jolma, A., Campitelli, L. F., Das, P. K., Yin, Y., Albu, M., Chen, X., Taipale, J., Hughes, T. R., and Weirauch, M. T. (2018). The Human Transcription Factors. Cell *172*, 650-665.

Lang, F., Wojcik, B., and Rieger, M. A. (2015). Stem Cell Hierarchy and Clonal Evolution in Acute Lymphoblastic Leukemia. Stem Cells Int *2015*, 137164.

Lara-Astiaso, D., Weiner, A., Lorenzo-Vivas, E., Zaretsky, I., Jaitin, D. A., David, E., Keren-Shaul, H., Mildner, A., Winter, D., Jung, S.*, et al.* (2014). Immunogenetics. Chromatin state dynamics during blood formation. Science *345*, 943-949.

Laramée, A. S., Raczkowski, H., Shao, P., Batista, C., Shukla, D., Xu, L., Haeryfar, S. M. M., Tesfagiorgis, Y., Kerfoot, S., and DeKoter, R. (2020). Opposing Roles for the Related ETS-Family Transcription Factors Spi-B and Spi-C in Regulating B Cell Differentiation and Function. Front Immunol *11*, 841.

Laurenti, E., Doulatov, S., Zandi, S., Plumb, I., Chen, J., April, C., Fan, J. B., and Dick, J. E. (2013). The transcriptional architecture of early human hematopoiesis identifies multilevel control of lymphoid commitment. Nat Immunol *14*, 756-763.

Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M. T., and Carey, V. J. (2013). Software for computing and annotating genomic ranges. PLoS Comput Biol *9*, e1003118.

Lawrence, M. S., Stojanov, P., Mermel, C. H., Robinson, J. T., Garraway, L. A., Golub, T. R., Meyerson, M., Gabriel, S. B., Lander, E. S., and Getz, G. (2014). Discovery and saturation analysis of cancer genes across 21 tumour types. Nature *505*, 495-501.

Lee, J. T. (2012). Epigenetic regulation by long noncoding RNAs. Science *338*, 1435-1439.

Lee, T. I., and Young, R. A. (2013). Transcriptional regulation and its misregulation in disease. Cell *152*, 1237-1251.

Li, M. A., Pettitt, S. J., Eckert, S., Ning, Z., Rice, S., Cadinanos, J., Yusa, K., Conte, N., and Bradley, A. (2013). The piggyBac transposon displays local and distant reintegration preferences and can cause mutations at noncanonical integration sites. Mol Cell Biol *33*, 1317-1330.

Li, S. K., Solomon, L. A., Fulkerson, P. C., and DeKoter, R. P. (2015). Identification of a negative regulatory role for spi-C in the murine B cell lineage. J Immunol *194*, 3798-3807.

Li, W., Notani, D., and Rosenfeld, M. G. (2016). Enhancers as non-coding RNA transcription units: recent insights and future perspectives. Nat Rev Genet *17*, 207-223.

Lin, K. H., Xie, A., Rutter, J. C., Ahn, Y. R., Lloyd-Cowden, J. M., Nichols, A. G., Soderquist, R. S., Koves, T. R., Muoio, D. M., MacIver, N. J.*, et al.* (2019). Systematic Dissection of the Metabolic-Apoptotic Interface in AML Reveals Heme Biosynthesis to Be a Regulator of Drug Sensitivity. Cell Metab *29*, 1217-1231.e1217.

Liu, Y., Easton, J., Shao, Y., Maciaszek, J., Wang, Z., Wilkinson, M. R., McCastlain, K., Edmonson, M., Pounds, S. B., Shi, L.*, et al.* (2017). The genomic landscape of pediatric and young adult T-lineage acute lymphoblastic leukemia. Nat Genet *49*, 1211-1218.

Long, H. K., Prescott, S. L., and Wysocka, J. (2016). Ever-Changing Landscapes: Transcriptional Enhancers in Development and Evolution. Cell *167*, 1170-1187.

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.

Lucic, B., Chen, H. C., Kuzman, M., Zorita, E., Wegner, J., Minneker, V., Wang, W., Fronza, R., Laufs, S., Schmidt, M.*, et al.* (2019). Spatially clustered loci with multiple enhancers are frequent targets of HIV-1 integration. Nat Commun *10*, 4059.

Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A. R., Kamitaki, N., Martersteck, E. M.*, et al.* (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell *161*, 1202-1214.

Mattick, J. S., and Rinn, J. L. (2015). Discovery and annotation of long noncoding RNAs. Nat Struct Mol Biol *22*, 5-7.

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.

McClintock, B. (1950). The origin and behavior of mutable loci in maize. Proc Natl Acad Sci U S A *36*, 344-355.

Miano, V., Codino, A., Pandolfini, L., and Barbieri, I. (2021). The non-coding epitranscriptome in cancer. Brief Funct Genomics.

Miskey, C., Kesselring, L., Querques, I., Abrusán, G., Barabas, O., and Ivics, Z. (2022). Engineered Sleeping Beauty transposase redirects transposon integration away from genes. Nucleic Acids Res *50*, 2807-2825.

Montalbano, A., Canver, M. C., and Sanjana, N. E. (2017). High-Throughput Approaches to Pinpoint Function within the Noncoding Genome. Mol Cell *68*, 44-59.

Moreau-Gachelin, F., Wendling, F., Molina, T., Denis, N., Titeux, M., Grimber, G., Briand, P., Vainchenker, W., and Tavitian, A. (1996). Spi-1/PU.1 transgenic mice develop multistep erythroleukemias. Mol Cell Biol *16*, 2453-2463.

Morse, H. C., 3rd, Anver, M. R., Fredrickson, T. N., Haines, D. C., Harris, A. W., Harris, N. L., Jaffe, E. S., Kogan, S. C., MacLennan, I. C., Pattengale, P. K., and Ward, J. M. (2002). Bethesda proposals for classification of lymphoid neoplasms in mice. Blood *100*, 246-258.

Muerdter, F., Boryń Ł, M., Woodfin, A. R., Neumayr, C., Rath, M., Zabidi, M. A., Pagani, M., Haberle, V., Kazmar, T., Catarino, R. R.*, et al.* (2018). Resolving systematic errors in widely used enhancer activity assays in human cells. Nat Methods *15*, 141-149.

Nagel, S., Meyer, C., Quentmeier, H., Kaufmann, M., Drexler, H. G., and MacLeod, R. A. (2008). MEF2C is activated by multiple mechanisms in a subset of T-acute lymphoblastic leukemia cell lines. Leukemia *22*, 600-607.

Neumann, M., Coskun, E., Fransecky, L., Mochmann, L. H., Bartram, I., Sartangi, N. F., Heesch, S., Gökbuget, N., Schwartz, S., Brandts, C.*, et al.* (2013). FLT3 mutations in early T-cell precursor ALL characterize a stem cell like leukemia and imply the clinical use of tyrosine kinase inhibitors. PLoS One *8*, e53190.

Newberg, J. Y., Black, M. A., Jenkins, N. A., Copeland, N. G., Mann, K. M., and Mann, M. B. (2018). SB Driver Analysis: a Sleeping Beauty cancer driver analysis framework for identifying and prioritizing experimentally actionable oncogenes and tumor suppressors. Nucleic Acids Res *46*, e94.

Novershtern, N., Subramanian, A., Lawton, L. N., Mak, R. H., Haining, W. N., McConkey, M. E., Habib, N., Yosef, N., Chang, C. Y., Shay, T.*, et al.* (2011). Densely interconnected transcriptional circuits control cell states in human hematopoiesis. Cell *144*, 296-309.

Nowell, P. C. (1976). The clonal evolution of tumor cell populations. Science *194*, 23-28.

Oliveira, L. H., Schiavinato, J. L., Fráguas, M. S., Lucena-Araujo, A. R., Haddad, R., Araújo, A. G., Dalmazzo, L. F., Rego, E. M., Covas, D. T., Zago, M. A., and Panepucci, R. A. (2015). Potential roles of microRNA-29a in the molecular pathophysiology of T-cell acute lymphoblastic leukemia. Cancer Sci *106*, 1264-1277.

Papaemmanuil, E., Gerstung, M., Bullinger, L., Gaidzik, V. I., Paschka, P., Roberts, N. D., Potter, N. E., Heuser, M., Thol, F., Bolli, N.*, et al.* (2016). Genomic Classification and Prognosis in Acute Myeloid Leukemia. N Engl J Med *374*, 2209-2221.

Parekh, S., Ziegenhain, C., Vieth, B., Enard, W., and Hellmann, I. (2016). The impact of amplification on differential expression analyses by RNA-seq. Sci Rep *6*, 25533.

Patel, J. L., Smith, L. M., Anderson, J., Abromowitch, M., Campana, D., Jacobsen, J., Lones, M. A., Gross, T. G., Cairo, M. S., and Perkins, S. L. (2012). The immunophenotype of T-lymphoblastic lymphoma in children and adolescents: a Children's Oncology Group report. Br J Haematol *159*, 454-461.

Petrovic, J., Zhou, Y., Fasolino, M., Goldman, N., Schwartz, G. W., Mumbach, M. R., Nguyen, S. C., Rome, K. S., Sela, Y., Zapataro, Z.*, et al.* (2019). Oncogenic Notch Promotes Long-Range Regulatory Interactions within Hyperconnected 3D Cliques. Mol Cell *73*, 1174-1190.e1112.

Pott, S., and Lieb, J. D. (2015). What are super-enhancers? Nat Genet *47*, 8-12.

Pui, C. H., Robison, L. L., and Look, A. T. (2008). Acute lymphoblastic leukaemia. Lancet *371*, 1030-1043.

Qian, M., Zhao, X., Devidas, M., Yang, W., Gocho, Y., Smith, C., Gastier-Foster, J. M., Li, Y., Xu, H., Zhang, S.*, et al.* (2019). Genome-Wide Association Study of Susceptibility Loci for T-Cell Acute Lymphoblastic Leukemia in Children. J Natl Cancer Inst *111*, 1350-1357.

Quinn, J. J., and Chang, H. Y. (2016). Unique features of long non-coding RNA biogenesis and function. Nat Rev Genet *17*, 47-62.

Rad, R., Rad, L., Wang, W., Cadinanos, J., Vassiliou, G., Rice, S., Campos, L. S., Yusa, K., Banerjee, R., Li, M. A.*, et al.* (2010). PiggyBac transposon mutagenesis: a tool for cancer gene discovery in mice. Science *330*, 1104-1107.

Rad, R., Rad, L., Wang, W., Strong, A., Ponstingl, H., Bronner, I. F., Mayho, M., Steiger, K., Weber, J., Hieber, M.*, et al.* (2015). A conditional piggyBac transposition system for genetic screening in mice identifies oncogenic networks in pancreatic cancer. Nat Genet *47*, 47-56.

Ramírez, F., Ryan, D. P., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A. S., Heyne, S., Dündar, F., and Manke, T. (2016). deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Res *44*, W160-165.

Ren, G., Jin, W., Cui, K., Rodrigez, J., Hu, G., Zhang, Z., Larson, D. R., and Zhao, K. (2017). CTCF-Mediated Enhancer-Promoter Interaction Is a Critical Regulator of Cell-to-Cell Variation of Gene Expression. Mol Cell *67*, 1049-1058.e1046.

Rheinbay, E., Nielsen, M. M., Abascal, F., Wala, J. A., Shapira, O., Tiao, G., Hornshøj, H., Hess, J. M., Juul, R. I., Lin, Z.*, et al.* (2020). Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. Nature *578*, 102-111.

Ribera, J. M., Ortega, J. J., Oriol, A., Bastida, P., Calvo, C., Pérez-Hurtado, J. M., González-Valentín, M. E., Martín-Reina, V., Molinés, A., Ortega-Rivas, F.*, et al.* (2007). Comparison of intensive chemotherapy, allogeneic, or autologous stem-cell transplantation as postremission

treatment for children with very high risk acute lymphoblastic leukemia: PETHEMA ALL-93 Trial. J Clin Oncol *25*, 16-24.

Roels, J., Thénoz, M., Szarzyńska, B., Landfors, M., De Coninck, S., Demoen, L., Provez, L., Kuchmiy, A., Strubbe, S., Reunes, L.*, et al.* (2020). Aging of preleukemic thymocytes drives CpG island hypermethylation in T-cell acute lymphoblastic leukemia. Blood Cancer Discov *1*, 274-289.

Rothenberg, E. V. (2019). Programming for T-lymphocyte fates: modularity and mechanisms. Genes Dev *33*, 1117-1135.

Rothenberg, E. V., Hosokawa, H., and Ungerbäck, J. (2019). Mechanisms of Action of Hematopoietic Transcription Factor PU.1 in Initiation of T-Cell Development. Front Immunol *10*, 228.

Rothenberg, E. V., Moore, J. E., and Yui, M. A. (2008). Launching the T-cell-lineage developmental programme. Nat Rev Immunol *8*, 9-21.

Sanda, T., Lawton, L. N., Barrasa, M. I., Fan, Z. P., Kohlhammer, H., Gutierrez, A., Ma, W., Tatarek, J., Ahn, Y., Kelliher, M. A.*, et al.* (2012). Core transcriptional regulatory circuit controlled by the TAL1 complex in human T cell acute lymphoblastic leukemia. Cancer Cell *22*, 209-221.

Sandoval-Villegas, N., Nurieva, W., Amberger, M., and Ivics, Z. (2021). Contemporary Transposon Tools: A Review and Guide through Mechanisms and Applications of Sleeping Beauty, piggyBac and Tol2 for Genome Engineering. Int J Mol Sci *22*.

Sarver, A. L., Erdman, J., Starr, T., Largaespada, D. A., and Silverstein, K. A. (2012). TAPDANCE: an automated tool to identify and annotate transposon insertion CISs and associations between CISs from next generation sequence data. BMC Bioinformatics *13*, 154.

Schlenner, S. M., and Rodewald, H. R. (2010). Early T cell development and the pitfalls of potential. Trends Immunol *31*, 303-310.

Seki, M., Kimura, S., Isobe, T., Yoshida, K., Ueno, H., Nakajima-Takagi, Y., Wang, C., Lin, L., Kon, A., Suzuki, H.*, et al.* (2017). Recurrent SPI1 (PU.1) fusions in high-risk pediatric T cell acute lymphoblastic leukemia. Nat Genet *49*, 1274-1281.

Shih, H. Y., Sciumè, G., Mikami, Y., Guo, L., Sun, H. W., Brooks, S. R., Urban, J. F., Jr., Davis, F. P., Kanno, Y., and O'Shea, J. J. (2016). Developmental Acquisition of Regulomes Underlies Innate Lymphoid Cell Functionality. Cell *165*, 1120-1133.

Shlyueva, D., Stampfel, G., and Stark, A. (2014). Transcriptional enhancers: from properties to genome-wide predictions. Nat Rev Genet *15*, 272-286.

Sidoli, S., Lopes, M., Lund, P. J., Goldman, N., Fasolino, M., Coradin, M., Kulej, K., Bhanu, N. V., Vahedi, G., and Garcia, B. A. (2019). A mass spectrometry-based assay using metabolic labeling to rapidly monitor chromatin accessibility of modified histone proteins. Sci Rep *9*, 13613.

Soodgupta, D., White, L. S., Yang, W., Johnston, R., Andrews, J. M., Kohyama, M., Murphy, K. M., Mosammaparast, N., Payton, J. E., and Bednarski, J. J. (2019). RAG-Mediated DNA Breaks Attenuate PU.1 Activity in Early B Cells through Activation of a SPIC-BCLAF1 Complex. Cell Rep *29*, 829-843.e825.

Stratton, M. R., Campbell, P. J., and Futreal, P. A. (2009). The cancer genome. Nature *458*, 719-724.

Su, W., Xu, M., Chen, X., Chen, N., Gong, J., Nie, L., Li, L., Li, X., Zhang, M., and Zhou, Q. (2017). Long noncoding RNA ZEB1-AS1 epigenetically regulates the expressions of ZEB1 and downstream molecules in prostate cancer. Mol Cancer *16*, 142.

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., and Bray, F. (2021). Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin *71*, 209-249.

Szczepański, T., Langerak, A. W., Willemse, M. J., Wolvers-Tettero, I. L., van Wering, E. R., and van Dongen, J. J. (2000). T cell receptor gamma (TCRG) gene rearrangements in T cell acute lymphoblastic leukemia refelct 'end-stage' recombinations: implications for minimal residual disease monitoring. Leukemia *14*, 1208-1214.

Tan, S. H., Bertulfo, F. C., and Sanda, T. (2017). Leukemia-Initiating Cells in T-Cell Acute Lymphoblastic Leukemia. Front Oncol *7*, 218.

Terwilliger, T., and Abdul-Hay, M. (2017). Acute lymphoblastic leukemia: a comprehensive review and 2017 update. Blood Cancer J *7*, e577.

Thakore, P. I., D'Ippolito, A. M., Song, L., Safi, A., Shivakumar, N. K., Kabadi, A. M., Reddy, T. E., Crawford, G. E., and Gersbach, C. A. (2015). Highly specific epigenome editing by CRISPR-Cas9 repressors for silencing of distal regulatory elements. Nat Methods *12*, 1143-1149.

Thibault, S. T., Singer, M. A., Miyazaki, W. Y., Milash, B., Dompe, N. A., Singh, C. M., Buchholz, R., Demsky, M., Fawcett, R., Francis-Lang, H. L.*, et al.* (2004). A complementary transposon tool kit for Drosophila melanogaster using P and piggyBac. Nat Genet *36*, 283-287.

Tottone, L., Lancho, O., Loh, J. W., Singh, A., Kimura, S., Roels, J., Kuchmiy, A., Strubbe, S., Lawlor, M. A., da Silva-Diz, V.*, et al.* (2021). A Tumor Suppressor Enhancer of PTEN in T-cell development and leukemia. Blood Cancer Discov *2*, 92-109.

Ungerbäck, J., Hosokawa, H., Wang, X., Strid, T., Williams, B. A., Sigvardsson, M., and Rothenberg, E. V. (2018). Pioneering, chromatin remodeling, and epigenetic constraint in early T-cell gene regulation by SPI1 (PU.1). Genome Res *28*, 1508-1519.

Vadillo, E., Dorantes-Acosta, E., Pelayo, R., and Schnoor, M. (2018). T cell acute lymphoblastic leukemia (T-ALL): New insights into the cellular origins and infiltration mechanisms common and unique among hematologic malignancies. Blood Rev *32*, 36-51.

Visvader, J. E. (2011). Cells of origin in cancer. Nature *469*, 314-322.

Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz, L. A., Jr., and Kinzler, K. W. (2013). Cancer genome landscapes. Science *339*, 1546-1558.

Volinia, S., Calin, G. A., Liu, C. G., Ambs, S., Cimmino, A., Petrocca, F., Visone, R., Iorio, M., Roldo, C., Ferracin, M.*, et al.* (2006). A microRNA expression signature of human solid tumors defines cancer gene targets. Proc Natl Acad Sci U S A *103*, 2257-2261.

Wada, H., Masuda, K., Satoh, R., Kakugawa, K., Ikawa, T., Katsura, Y., and Kawamoto, H. (2008). Adult T-cell progenitors retain myeloid potential. Nature *452*, 768-772.

Wang, W., Lin, C., Lu, D., Ning, Z., Cox, T., Melvin, D., Wang, X., Bradley, A., and Liu, P. (2008). Chromosomal transposition of PiggyBac in mouse embryonic stem cells. Proc Natl Acad Sci U S A *105*, 9290-9295.

Weber, J., Braun, C. J., Saur, D., and Rad, R. (2020). In vivo functional screening for systems-level integrative cancer genomics. Nat Rev Cancer *20*, 573-593.

Weber, J., de la Rosa, J., Grove, C. S., Schick, M., Rad, L., Baranov, O., Strong, A., Pfaus, A., Friedrich, M. J., Engleitner, T.*, et al.* (2019). PiggyBac transposon tools for recessive screening identify B-cell lymphoma drivers in mice. Nat Commun *10*, 1415.

Wei, G., Abraham, B. J., Yagi, R., Jothi, R., Cui, K., Sharma, S., Narlikar, L., Northrup, D. L., Tang, Q., Paul, W. E.*, et al.* (2011). Genome-wide analyses of transcription factor GATA3-mediated gene regulation in distinct T cell types. Immunity *35*, 299-311.

Weinhold, N., Jacobsen, A., Schultz, N., Sander, C., and Lee, W. (2014). Genome-wide analysis of noncoding regulatory mutations in cancer. Nat Genet *46*, 1160-1165.

Welner, R. S., Esplin, B. L., Garrett, K. P., Pelayo, R., Luche, H., Fehling, H. J., and Kincade, P. W. (2009). Asynchronous RAG-1 expression during B lymphopoiesis. J Immunol *183*, 7768-7777.

Weng, A. P., Ferrando, A. A., Lee, W., Morris, J. P. t., Silverman, L. B., Sanchez-Irizarry, C., Blacklow, S. C., Look, A. T., and Aster, J. C. (2004). Activating mutations of NOTCH1 in human T cell acute lymphoblastic leukemia. Science *306*, 269-271.

Wenzinger, C., Williams, E., and Gru, A. A. (2018). Updates in the Pathology of Precursor Lymphoid Neoplasms in the Revised Fourth Edition of the WHO Classification of Tumors of Hematopoietic and Lymphoid Tissues. Curr Hematol Malig Rep *13*, 275-288.

Whyte, W. A., Orlando, D. A., Hnisz, D., Abraham, B. J., Lin, C. Y., Kagey, M. H., Rahl, P. B., Lee, T. I., and Young, R. A. (2013). Master transcription factors and mediator establish super-enhancers at key cell identity genes. Cell *153*, 307-319.

Wu, H., Liu, X., and Jaenisch, R. (1994). Double replacement: strategy for efficient introduction of subtle mutations into the murine Col1a-1 gene by homologous recombination in embryonic stem cells. Proc Natl Acad Sci U S A *91*, 2819-2823.

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.

Yoshida, H., Lareau, C. A., Ramirez, R. N., Rose, S. A., Maier, B., Wroblewska, A., Desland, F., Chudnovskiy, A., Mortha, A., Dominguez, C.*, et al.* (2019). The cis-Regulatory Atlas of the Mouse Immune System. Cell *176*, 897-912.e820.

Yoshida, J., Akagi, K., Misawa, R., Kokubu, C., Takeda, J., and Horie, K. (2017). Chromatin states shape insertion profiles of the piggyBac, Tol2 and Sleeping Beauty transposons and murine leukemia virus. Sci Rep *7*, 43613.

Yue, F., Cheng, Y., Breschi, A., Vierstra, J., Wu, W., Ryba, T., Sandstrom, R., Ma, Z., Davis, C., Pope, B. D.*, et al.* (2014). A comparative encyclopedia of DNA elements in the mouse genome. Nature *515*, 355-364.

Yusa, K., Zhou, L., Li, M. A., Bradley, A., and Craig, N. L. (2011). A hyperactive piggyBac transposase for mammalian applications. Proc Natl Acad Sci U S A *108*, 1531-1536.

Zeitlinger, J. (2020). Seven myths of how transcription factors read the cis-regulatory code. Curr Opin Syst Biol *23*, 22-31.

Zhang, J., Ding, L., Holmfeldt, L., Wu, G., Heatley, S. L., Payne-Turner, D., Easton, J., Chen, X., Wang, J., Rusch, M.*, et al.* (2012). The genetic basis of early T-cell precursor acute lymphoblastic leukaemia. Nature *481*, 157-163.

Zhao, J. J., Lin, J., Lwin, T., Yang, H., Guo, J., Kong, W., Dessureault, S., Moscinski, L. C., Rezania, D., Dalton, W. S.*, et al.* (2010). microRNA expression profile and identification of miR-29 as a prognostic marker and pathogenetic factor by targeting CDK6 in mantle cell lymphoma. Blood *115*, 2630-2639.

Zuurbier, L., Gutierrez, A., Mullighan, C. G., Canté-Barrett, K., Gevaert, A. O., de Rooi, J., Li, Y., Smits, W. K., Buijs-Gladdines, J. G., Sonneveld, E.*, et al.* (2014). Immature MEF2C-dysregulated T-cell leukemia patients have an early T-cell precursor acute lymphoblastic leukemia gene signature and typically have non-rearranged T-cell receptors. Haematologica *99*, 94-102.

6. Publications

Publications

Fischer A*, Lersch R*, de Andrade Krätzig N, Strong A, Friedrich MJ, Weber J, Engleitner T, Öllinger R, Yen HY, Kohlhofer U, Gonzalez-Menendez I, Sailer D, Kogan L, Lahnalampi M, Laukkanen S, Kaltenbacher T, Klement C, Rezaei M, Ammon T, Montero JJ, Schneider G, Mayerle J, Heikenwälder M, Schmidt-Supprian M, Quintanilla-Martinez L, Steiger K, Liu P, Cadiñanos J, Vassiliou GS, Saur D, Lohi O, Heinäniemi M, Conte N**, Bradley A**, Rad L**, Rad R**. In vivo interrogation of regulatory genomes reveals extensive quasi-insufficiency in cancer evolution. **Cell Genom.** 2023 Mar 8;3(3):100276.

Bredthauer C, **Fischer A**, Ahari AJ, Cao X, Weber J, Rad L, Rad R**, Wachutka L**, Gagneur J**. Transmicron: accurate prediction of insertion probabilities improves detection of cancer driver genes from transposon mutagenesis screens. **Nucleic Acids Res.** 2023 Feb 28;51(4):e21.

Kaltenbacher T*, Löprich J*, Maresch R, Weber J, Müller S, Oellinger R, Groß N, Griger J, de Andrade Krätzig N, Avramopoulos P, Ramanujam D, Brummer S, Widholz SA, Bärthel S, Falcomatà C, **Pfaus A**, Alnatsha A, Mayerle J, Schmidt-Supprian M, Reichert M, Schneider G, Ehmer U, Braun CJ, Saur D, Engelhardt S, Rad R. CRISPR somatic genome engineering and cancer modeling in the mouse pancreas and liver. **Nat Protoc.** 2022 Apr;17(4):1142-1188.

Weber J*, de la Rosa J*, Grove CS*, Schick M, Rad L, Baranov O, Strong A, **Pfaus A,** Friedrich MJ, Engleitner T… Saur D, Liu P, Steiger K, Chudakov DM, Lenz G, Quintanilla-Martinez L, Keller U, Vassiliou GS**, Cadiñanos J**, Bradley A**, Rad R**. PiggyBac transposon tools for recessive screening identify B-cell lymphoma drivers in mice. **Nature Commun.** 2019 Mar 29;10(1):1415.

Rudat S, **Pfaus A**, Cheng YY, Holtmann J, Ellegast JM, Bühler C, Di Marcantonio D, Martinez E, Göllner S, Wickenhauser C, Müller-Tidow C, Lutz C, Bullinger L, Milsom MD, Sykes SM, Fröhling S, Scholl C. RET-mediated autophagy suppression as targetable co-dependence in acute myeloid leukemia. **Leukemia** 2018. Oct;32(10):2189-2202.

Friedrich M, Rad L, Bronner I, Strong A, Wang W, Weber J, Mayho M, Ponstingl H, Engleitner T, Grove C, **Pfaus A**, Saur D, Cadinanos J, Quail MA, Vassiliou GS, Liu P, Bradley A, Rad R. Genome-wide transposon screening and quantitative insertion site sequencing (QiSeq) for cancer gene discovery in mice. **Nature Protocols** 2017, Feb;12(2):289-309.

Conference contribution

Selected Oral Presentation: 23rd European Hematology Association Congress, Stockholm

Pfaus A*, Lersch R*, Weber J, de la Rosa J, Strong A, Rad L, Geumann U, Öllinger R, Engleitner T, Friedrich MJ, de Andrade Krätzig N, Steiger K, Kohlhofer U, Gonzalez-Menendez I, Quintanilla-Martinez L, Heikenwälder M, Liu P, Cadiñanos J, Bradley A**, Vassiliou GS**, Conte N**, Rad R**. PIGGYBAC TRANSPOSON SCREENING IDENTIFIES NOVEL CANCER GENES AND REGULATORY ELEMENTS IN T CELL LEUKEMIA. EHA Library. Pfaus A. 06/16/18; 214552; S862.

* Shared first authorship; ** Shared senior authorship

7. Acknowledgement

Es gibt sehr viele Menschen, denen ich an dieser Stelle für ihre Unterstützung und ihren Beitrag zu dieser Arbeit danken möchte:

Zuerst, danke an Professor Dr. Roland Rad. Dafür, dass Du mir immer vertraut hast mit dem, was ich tue, für Deine Unterstützung und auch dafür, dass Du mir viele Freiheiten gelassen hast. Dadurch habe ich sehr viel gelernt. Danke für Deine Ratschläge und Deine Mühe.

Danke an die Mitglieder meines TACs für die Zeit und die guten Diskussionen: Professor Dr. Angelika Schnieke, PD Dr. Ursula Zimber-Strobl und Professor Dr. Stefan Fröhling. Danke Stefan, dass Du auch außerhalb der TAC Meetings hilfreiche Ratschlägen gegeben hast.

Ich danke allen Ko-Autoren des aus dieser Arbeit entstandenen Manuskripts für deren wertvollen Beitrag. Hier vor allem dem Team von Professor Dr. Allan Bradley für die wegweisenden Vorarbeiten.

Danke an alle Kollaborationspartner, ohne die meine Projekte nicht so erfolgreich gewesen wären: Danke Laura für unsere gute Kollaboration im SFB Projekt. Danke Merja für Deine ausführlichen Mails und Deine Ratschläge in unseren Skype Treffen. Danke Michele und Franzi für Eure Hilfe beim FACSen und die schönen Retreats. Danke an alle Tierpfleger, die sich immer gut um die Mäuse gekümmert haben.

Danke Elizabeth und Elke vom SFB1243. Mit Euch war es immer nett und eure Unterstützung sehr wertvoll. Danke für Euren Einsatz und Eure Hilfe.

Besonderen Dank geht an meine Kollegen. Danke, dass ihr immer zugehört und mitgefühlt habt, beziehungsweise zur richtigen Zeit einfach ein Bier mit aufgemacht habt. Danke, dass auch spät abends immer jemand da war. Wenn man Freunde als Kollegen hat, ist alles viel einfacher. Danke Wolle, Robert, Thorsten, Roman, Miguel, Katha, Jessi, David, Christine und Nina. Danke natürlich auch an Leute außerhalb der Gruppe: Christian V., Christian S., Stefanie, Mariel, Chiara, Fabio. Danke für unsere schönen Abende und Ausflüge durch halb Europa.

Besonderen Dank geht hier an Wolle, für alle Obstzeiten, ohne die der erste Teil dieser Zeit sehr langweilig gewesen wäre. Schön, dass ich jemanden in meiner Nähe hatte, den ich lange kannte und auf den ich immer zählen konnte.

Danke an Deine fachliche Unterstützung, Deine Meinung und schnellen Antworten (auch aus der Ferne) an Julia.

Und danke Robi! Dafür, dass Du mich zwar manchmal in den Wahnsinn getrieben hast, aber Deine unglaubliche positive Art war immer sehr motivierend. Danke, dass Du zu jeder Zeit da warst, geholfen hast und ich mich immer auf Dich verlassen konnte. Unsere Zusammenarbeit hat immer sehr viel Spaß gemacht!

Danke an die Bioinformatiker Thomas, Niklas und Mathias für viele interessante Gespräche. Vor allem auch tausend Dank an Justus und Leo! Es macht super viel Spaß mit Euch zu arbeiten und ich bin wirklich dankbar, dass man sich bei allen Fragen immer auf Euch verlassen kann.

Und danke an alle anderen Kollegen der AG Rad, die ich nicht alle einzeln hier erwähnen kann, die aber immer sehr hilfsbereit waren.

Danke an meine neuen Kollegen in Ulm - besonders an Professor Dr. Reiner Siebert - die während der Fertigstellung dieser Arbeit auf verschiedenste Weise sehr entgegenkommend waren.

Danke an mein Doktorarbeits-Korrektur-Team Robi, Annika, Mella, Mareike und Eike.

Danke an alle meine Freunde aus den verschiedensten Abschnitten meines Lebens. Tausend Dank für alle schönen Momente in den letzten Jahren, die ein wichtiger Ausgleich zur Laborarbeit waren. Danke an meine Ulmer me-gusta Mädels Kim, Annika, Mia, Carina und Lena. Ihr seid schon so lange Teil meines Lebens und ich bin sehr froh, dass ihr das trotz wenig Zeit meinerseits immer noch seid. Danke Mareike für viele schöne Abende in und um München und deine endlose Unterstützung am Telefon. Danke an meine Erlanger Julia, Jassi, Thomas, Christine, Jana und Anne! Das regelmäßige Treffen und der Austausch mit Euch war immer schön. Danke auch, dass ihr dafür gesorgt habt, dass Benjamin der letzte ist, der fertig wird und nicht ich. Danke an die (ehemaligen) Münchner Thanos, Felix, Annika, Kim, Jana, Christine, Lilli und Alice, die mir vor allem meine Anfangszeit hier in München echt einfach gemacht haben.

Herzlichsten Dank auch an meine ganze Familie und Familie Fischer für Eure Unterstützung. Ohne diesen Halt zu Hause wäre das ganze deutlich schwieriger gewesen. Ich bedanke mich bei Euch allen für Euren Glauben an mich und für Eure Nachsicht, dass ich auch oft nicht da war.

Besonderen Dank geht an meine Schwester Mella und meine Mama Ute. Für’s Nach-Hause-Telefonieren, wenn es mal wieder spät wurde und dafür, dass ihr euch immer alles angehört habt. Ohne Euch wäre ich nicht annähernd da, wo ich heute bin. Danke für Eure Unterstützung, auch wenn ich manchmal mit den verrücktesten Plänen ankomme. Ich danke euch aus tiefstem Herzen.

Zu allerletzt: Danke, Eike! Ich denke, dass es sich kaum in Worte fassen lässt, wie dankbar ich Dir bin. Für Deine pausenlose und unschätzbare Unterstützung. Dafür, dass Du wirklich immer für mich da warst und das, obwohl Du so genau weißt, wie sich das alles anfühlen kann.

Danke, dass ich so viele herzliche Leute um mich habe, die mich begleiten und denen ich hier danken kann. Und danke dafür, dass Ihr Euch diese Arbeit angeschaut habt (wenn auch nur teilweise), obwohl die meisten wenig Ahnung davon haben, um was es geht. Das ist die Art der Unterstützung, die man eigentlich nicht beschreiben kann.