DOCTORAL DISSERTATION

EPIGENETIC BLUEPRINT OF HUMAN THYMOPOIESIS AND ADULT T-CELL ACUTE LYMPHOBLASTIC LEUKEMIA

ANAND MAYAKONDA
DIVISION OF CANCER EPIGENOMICS (B370)
Deutsches Krebsforschungszentrum (DKFZ)

2021

Dissertation

Epigenetic blueprint of human thymopoiesis and adult T-cell Acute Lymphoblastic Leukemia

Anand Mayakonda Thippeswamy

2021

Dissertation

submitted to the

Combined Faculty of Natural Sciences and Mathematics

of the Ruperto Carola University Heidelberg, Germany

for the degree of

Doctor of Natural Sciences

Presented by

Anand Mayakonda Thippeswamy
Master of Science, Bioinformatics

Born in Hubli, India

Oral examination: July 5th 2021

Epigenetic blueprint of human thymopoiesis and adult T-cell Acute Lymphoblastic Leukemia

Referees

Prof. Dr. Benedikt Brors

Prof. Dr. Christoph Plass

ಈ ಪ್ರಬಂಧವನ್ನು ನಾನು ನನ್ನ ಪ್ರೀತಿಯ ಪೋಷಕರಾದಂತಹ

ಎಂ. ಆರ್. ತಿಪ್ಪೇಸ್ವಾಮಿ ಹಾಗು ಎಂ. ಟಿ. ಲಲಿತಾ,

ಮತ್ತು ಪೋಷಕ ಸಮಾನರಾದಂತಹ

ಶಾಂತೇಶ್ ಸಜ್ಜನ್ ಹಾಗು ನಾಗಮ್ಮ ಸಜ್ಜನ್ ರವರಿಗೆ ಅರ್ಪಿಸುತ್ತೇನೆ.

This work is dedicated to my beloved parents

M. R. Thippeswamy and M. T. Lalita,

and to my guardians

Shantesh Sajjan and Nagamma Sajjan.

Division of Cancer Epigenomics (B370)

Head of Division: Prof. Dr. Christoph Plass

German Cancer Research Center (DKFZ)

Heidelberg, Germany

CONTRIBUTIONS

This thesis was performed in collaboration with Prof. Dr. Vahid Asnafi from Université de Paris. The analyzed cohort of 143 primary adult T-ALLs was primarily generated as part of the French GRAALL 2003-2005 clinical trials[i] [ii]. Data generated by EPIC arrays and thymic cell sorting were performed by the same group.

Contributions (ordered alphabetically):

Bioinformatic analysis: Anand Mayakonda[1,2]

I am also grateful for the helpful suggestions received from the following people who have been acknowledged wherever necessary in the manuscript: Guillaume P. Andrieu[3], Joschka Hey[1,2,4], Pavlo Lutsik[1,5], Reka Toth[1]

Thymic cell isolation and experimental work: Agata Cieslak[3], Aurore Touzart[1,3], Charlotte Smith[3]

Sequencing library preparation: Aurore Touzart[1,3] Dieter Weichenhan[1], Marion Bähr[1]

Affiliations:

[1]Cancer Epigenomics, German Cancer Research Center (DKFZ), Heidelberg, Germany

[2]Faculty of Biosciences, Heidelberg University, 69120 Heidelberg, Germany

[3]Université de Paris, Institut Necker -Enfants Malades (INEM), Institut national de la santé et de la recherche médicale (Inserm) U1151, and Laboratory of Onco-Hematology, Assistance Publique-Hôpitaux de Paris, Hôpital Necker Enfants-Malades, Paris, France

[4]Germany-Israeli Helmholtz Research School in Cancer Biology

[5]German Cancer Research Consortium (DKTK)

The vast majority of the **results** in **section 3.2**, are integrated into the manuscript *"Epigenetic blueprint identifies poor outcome and hypomethylating agent-responsive T-ALL subgroup,"* which has been accepted as an *original research article* in the journal *Science Translational Medicine* and scheduled for publication during May 2021. The author of the present thesis, Anand Mayakonda, is a co-first author of the soon-to-be-published article and contributed significantly to data analysis, data interpretation, figure design, and manuscript writing.

[i] https://clinicaltrials.gov/ct2/show/NCT00222027

[ii] https://clinicaltrials.gov/ct2/show/NCT00327678

Section **5.2.2** of the **material and methods,** describing a computational framework for the whole-genome bisulfite sequencing data analysis, is now published as an *application note* in the journal *Bioinformatics*[iii] under the title *"Methrix: an R/Bioconductor package for systematic aggregation and analysis of bisulfite sequencing data."* The author of the present thesis, Anand Mayakonda, is the first author of the published article and contributed significantly to the tool development, data analysis, data interpretation, figure design, and manuscript writing.

The results section contains text from the above mentioned publications originally written by myself, and may contain suggestions from co-authors. In compliance with good scientific practice of the University of Heidelberg, these text sections are therefore put between quotation marks wherever necessary. Contributions from co-authors - either experimental or data analysis, are acknowledged in the figure/table legends wherever necessary.

[iii] https://doi.org/10.1093/bioinformatics/btaa1048

DECLARATIONS

Declarations according to § 8 (3) b) and c) of the doctoral degree regulations:

a) I hereby declare that I have written the submitted dissertation myself and, in this process, have used no other sources or materials than those expressly indicated,

b) I hereby declare that I have not applied to be examined at any other institution, nor have I used the dissertation in this or any other format any other institution as an examination paper, nor submitted it to any other faculty as a dissertation.

Anand Mayakonda

TABLE OF CONTENTS

SUMMARY

Thymopoiesis is a process by which bone-marrow-derived lymphoid progenitor cells migrate to the thymus and undergo multi-step differentiation into mature CD4+ or CD8+ T-lymphocytes. The entire process is tightly regulated and governed by the transcriptional/epigenetic changes necessary for lineage commitment and cellular identity. Genetic lesions such as somatically acquired point mutations or chromosomal rearrangements lead to differentiation blockade resulting in hematological malignancy known as *T-cell acute lymphoblastic leukemia* (T-ALL). While the thymopoiesis and T-ALL are well characterized by transcriptional studies, high-resolution mapping of the epigenetic changes is still lacking.

DNA methylation (DNAm) changes involving the addition of de-novo, or the erasure of existing methyl groups from the cytosine nucleotide, are dynamic during cellular differentiation and form the cell-type-specific signatures. In this doctoral thesis, DNAm dynamics during the human thymopoiesis is studied by whole-genome bisulfite sequencing of seven distinct intra-thymic cell types. DNAm changes during the thymopoiesis are characterized by the uni-directional and irreversible loss methylation primarily occurring at the transcription factor binding sites critical for T-cell lineage commitment (e.g., *NOTCH1* and *MYB*) and T-cell receptor rearrangements. A DNAm atlas of thymopoiesis is established by identifying 381 *de-novo* differentially methylated regions (tDMRs) that are highly conserved across cell-types originating from the thymic lineage. The tDMRs can recapitulate the in-silico ontogeny of T-cell differentiation and are validated in an independent dataset. Remarkably, combined analysis with bone-marrow-derived hematopoietic progenitors and peripheral derived mature blood cells shows the hypermethylation of tDMRs among non-lymphoid cell types suggesting the epigenetic silencing of pathways necessary for thymic lineage commitment.

To further highlight the role of tDMRs in disease development, a combined array-centric analysis of intra-thymic cell types and a well-defined cohort of 143 primary adult T-ALLs was performed. Interestingly, DNAm classified the T-ALL cohort into five distinct subtypes (C_1-C_5) with characteristic levels of DNAm (C_1 lowest level and C_5 the highest). Moreover, each

subtype is correlated with a specific somatic event, including a novel adult T-ALL specific subtype with co-occurring *DNMT3A/IDH2* mutations (C_1), and well known transcriptionally deregulated subtypes resulting from *TAL1* (C_2), *TLX3* (C_3), *TLX1*/in *cis-HOXA9* (C_4), or in *trans-HOXA9* (C_5) overexpression. Utilizing tDMRs as the blueprint, maturation arrest stages of T-ALL subtypes are established, revealing a hierarchical ordering, with C_1 and C_5 arising earlier during the T-cell development followed by *TLX3/1* overexpression (C_3, C_4) and *TAL1* deregulation (C_5). Although tDMRs highlight the ontogeny of T-ALL subtypes, global DNAm levels did not correlate with the maturation arrest stages suggesting a non-linear association of DNAm and differentiation blockade. Subsequent integrative analysis with epigenetic marks associated with active transcription (H3K27ac and H3K4me1) revealed the hypomethylation of pathogenic enhancer elements. Importantly, careful survival analysis identified an unexpected, clinically aggressive hypermethylated subtype (C_5) that can be targeted with DNA hypomethylating agents. Finally, using machine learning models, a 79 CpG classifier was developed for *de-novo* classification of newly diagnosed adult T-ALLs.

In summary, results from the comprehensive analysis of DNAm changes during human thymopoiesis and the subsequent modifications leading to T-ALL provide meaningful insights into the role of DNAm in maintaining the cellular identity and disease development. Furthermore, the identification of clinically actionable hypermethylated T-ALL subtype paves the way for targeted epigenetic therapies.

ZUSAMMENFASSUNG

Die Thymopoese ist ein Prozess, bei dem aus dem Knochenmark stammende lymphoide Vorläuferzellen in den Thymus wandern und eine mehrstufige Differenzierung zu reifen CD4+ oder CD8+ T-Lymphozyten durchlaufen. Der gesamte Prozess ist streng reguliert und wird von den transkriptionellen/epigenetischen Veränderungen gesteuert, die für das Lineage Commitment und die zelluläre Identität notwendig sind. Genetische Läsionen wie somatisch erworbene Punktmutationen oder chromosomale Rearrangements führen zu einer Blockade der Differenzierung und damit zu einer hämatologischen Malignität, die als akute lymphoblastische T-Zell-Leukämie (T-ALL) bekannt ist. Während die Thymopoese und die T-ALL durch transkriptionelle Studien gut charakterisiert sind, fehlt es noch an einer hochauflösenden Kartierung der epigenetischen Veränderungen.

DNA-Methylierungs (DNAm)-Veränderungen, die das Hinzufügen von de-novo oder das Löschen bestehender Methylgruppen vom Cytosin-Nukleotid beinhalten, sind während der zellulären Differenzierung dynamisch und bilden die zelltypspezifischen Signaturen. In dieser Dissertation wird die DNAm-Dynamik während der menschlichen Thymopoese durch Ganzgenom-Bisulfit-Sequenzierung von sieben verschiedenen intra-thymischen Zelltypen untersucht. DNAm-Veränderungen während der Thymopoese sind durch uni-direktionale und irreversible Verlust-Methylierung charakterisiert, die hauptsächlich an den Transkriptionsfaktor-Bindungsstellen auftritt, die für die T-Zell-Linienbindung (z.B. NOTCH1 und MYB) und T-Zell-Rezeptor-Rearrangements entscheidend sind. Ein DNAm-Atlas der Thymopoese wird durch die Identifizierung von 381 de-novo differentiell methylierten Regionen (tDMRs) erstellt, die über Zelltypen, die aus der thymischen Abstammung stammen, hoch konserviert sind. Die tDMRs können die in-silico Ontogenie der T-Zell-Differenzierung rekapitulieren und werden in einem unabhängigen Datensatz validiert. Bemerkenswerterweise zeigt die kombinierte Analyse mit aus dem Knochenmark stammenden hämatopoetischen Vorläufern und aus der Peripherie stammenden reifen Blutzellen die Hypermethylierung der tDMRs unter den nicht-lymphoiden Zelltypen, was auf die epigenetische Stilllegung von Signalwegen hindeutet, die für die thymische Abstammung notwendig sind.

Um die Rolle der tDMRs bei der Krankheitsentwicklung weiter zu beleuchten, wurde eine kombinierte Array-zentrierte Analyse von intra-thymischen Zelltypen und einer gut definierten Kohorte von 143 primären erwachsenen T-ALLs durchgeführt. Interessanterweise klassifizierte DNAm die T-ALL-Kohorte in fünf verschiedene Subtypen (C_1-C_5) mit charakteristischen DNAm-Werten (C_1 der niedrigste Wert und C_5 der höchste). Darüber hinaus ist jeder Subtyp mit einem spezifischen somatischen Ereignis korreliert, einschließlich eines neuen erwachsenen T-ALL-spezifischen Subtyps mit gleichzeitig auftretenden DNMT3A/IDH2-Mutationen (C_1) und gut bekannten transkriptionell deregulierten Subtypen, die aus TAL1 (C_2), TLX3 (C_3), TLX1/in cis-HOXA9 (C_4) oder in trans-HOXA9 (C_5) Überexpression resultieren. Unter Verwendung der tDMRs als Blaupause werden Reifungsarrest-Stadien von T-ALL-Subtypen etabliert, die eine hierarchische Anordnung aufzeigen, wobei C_1 und C_5 früher während der T-Zell-Entwicklung auftreten, gefolgt von TLX3/1-Überexpression (C_3, C_4) und TAL1-Deregulation (C_5). Obwohl die tDMRs die Ontogenese der T-ALL-Subtypen hervorheben, korrelierten die globalen DNAm-Spiegel nicht mit den Stadien des Reifungsstopps, was auf eine nicht-lineare Assoziation von DNAm und Differenzierungsblockade hindeutet. Eine anschließende integrative Analyse mit epigenetischen Markierungen, die mit aktiver Transkription assoziiert sind (H3K27ac und H3K4me1), zeigte die Hypomethylierung von pathogenen Enhancer-Elementen. Wichtig ist, dass eine sorgfältige Überlebensanalyse einen unerwarteten, klinisch aggressiven hypermethylierten Subtyp (C_5) identifizierte, der mit DNA-hypomethylierenden Wirkstoffen gezielt behandelt werden kann. Schließlich wurde mit Hilfe von Machine-Learning-Modellen ein 79 CpG-Klassifikator für die de-novo-Klassifikation von neu diagnostizierten erwachsenen T-ALLs entwickelt.

Zusammenfassend lässt sich sagen, dass die Ergebnisse der umfassenden Analyse der DNAm-Veränderungen während der menschlichen Thymopoese und der nachfolgenden Modifikationen, die zu T-ALL führen, aussagekräftige Einblicke in die Rolle der DNAm bei der Aufrechterhaltung der zellulären Identität und der Krankheitsentwicklung liefern. Darüber hinaus ebnet die Identifizierung von klinisch wirksamen hypermethylierten T-ALL-Subtypen den Weg für gezielte epigenetische Therapien.

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

Acronym	Definition
4ISP	Immature single CD4 positive
5-hmC	5-Hydroxy methyl cytosine
5mC	5-methyl cytosine
ALL	Acute Lymphoblastic Leukemia
APC	Antigen presenting cells
BCP	B-cell precursors
bHLH	Basic helix loop helix
BM	Bone marrow
CGI	CpG island
ChIP	Chromatin immunoprecipitation
CI	Confidence interval
CIMP	CpG Island Methylation Phenotype
CLP	Common lymphoid progenitor
CMP	Common myeloid progenitor
CNS	Central nervous system
CNV	Copy number variation
CpG	Cytosine-phosphate-guanine
CR	Complete remission
CTCF	CCCTC-binding factor
DEG	Differentially expressed genes
DLL4	Delta like ligand 4
DMP	Differentially methylated probe
DMR	Differentially methylated region
DN	Double negative
DNA	Deoxyribonucleic acid
DNAm	DNA methylation
DNMT	DNA methyl transferase
EC	Early cortical
EFS	Event free survival
EPIC	Illumina methylationEPIC arrays
ETP	Early thymic progenitor
FACS	Florescence assisted cell surface

FDR	False discovery rate
GISTIC	Genomic identification of significant targets in cancer
GMP	Granulocyte monocyte progenitor
H3K27ac	Histone 3 lysine 27 acetylation
H3K4me1	Histone 3 lysine 4 monomethylation
H3K4me3	Histone 3 lysine 4 trimethylation
HM450K	Illumina Human methylation 450K arrays
HSC	Hematopoietic stem cells
ICN	Intra cellular NOTCH
IDH	Isocitrate dehydrogenase
IL7	Interleukin 7
L-MPP	Lymphoid multipotent progenitors
LC	Late cortical
LOLA	Locus overlap enrichment analysis
MEP	Megakaryocyte-erythroid progenitor
MHC-1	Major histocompatibility class 1
MHC-2	Major histocompatibility class 2
MPP	Multipotent progenitor
MRD	Minimal residual disease
NK-cells	Natural killer
NKP	Natural killer progenitor
NMF	Non-negative matrix factorization
OS	Overall survival
PB	Peripheral blood
PBAT	Post bisulfite adapter tagging
PCA	Principal component analysis
PRC2	Polycomb repressor complex 2
RF	Random forest
RNA	Ribonucleic acid
RRBS	Reduced representation bisulfite sequencing
SCT	Stem cell transfer
SE	Super enhancer
SNP	Single nucleotide polymorphism
T-ALL	T-cell acute lymphoblastic leukemia
TCGA	The cancer genome atlas
TCP	T-cell progenitor
TCR	T-cell receptor

tDMRs	Thymic differentially methylated regions
TE	Typical enhancer
TEC	Thymic epithelial cells
TET	Ten eleven translocations
TF	Transcription factor
TFBS	Transcription factor binding site
TSG	Tumor suppressor gene
TSS	Transcription start site
UPD	Uniparental disomy
VAF	Variant allele frequency
WBC	White blood cell
WGBS	Whole genome bisulfite sequencing

1 INTRODUCTION

1.1 Thymopoiesis

1.1.1 Immune System

The human immune system comprises two constituents: the *innate* and the *adaptive* immune system, originating from the hematopoiesis (**Figure 1)**. The innate immune system is generated by the hematopoietic system's myeloid compartment and consists of monocytes, neutrophils, eosinophils, basophils, and dendritic cells. It forms the hosts' immune system's primary defense and is responsible for immediate response against invading pathogens. The innate immune system is restricted in identifying the pathogens and limited by the repertoire of receptors, primarily consisting of conserved domains across a large group of pathogens. Recognition of microorganisms by the innate compartment utilizes special receptors known as *pattern recognition receptors,* which further activates defense mechanisms such as phagocytosis (by macrophages and neutrophils) or the secretion of interferons (Akira et al. 2006). Of note, the innate immune system's defense mechanism is non-specific and often affects the host homeostasis (such as inflammation and fever).

On the contrary, the adaptive immune system is highly specific and mediated by the specialized cells called *lymphocytes* originating from the hematopoietic system's lymphoid lineage. Unlike innate immunity, the adaptive immune system can recognize a wide array of pathogens and can result in long-term immunity against the same pathogen due to its ability to *memorize.* The adaptive immune system's diversity is facilitated by the somatic recombination of gene segments resulting in highly target specific receptors.

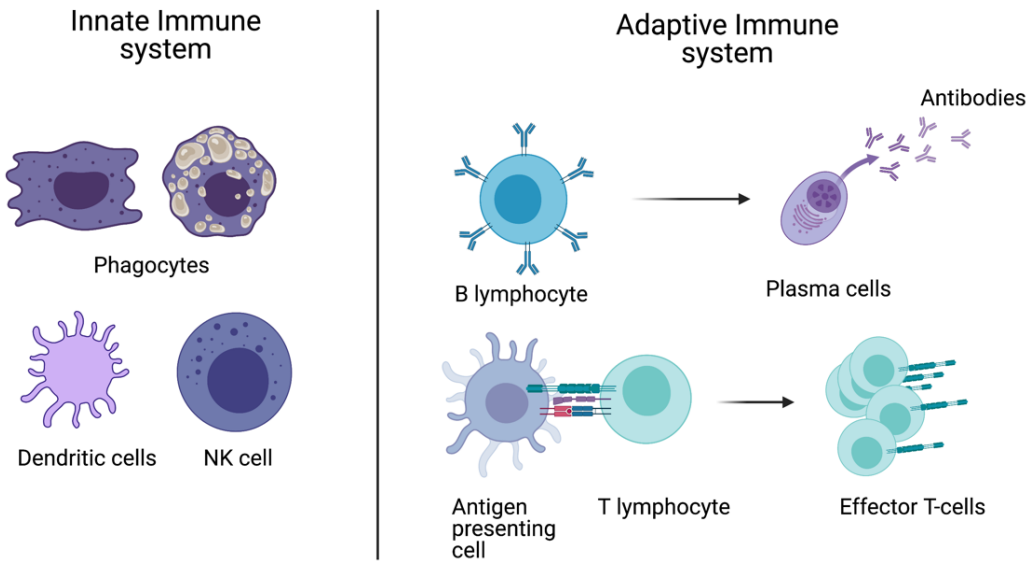

Figure 1. Comparison of Innate and Adaptive immune system.

Left panel shows cell types involved in innate immunity whereas, right panel shows brief mode of action for B-lymphocytes and T-lymphocytes. Created with BioRender.com

Broadly, lymphocytes from the adaptive immune system consisting of *B-cells* and *T-cell*s, both of which serve a specific purpose and significantly differ in action mode. B-cells originate from bone marrow (BM) and form the *humoral* element of the adaptive immunity. B-cells themselves do not interact with the invading pathogens but generate *antibodies* that can bind to the specific antigens and neutralize the invading pathogens. Moreover, upon antigen recognition, a subset of B-cells differentiates into memory B-cells, recognizing and rapidly acting against the repeated stimuli.

T-cells, however, originate from BM-derived lymphoid progenitors, which further undergo maturation in the *Thymus* (Schmitt et al. 1995). T-cell maturation in the thymus is a hierarchical process and involves *positive selection* – a process by which self-reacting T-cells are eliminated. A detailed process of T-cell maturation is discussed in subsequent chapters. Successfully differentiating T-cells leave the thymus and enter the periphery. Peripheral T-cells are under constant surveillance and are activated upon interacting with the antigens presented by *antigen-presenting cells* (APC). Upon activation, T-cells undergo proliferation resulting in a clonal population of *effector T-cells* which can identify and eliminate the infected cells carrying the surface foreign antigen.

1.1.2 Thymus: structure and function

Anatomically, the human thymus is a bi-lobed, granulated organ located above the heart. Historically, the thymus had been considered a vestigial organ with no functional properties and was observed to undergo *involution* - a process by which an organ loses its tissue mass along with age (Geenen 2017). Until the 1960's when several experiments showed that thymectomy in neo-natal mice was fatal, and often the progeny lacked the proper functioning immune system. Further experiments over the years proved that the thymus is a critical immune organ and provides a required niche for developing a particular subset of immune cells, which was later identified as T-cells (Miller 2020).

The thymus is a capsulated structure consisting of an outer cortical layer and the medulla's inner mass. Both cortex and medulla have a unique role in T-cell development and are distinguished by the sub-capsular zone (**Figure 2)**. Specialized epithelial cells known as *thymic epithelial cells* (TECs) densely pack the tissue forming a mesh-like structure that allows maturing lymphocytes to move between the thymic compartments. The inner medulla region also houses APCs called dendritic cells, critical for modulating self-reacting T-cells. Lymphocytes are abundant in the cortex, whereas; dendritic cells significantly occur within the medulla.

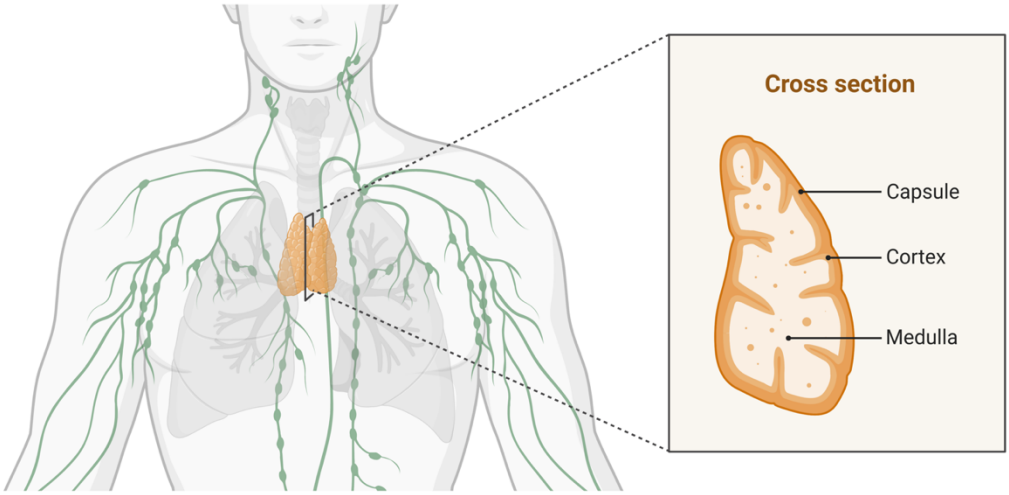

Figure 2. The anatomical structure of the human thymus.

The thymus is a bi-lobed structure located above the heart. The thymus is a capsulated tissue consisting of an outer cortical layer and an inner mass called the medulla. Created with BioRender.com

Altogether, the thymus is a complex organ that provides the necessary niche for T-cell development while restricting the formation of intolerant and self-reactive immune cells. The overall process by which the progenitor cells migrating from the bone marrow enter the thymus, undergo multi-stage differentiation and selection to form functional and immune-competent T-cells is known as *thymopoiesis*.

1.1.3 Developmental stages of αβ T-cells

Classical hematopoiesis involves specialized, self-renewing stem cells known as Hematopoietic Stem Cells (HSCs) – that undergo hierarchical and uni-directional differentiation process resulting in the formation of lineage-restricted cell types. Broadly, hematopoiesis gives rise to the Myeloid and Lymphoid lineages, which form the entire lymphatic system (Kondo et al.). Myeloid lineage constitutes cell types responsible for oxygen transportation (by erythrocytes), blood clotting (by platelet producing megakaryocytes), and the innate immune system. Lymphoid lineage involves cell types responsible for the adaptive immune compartment, namely, antibody-secreting B-cells and cytotoxic T-cells. While the formation of the myeloid derived hematopoietic cell types occurs in the bone marrow itself, mature T-cells develop in the thymus. Progenitor cells primed towards T-cells (known as TCPs) leave the bone marrow and migrate to the thymus via the lymphatic system. Migration of TCPs from BM to the thymus is guided by the signaling molecule Interleukin-7 (IL7) and depends on the expression of the IL7 receptor by TCPs.

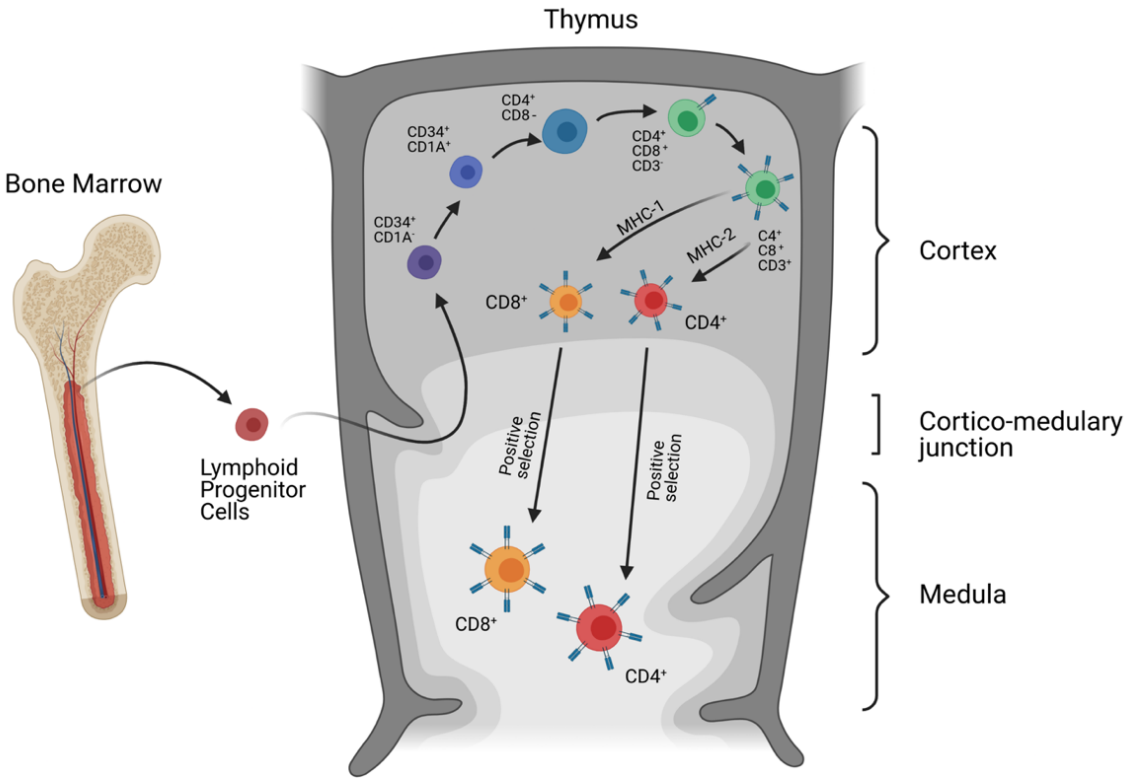

Figure 3. Intrathymic T-cell development.

Lymphoid progenitor cells primed towards T-cell development migrate from bone marrow and enter the thymus through the cortico-medullary junction. Early CD34+ thymic precursor (ETP) cells lacking CD4- and CD8- receptors commit to T-lineage in subsequent stages. Lineage commitment is proceeded by expressing T-cell receptors resulting in double-positive (DP; CD4+ CD8+ CD3+) cells. Depending on their interaction with class-1 MHC or class-2 MHC, DP cells differentiate into single-positive CD3+ or CD8+ cells. SP cells successfully surviving the selection, migrate to the medulla and leave the thymus. Created with BioRender.com

Successfully responding TCPs enter the thymus through the outer cortex layer and are known as Early Thymic Progenitor cells (ETPs). ETPs then undergo a cascade of uni-directional developmental changes characterized by the expression of cell surface receptors and transcriptional profiles. These events are tightly regulated and occur in an orderly manner, ultimately resulting in fully functional $CD4^+$ helper and $CD8^+$ cytotoxic T-cells (Koch and Radtke). ETPs often show the expression of CD34 and still retain the stem cell-like properties. Moreover, ETPs also lack CD4 and CD8 markers (known as double negative or DN cells). In general, critical stages of development of human αβ T-cells involves differentiation of ETP cells ($CD34^+$ $CD1A^-$) into immature single-positive cells (ISPs; $CD34^-$ $CD1A^+$ $CD4^+$ $CD8^-$ $CD3^-$), then into double-positive (DP; $CD4^+$ $CD8^+$ $CD3^+$) cells, and finally forming a mature single-positive (SP; $CD4^+$ $CD8^-$ $CD3^+$, $CD4^-$ $CD8^+$ $CD3^+$) cells (**Figure 3**).

It is worth noting that the ETP cells, albeit migrating to the thymus, still retain the capacity to give rise to non-T-cell populations such as B-cells and some of the myeloid cells (Bell and Bhandoola 2008). During the subsequent stages, ETPs lose their multi-potency and commit towards T lineage in a process often referred to as *lineage commitment*. T-cell lineage commitment involves activating specific cellular signaling pathways and transcription factors (TFs), critical for T-cell development. A receptor-ligand-driven NOTCH pathway plays a crucial role in T-cell lineage commitment (Radtke et al. 1999). Briefly, NOTCH receptors expressed by the early lymphocytes interact with the delta-like protein 4 (*DLL4*) ligands found on the cortical TECs, thereby initiating the NOTCH signaling system. The binding of NOTCH to DLL4 results in the cleavage and migration of intracellular NOTCH (ICN) into the nucleus, further activating the target genes necessary for T-cell development (Schmitt and Zuniga-Pflucker 2002). Additionally, downstream expression of key transcription factors such as *GATA3, TCF7, BCL11B, E2A, HEB*, and RUNX form regulatory networks to facilitate T-cell development (Hosoya et al. 2009; Ashworth et al. 2010).

1.2 T-cell Acute Lymphoblastic Leukemia

1.2.1 Molecular biology and pathogenesis of T-ALL

T-cell Acute Lymphoblastic Leukemia (T-ALL) is a sub-type of lymphoblastic leukemia characterized by T-cell differentiation defects resulting in the formation and accumulation of immature, partially differentiated, and non-functional thymocytes (Raetz and Teachey 2016). Clinically, T-ALL patients show massive infiltration of bone marrow with immature thymocytes, enlarged thymus, increased white blood cells, neutropenia, anemia, and thrombocytopenia (You et al. 2015). T-ALL is common among children (10-15% of all ALLs) and young adults (20% of all ALLs) (Hunger and Mullighan 2015; Litzow and Ferrando 2015). Recent advances in the intense chemotherapy regimens have resulted in high cure rates for T-ALL (up to 80% for pediatric T-ALL and 50% for adults) (Pui et al. 2008; Stock et al. 2013). However, a substantial portion of patients displays clinical relapse with resistance to further therapy.

The primary pathogenesis involves somatically acquired genetic lesions such as point mutations, copy number aberrations, and chromosomal translocations. Earlier work by several groups identified activating mutations in *NOTCH1* and its downstream target *FBXW7*

in ca. 60% of the T-ALL cases (Weng et al. 2004). Mutations in NOTCH pathways now serve as T-ALL hallmarks and tend to co-occur with deletions in *CDKN2A*. However, recent genomic studies backed by massively parallel sequencing have allowed further comprehensive characterization of the disease, thereby identifying somatic mutations in over 100 driver genes affecting several previously overlooked pathways. Especially genes involved in PI3K-AKT (29%), JAK-STAT (5%), Ras (4%), Ribosomal (3%), Ubiquitination (9%), and RNA processing (9%) were identified (Liu et al. 2017).

In addition to the somatically acquired mutations, T-ALL is characterized by chromosomal translocations, resulting in the expression of undesired transcription factors which impede the T-cell development. Some of the critical transcription factors that are over-expressed in T-ALL include basic helix-loop-helix (bHLH) family members (*TAL1, TAL2, LYL1*) (Powell-Jones et al. 1976; Begley et al. 1989), LMO genes (*LMO1, LMO2*) (McGuire et al. 1989; Kennedy et al. 1991; Royer-Pokora et al. 1991), and developmental associated HOX genes (*HOXA9, HOXA10, TLX1, TLX2,* and *TLX3*) (Soulier et al. 2005). The most common mechanism of overexpression of these transcription factors involves *enhancer hijacking*, placing a gene under the influence of strongly expressed enhancers. In T-ALL, transcription factors are identified most commonly under the T-cell developmental associated genes (e.g., TCR-α or TCR-δ enhancer). In addition to the enhancer hijacking, a small portion of T-ALL patients show overexpression of TFs mediated by mutations in upstream non-coding regions, which creates binding pockets for its target genes by acting as neo-enhancers (e.g., de-novo MYB binding sites created by promoter mutations in *TAL1*) (Mansour et al. 2014).

1.2.2 Oncogenic transcription factors in T-ALL

bHLH transcription factors

Aberrantly expressed bHLH family transcription factors include *TAL1, TAL2*, and LYL genes (*LYL1/LYL2*). *TAL1* is deregulated in over 25% of T-ALL while the rest are expressed in less than 2% of T-ALL. A common overexpression mechanism involves chromosomal rearrangements, which place *TAL1* near the regulatory regions associated with T-cell developmental specific genes TCRα and TCR-δ (Begley et al. 1989; Bernard et al. 1990; Chen et al. 1990). Besides, mutations in upstream regions of *TAL1* are shown to create binding sites for *MYB* transcription factors leading to the monoallelic expression of *TAL1* (Mansour et al. 2014). Mechanistically, conditional expression of *TAL1* in T-cells leads to leukemic transformation in

mice models (Condorelli et al. 1996). Moreover, *TAL1* forms a regulatory circuit with *RUNX1*, *GATA3*, forming a feed-forward loop that drives the leukemic oncogenic programs (Sanda et al. 2012). *TAL1* deregulated T-ALLs often occur at the later stage of T-cell development and are associated with a poor prognosis (Sanda and Leong 2017). Epigenetically, DNA methylation (DNAm) based studies have shown TAL1 deregulation leads to hypomethylated CpG islands and resembles normal thymic cell types (Touzart et al. 2020).

HOX family transcription factors

HOX family genes are well described in *Drosophila* as the associated developmental genes involved in body segmentation and body parts formation (Garcia-Fernandez 2005; Pearson et al. 2005). Similar to *TAL1*, chromosomal translocations place the cluster of HOX genes (*HOXA9/HOXA10*) near the regulatory genes associated with the TCR-α/TCR-δ genes leading to its overexpression (Soulier et al. 2005). HOXA genes are aberrantly expressed in about 3% of the T-ALL and are related to poor outcomes. Developmentally, HOXA deregulation is characteristic of ETP-ALL and occurs earlier during the T-cell development (Soulier et al. 2005).

In addition to *HOXA19/10*, *TLX1/3* are the well-known deregulated HOX family genes frequently overexpressed in T-ALL. *TLX1* and *TLX3* expression also show distinct enrichment among pediatric and adult T-ALL, with *TLX1* being more frequent in adult T-ALL (25%), whereas *TLX3* is more frequent in pediatric samples (25%) (Ferrando et al. 2004). Moreover, TLX3 overexpression is driven by translocations placing the gene near the *BCL11B* gene, whereas *TLX1* is placed near TCR-δ enhancers (Bernard et al. 2001). Clinically, *TLX1/3* positive patients show a better prognosis and overall survival (De Keersmaecker and Ferrando 2011). Moreover, gene expression studies have identified a standard set of genes and pathways altered in both the subgroups suggesting a traditional mode of action (Della Gatta et al. 2012). Developmentally, *TLX1/3* deregulation occurs post-T-cell commitment stage of T-cell development.

1.2.3 T-ALL classification

Broadly, T-ALL is classified based on the oncogenic TF expression or immunophenotypes, both of which reflect the T-cell maturation arrest stages. In particular, both arrays and sequencing-based gene expression studies have observed sample grouping according to the expression of oncogenic TFs with subtype-specific gene expression signatures (Ferrando et al. 2002; Chen

et al. 2018b). TF-based oncogenic subgroups also form distinct risk categories that reflect the therapeutic response and clinical outcomes **(Figure 4)**.

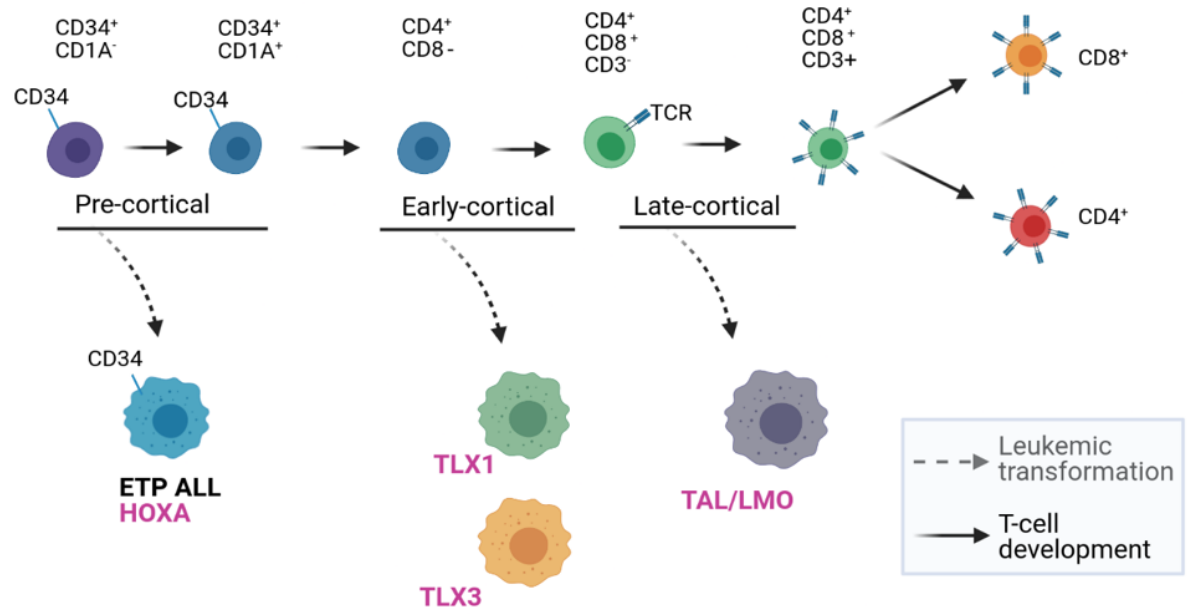

Figure 4. Gene expression-based distinct molecular subgroups of T-ALL.

Early T-cell progenitor (ETP) ALLs arise much early during the T-cell development and show the expression of CD34. ETP ALLs carry genetic aberrations leading to the expression of HOXA. Early cortical T-ALL does not show expression of CD34 and express TLX1/TLX3. Late cortical T-ALL shows high TCR and express bHLH family TF such as TAL and LMO genes. Created with BioRender.com

T-ALLs with early maturation arrest during T-cell development are often referred to as ETP-ALL and are known to be clinically aggressive (Coustan-Smith et al. 2009). ETP-ALLs show gene expression signature similar to myeloid progenitors and often carry less frequent mutations in *NOTCH1*. Besides, mutations in myeloid-like genes such as *DNMT3A, RUNX, RAS, IDH* are found in ETP-ALL (Van Vlierberghe et al. 2013). On the contrary, T-ALLs with maturation arrest during the early cortical stage of T-cell development show overexpression of TLX proteins (*TLX1/TLX3*) and are clinically responsive. Studies have indicated that the overall survival of TLX deregulated T-ALLs is significantly higher than the rest of the subtypes (De Keersmaecker and Ferrando 2011). Finally, T-ALLs with maturation arrest stage during the late cortical stage of T-cell development include overexpression of bHLH family TFs - *TAL1, LMO1/LMO2*.

Genomically, this subtype harbors enriched mutations in the *PTEN* tumor suppressor gene and is known to be clinically aggressive (D'Angio et al. 2015).

In addition to gene expression-based subtype classification, T-ALLs have also formed distinct groups according to their DNAm profile. An array-based study measuring DNAm levels of 27,000 CpG sites across a cohort of T-ALLs, classified the disease into two groups, namely $CIMP^+$ and $CIMP^-$ (CIMP: CpG Island Methylation Phenotype) (Haider et al. 2019). According to the classification, $CIMP^+$ T-ALLs show hypermethylation in CpG rich promoter regions whereas, $CIMP^-$ T-ALL showed hypomethylation. Moreover, $CIMP^-$ T-ALLs were associated with the worst survival, thereby acting as a biomarker for T-ALL risk stratification (Kraszewska et al. 2012).

1.2.4 Pediatric T-ALL v/s Adult T-ALL

Significant differences between pediatric and young adult T-ALL arise in the overall frequency of the disease itself, alterations in epigenetic mutations, and TLX positive incidences. Importantly, clinical outcomes associated with adult T-ALL are significantly worse when compared to the pediatric counterpart (80% five years overall survival for pediatric v/s 50% for young adult) (Hunger and Mullighan 2015; Litzow and Ferrando 2015).

Genetically, young adult T-ALLs harbor a significantly higher fraction of mutations in epigenetic factors (*DNMT3A, IDH1/2*) and *JAK1/JAK3* (Liu et al.). *DNMT3A/IDH* mutations are most common among myeloid malignancies (Ley et al. 2010). Studies have speculated possible age-related, clonal hematopoiesis-associated mutations based on their allelic frequencies and elderly age group associated with *DNMT3A* mutant T-ALLs (Bond et al. 2019). Moreover, these patients belong to ETP-ALL and display shorter event-free survivals. In addition to the differences in mutational patterns, adult T-ALLs harbor a significantly higher proportion of *TLX1* positive T-ALLs whereas, *TLX3* positive T-ALLs are more frequent in pediatric T-ALLs. Considering these differences, adult T-ALL forms a distinct subtype with a characteristic molecular profile and clinical outcomes **(Table 1).**

Oncogenic events	Adult T-ALL	Pediatric T-ALL
TLX1 overexpression	30%	5-10%
TLX3 overexpression	5%	20-25%
DNMT3A mutations	10%	0%
IDH1/IDH2 mutations	5-7%	0%
JAK1/JAK3 mutations	18-30%	<3%

Table 1. Significant genetic differences between pediatric and adult T-ALL

1.3 Epigenetics

Epigenetics, in broad terms, can be defined as the study of observable changes (e.g., in gene expression) that are not due to the direct consequences of genetic changes such as single nucleotide polymorphisms. Often epigenetics involves modifications to the DNA or to the histones around which it is wrapped. Epigenetic changes are heritable and sensitive to environmental changes (Weinhold 2006). Epigenetic modifications can be broadly categorized into three forms: DNA methylation, post-translational modifications to the histones such as acetylation, methylation, ubiquitination, and the chromatin organization itself **(Figure 5)**. All of these modifications greatly influence gene expression and play a significant role in providing cellular identity.

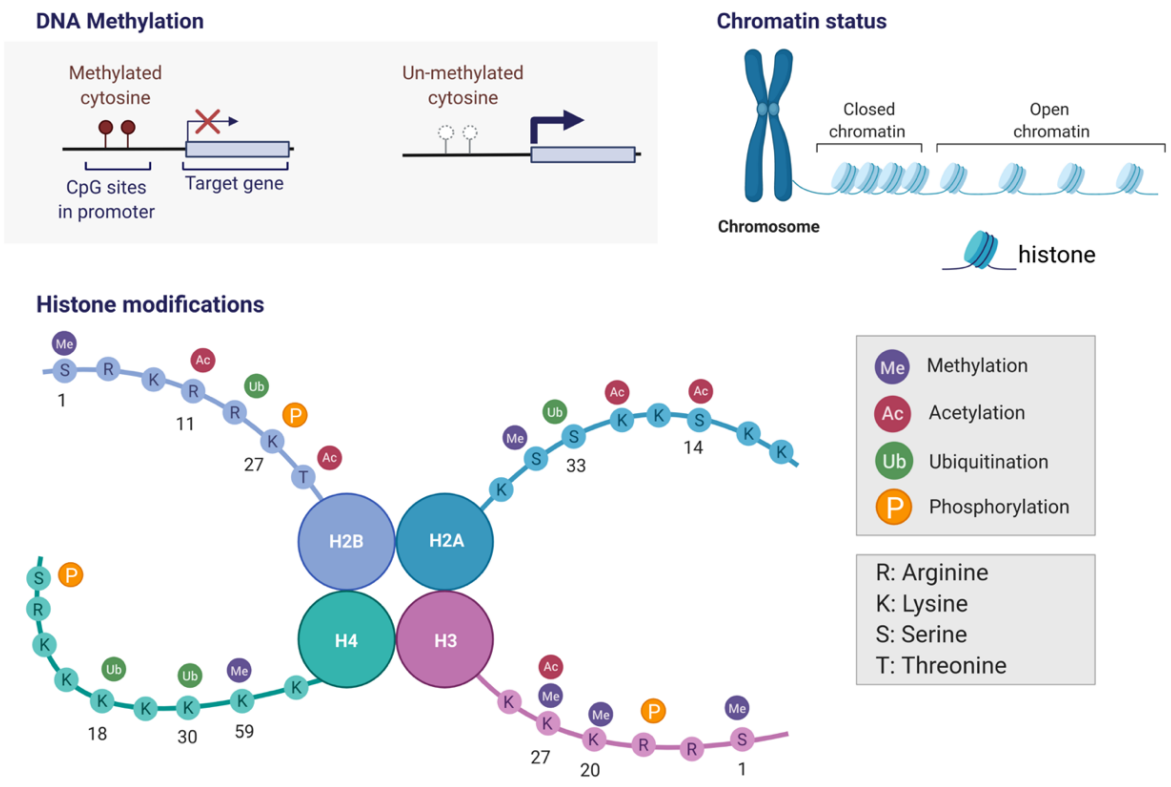

Figure 5. Epigenetic modifications.

DNA methylation involves the addition of methyl groups to the cytosine residues, the levels of which correlate with gene expression. DNA is tightly coiled around a histone, which, when densely packed, leads to a condensed form called heterochromatin (closed state). On the contrary, when histones are loosely separated, they form euchromatin (open state). Open chromatin is often occupied by transcription factors or RNA polymerases controlling the gene expression of target genes. Histone modifications involve post-translational modifications to the histone tail, such as methylation, acetylation, ubiquitination, and phosphorylation. These modifications broadly mark the regulatory regions which control the gene expression. Created with BioRender.com

1.3.1 DNA methylation

Of all the three epigenetic modifications, DNAm is the simplest to study. DNAm involves a covalent addition of a methyl group to the 5th carbon of the cytosine nucleotide base resulting in 5mC. DNA methylation is often associated with transcriptional repression and acts as a proxy for transcriptional activity (Watt and Molloy 1988). DNAm is also found to be heritable and is copied to the daughter cells upon cell division (Greenberg and Bourc'his 2019). In addition to transcriptional regulation, DNAm plays a diverse role in developmental biology and disease development (Greenberg and Bourc'his 2019).

Several enzymes of the DNA methyltransferase (DNMTs) family facilitate the addition of de-novo (by *DNMT3A, DNMT3B*) and copying of existing methyl groups (by *DNMT1*) whereas, TET (*TET1, TET2, TET3*) enzymes play a crucial role in the erasure of existing methyl groups (Li and Zhang 2014) **(Figure 6)**. DNAm prominently occurs at cytosines preceded by the guanine base in 5'-3' direction - known as CpG sites. Although CpG sites occur throughout the genome, selected genomic regions such as promoters harbor a cluster of CpG sites known as CpG Islands (CGI). CGIs show a lack of DNAm compared to the rest of the genome and significantly correlate with the gene expression (Jeong et al. 2014). Similarly, several studies have also found a lack of DNAm of CpG sites in non-coding enhancer regions, which is interpreted and controls the binding of TFs, thereby regulating the target gene expression (Angeloni and Bogdanovic 2019). Overall, DNAm plays a crucial role in suppressing non-lineage-specific genes, thereby maintaining cellular identity.

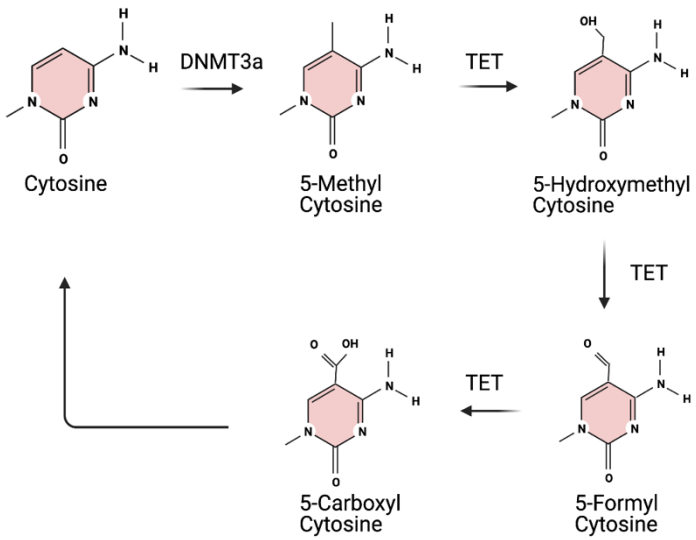

Figure 6. The DNA methylation pathway.

The DNA methyltransferases (DNMTs) add a methyl group to the 5th carbon of cytosine resulting in 5-methylcytosine (5-mC). 5mC can be oxidized to 5-hydroxymethylCytosine (5-hmC) by the TET family of enzymes (TET1/2/3) which is iteratively converted to unmethylated cytosine. Created with BioRender.com

1.3.2 Mutations of cytosine modifiers in leukemia

Feinberg et al. argue that cancers arising from distinct cells of origin are more similar since they all harbor unstable and disturbed epigenome (Feinberg et al. 2016). The inconsistent

epigenome theory is backed by the vast number of mutations in the genes associated with chromatin remodelers and cytosine modifiers in solid tumors and leukemias (Baylin and Jones 2016). Moreover, Leukemias carry a significant portion of mutations in cytosine modifiers *DNMT3A, TET*, and *IDH*, thereby affecting the DNAm pathway **(Figure 7)**. Myeloid leukemias, in particular, harbor mutations in all three before-mentioned genes whereas, T-ALLs show mutations in *DNMT3A* and IDH genes (Cancer Genome Atlas Research 2013; Liu et al. 2017). In both the leukemias, mutations in these genes are associated with a poor prognosis and form distinct risk groups.

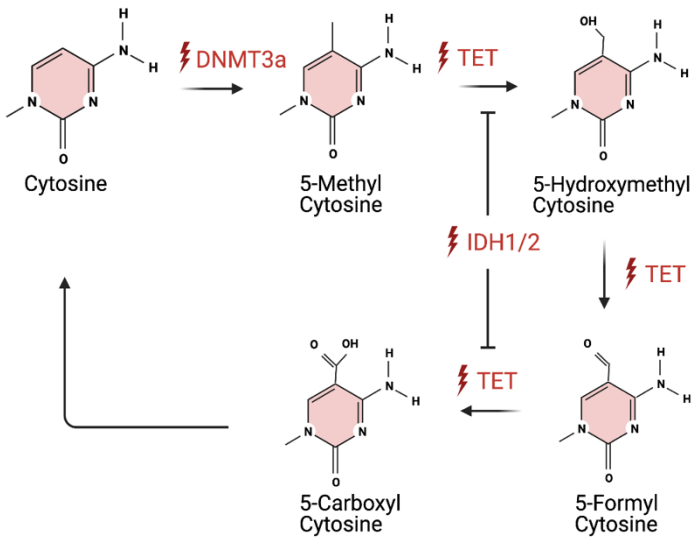

Figure 7. Deregulation of DNA methylation pathway.

Mutations in cytosine modifiers such as DNMT3A and TET enzymes can dysregulate standard DNAm pathways. The oncometabolite 2-hydroxygluterate produced by IDH1/2 mutants can indirectly affect DNAm by inhibiting a-ketoglutarate essential for TET enzymatic activity. Created with BioRender.com

DNMT3A belongs to the DNA methyltransferase family consisting of *DNMT31, DNMT3A, DNMT3B*, and *DNMT3L*. Of the four DNMTs, *DNMT3A* is mutated in up to 20-25% of myeloid/adult T-cell leukemias and is often associated with poor clinical outcomes (Cancer Genome Atlas Research et al. 2013). The mutational profile of *DNMT3A* includes a well-known hotspot R882 which accounts for ~60% and ~20% of all variants in AML and T-ALL, respectively (Ley et al. 2010; Grossmann et al. 2013). Immune compromised mice models transferred with *Dnmt3a* deficient HSCs develop myeloid/lymphoid leukemia with characteristic secondary mutations and methylation patterns (Mayle et al. 2015). Moreover, elderly healthy individuals

harbor *DNMT3A* mutations without leukemia's typical symptoms, resulting in a condition now widely known as clonal hematopoiesis (Jaiswal et al. 2014). These observations suggest that *DNMT3A* mutations predispose HSCs for malignant transformation.

Functionally, *TET2* (Ten Eleven Translocation-2) behave opposite to that of DNMTs and are involved in catalyzing methylated cytosine (5mC) to 5-hydroxymethylcytosine (5hmC), which is converted back to unmethylated cytosine. *TET2* mutations are found in ca. 10% of leukemias, and unlike *DNMT3A*, *TET2* mutations lack any functional hotspots. Phenotypically, *TET2* mutants have decreased 5hmC (Ko et al. 2010; Ley et al. 2010). Mice models with *Tet2* deletions resulted in increased HSC renewal and progressive development of proliferating myeloid malignancies (Moran-Crusio et al. 2011).

IDH1/2 (Iso-citrate dehydrogenase) genes are key players in the citric-acid cycle and convert isocitrate to a-ketoglutarate. *IDH1/2* are mutated in ca. 10% of leukemias and are mutated mutually exclusively (Cancer Genome Atlas Research et al. 2013). Similar to *DNMT3A*, *IDH1* and *IDH2* genes carry functional hotspots at R132(H) and R140(Q), respectively. Mutations in IDH result in the accumulation of oncometabolite 2-hydroxygluterate (2HG), disrupting the pathways dependent on the α-ketoglutarate through competitive inhibition. Moreover, 2HG can inhibit TET enzymes from de-methylating the 5mC, resulting in hypermethylation phenotype in AML (Figueroa et al. 2010a).

Altogether, it is abundantly clear that leukemias' epigenome, in particular, is highly unstable with preferential mutations in cytosine modifiers which affects the DNAm, and thereby possible gene expression.

1.3.3 DNA methylation as a biomarker and targets for therapy

Given the critical role of DNAm in maintaining the cellular identity and epigenomic stability, several studies have utilized DNAm as a marker to characterize multiple cancers (Figueroa et al. 2010b; Capper et al. 2018). Moreover, the relative stability of DNAm makes it a stable epigenetic marker in studying relatively older samples preserved in paraffin-embedded tissues.

Figueroa et al. analyzed DNAm from over 100 AML patients. They found that a seemingly homogenous cohort consists of 14 distinct clusters correlating with the specific genetic

aberrations (Figueroa et al. 2010b). Similarly, Nordlund et al. analyzed DNAm data from over 750 pediatric ALL samples and found characteristic methylation patterns associated with several genomic regions (Nordlund et al. 2013). The cancer genome atlas (TCGA) program also houses DNAm data for over 10,000 tumors from 30 distinct tumor types. Recently, (Capper et al. 2018) performed DNAm analysis over a cohort of 10,000 glioblastomas, which now serves as a classifier to diagnose and predict tumor type. In addition to using DNAm in disease classification, DNAm has been used to discover cancer risk prediction biomarkers. For example, Borssen et al. performed DNAm analysis of 43 pediatric T-ALL samples and classified the disease into two blocks based on the methylation levels of CpG islands (Borssen et al. 2013). Of the two, a subtype with hypermethylation in CGI regions - termed as CpG Island Methylator Phenotype (CIMP) showed a favorable prognosis whereas, CIMP negative samples were associated with poor clinical outcome. Similar attempts are made in aggressive melanoma in predicting the overall survival (Guo et al. 2019).

Due to its reversible nature and aberrant levels under disease conditions, DNAm forms an attractive candidate for targeted therapies. Accordingly, several drugs are used in the clinical setting to treat multiple leukemia subtypes **(Table 2)**. 5-Azacytidine and Decitabine - two of the well-known drugs that target and act as DNA hypomethylating agents, bringing down the tumorigenesis potential (Gardin and Dombret 2017). DNA hypomethylating agents are primarily beneficial in treating elderly patients (>60 years old) classified as unfit for intense chemotherapies (DiNardo and Wei 2020). Moreover, several in-vivo and clinical studies have demonstrated the benefits of hypomethylating agents among specific subgroups of hypermethylated leukemias.

Drug	target	Interaction type	Reference
Decitabine	*DNMT3A/TET2/ IDH1/TET2*	Inhibitor	(Metzeler et al. 2012)
Procaine	*DNMT3A*	Inhibitor	(Li et al. 2018)
Azacitidine	*DNMT3A/TET2*	Inhibitor	NA
Ivosidenib	*IDH1*	Inhibitor	(Rohle et al. 2013; 2015; DiNardo et al. 2018; Lowery et al. 2019)
Enasidenib	*IDH2*	Inhibitor	(Medeiros et al. 2017; Stein et al. 2017; Yen et al. 2017)
Venetoclax	*IDH1/IDH2*	NA	(Konopleva et al. 2016; Huemer et al. 2019)

Table 2. FDA approved epigenetic drugs potentially targeting cytosine modifiers (direct or indirect)

1.3.4 Assays for measuring DNA methylation

Given the importance of DNAm in cancer and its application in clinical settings, several assays have been developed to quantify the methylation changes. Among the available options **(Table-3)**, the three primary assays include:

1. Bisulfite conversion of cytosine to uracil followed by,
    a. Array-based methylation quantification
    b. Massive parallel sequencing
2. Nanopore mediated single-molecule DNA sequencing

Treating DNA with sodium-bisulfite converts unmethylated cytosines to uracil whereas, methylated cytosines remain unchanged. The methylation status from the modified bases can be measured by high throughput assays such as DNAm arrays or whole-genome bisulfite sequencing (WGBS). Array-based assays have been historical of popular choice, and can measure DNAm of hundreds of thousands of CpG sites. Illumina Infinium arrays (Illumina Inc, San Diego, CA) employ two loci-specific probes, namely, M-probe for methylated loci and U-probe for unmethylated loci. The DNAm level is quantified by the ratio of fluorescence signal emitted by the two probes. Earlier versions of arrays measured DNAm levels of ca. 450K sites primarily located at CpG rich promoters (Illumina Human Methylation 450K; also known as 450K arrays) whereas, recent updates have increased the coverage to ca. 850K arrays

targeting the regions of known enhancer elements (Infinium MetthylationEPIC; also known as EPIC arrays).

Although arrays provide a cost-effective solution for DNAm analysis, they are hindered by the limited coverage. Human DNA, for example, contains ca. 28 million CpG sites distributed across promoters and regulatory elements. EPIC arrays, however, only target ca. 1.6% of the same, thereby missing critical information. These limitations are overcome by massively parallel sequencing approaches such as WGBS, which can measure genome-wide DNAm level at a base-pair resolution (Lister et al. 2009). WGBS can provide the DNAm status of regulatory elements such as enhancers and repressors in greater detail and allows robust integration with gene expression. Compared to arrays, WGBS is expensive, and the downstream analysis is computationally intensive due to the massive coverage.

In contrast, the second approach involving third generation *nanopore* sequencing allows low cost and rapid quantification of DNAm (Jain et al. 2018). Moreover, nanopore sequencing allows direct detection of methylated cytosines from naive DNA and avoids complex and expensive bisulfite conversion (Rand et al. 2017; Simpson et al. 2017). However, current nanopore sequencing still suffers from high error rates (~10%) and accurate discrimination of ionic currents from cytosine and 5mC is an active area of on-going research (Schatz 2017).

	Array-based	**Second gen. sequencing based**	**Third gen. sequencing based**
C and 5mC discrimination	Bisulfite treatment	Bisulfite treatment	electrolytic current signals from naïve DNA
Platforms	Infinium HM450K, Infinium EPIC	Illumina short read sequencer	Oxford Nanopore MinION
Library prep.	-	WGBS, RRBS, PBAT	-
Cost (Relative to array-based)	*	WGBS (***)/ RRBS (**)	*
Coverage^	Infinium HM450K (ca. 1.5%), Infinium EPIC (ca. 3%)	WGBS (100%), RRBS (5-10%)	100%
Computational analysis	Easy	Complex	Complex
Accuracy	High	High	Low

^relative to human methylome consisting of ca. 28 million CpGs.; **C** = Cytosine; **5mC** = methylated cytosine; **WGBS** = whole genome bisulfite sequencing; **RRBS** = Reduced representation bisulfite sequencing; **PBAT** = Post bisulfite adapter tagging

Table 3. Comparison of assays for quantifying DNA methylation.

2 AIMS OF THE THESIS

Understanding the role of DNAm in thymopoiesis provides insights into epigenetic modulation during T-cell development. Like hematopoiesis, thymopoiesis involves progenitor cells undergoing multi-step uni-directional differentiation resulting in mature and fully functional thymocytes. Several studies have already characterized the role of DNAm in hematopoiesis using both array-based and sequencing-based approaches (Bock et al. 2012). However, such attempts have been sobering for thymopoiesis and lack extensive analysis (Rodriguez et al. 2015). Similarly, in leukemia, a significant fraction of either transcriptional or DNAm published research has mainly focused on pediatric T-ALL (Borssen et al. 2013; Liu et al. 2017). It is now clear that adult T-ALL differs significantly from pediatric T-ALL at the molecular and clinical outcomes. Extensive characterization of adult T-ALL is lacking to date and deserves much-needed attention. The current thesis addresses these two issues by utilizing multiple omics data generated for distinct intra-thymic cell types and a rare adult T-ALL cohort.

2.1 DNA methylation dynamics of human αβ T-cell development

DNAm provides an epigenetic history of cellular development and is inherited to the daughter cells. These epigenetic signatures are retained post differentiation and can be traced back to the parental cell of origin. Using DNAm as an epigenetic signature, here we address the following questions:

1. Changes in global methylation levels during thymopoiesis
2. Characteristics of regulatory regions associated with the differentiation
3. Comparative analysis with the hematopoiesis

We utilize WGBS to measure DNAm levels of seven distinct intra-thymic cell types. To facilitate the analysis of the complex data types resulting from WGBS, we describe a software framework called *methrix*. Results from the study were used to perform a comparative analysis between intra-thymic cell types and the hematopoietic cells. Finally, we describe the thymus lineage-specific, developmentally associated genomic regions, which constitute as an *epigenetic atlas* for thymopoiesis.

2.2 Epigenetic blueprint of adult T-ALL

Adult T-ALLs make up to 30% of all ALLs among adults, and much is unknown about the epigenetic profiles governing T-ALL subtypes and their association with the clinical outcome. Studies in pediatric ALL have shown hypo methylated groups to be associated with poor clinical outcomes (Borssen et al. 2013). Similarly, gene expression data has identified distinct transcriptional subgroups (Liu et al. 2017). Here, we use DNAm data to further characterize adult T-ALL by addressing the following issues:

1) Identification of distinct epigenetic T-ALL subtypes
2) Somatic landscapes of adult T-ALL
3) Enhancer landscapes of T-ALL
4) Deciphering the maturation arrest stages of T-ALL subgroups

To achieve these, we employ Illumina Infinium EPIC arrays to measure DNAm from a cohort of well-characterized adult T-ALLs. Using the DANm profiles, we use a robust clustering strategy to describe epigenetic subtypes. Subtypes are further characterized using mutational profiles, histone modifications, and gene expression. Finally, we integrate normal thymic developmental-associated genomic regions to describe maturation arrest stages of T-ALL subtypes.

3 RESULTS

3.1 DNA methylation dynamics of human αβ T-cell development

The hematopoietic system is responsible for the generation of myeloid and lymphoid cells that make up the entire lymphatic system. Although hematopoiesis gives rise to lymphoid progenitor cells, the formation of mature T-lymphocytes occurs within the thymus – a process known as *thymopoiesis*. Handful of studies have attempted to gauge the DNAm changes during thymopoiesis using either array-based (Rodriguez et al. 2015) or sequencing-based approaches (Cieslak et al. 2020). However, they have suffered from limitations such as restricted coverage of arrays (ca. 3% of DNA methylome) or by the lack of biological replicates. Furthermore, comparative analysis of intrathymic cell types with lymphatic T-cells and BM-derived progenitor cells is lacking and can provide the thymic signatures necessary for cellular identity.

To map the DNAm dynamics of thymopoiesis, we applied WGBS to generate a genome-wide DNAm map of seven distinct intrathymic cell types from multiple neo-natal thymi collected from patients undergoing thymectomy. We compared the results to BM and peripheral cell types and defined the *thymic signature,* which can recapitulate the thymopoiesis differentiation trajectory in an independent dataset.

3.1.1 DNA methylation map of intra-thymic cell types

Using Fluorescence-activated cell sorting (FACS), we isolated seven distinct intrathymic cell types from multiple thymi collected from neonates undergoing cardiac surgery. The cell types include; immature early thymic CD34+ precursors (CD34+ CD1A-, CD34+ CD1A+), immature single CD4+ cells (ISP CD4+), early cortical cell types with low TCR expression (CD4+ CD8+ CD3-), late cortical cells with high TCR expression (CD4+ CD8+ CD3+), and finally mature single CD4+ and CD8+ positive cells **(Figure 8A) (Table 5).**

The isolated cell types represent T-cell differentiation's hierarchical organization and provide a valuable resource to study the dynamics of DNAm governing the thymopoiesis. Following WGBS using the SWIFT protocol, the initial analysis revealed a non-significant yet gradual loss of methylation along with the T-cell differentiation. Loss of DNAm is prominent post-beta-selection leading to the TCR expression **(Figure 8B).** To further analyze the DNAm at

regulatory regions, we used publicly available and well-defined BLUEPRINT regulatory regions (Zerbino et al. 2015). BLUEPRINT regulatory build is an effort from Ensembl genome browser which attempts to summarize regulatory regions in human genome by analyzing publicly available datasets encompassing epigenetic marks and transcription factor binding sites (TFBS). It consists of ca. 500,000 genomic loci classified into six distinct categories: CTCF, distal enhancers, TFBS, open chromatin regions, promotes and its flanking regions. BLUEPRINT loci's aggregated DNAm levels revealed a common hypomethylation trend among promoters, whereas the rest showed varying degrees of hypermethylation **(Figure 8C)**.

Moreover, several of the regulatory regions associated with the T-cell development genes showed a distinct DNAm pattern. For example, a CTCF binding site near the *CD34* promoter showed a gradual methylation gain (Schmitt et al. 1995). The promoter flanking region of *BCL11B* – a master TF associated with the T-cell fate decision - showed severe hypomethylation during the terminal stages of T-cell development (Kastner et al. 2010) **(Figure 8D)**.

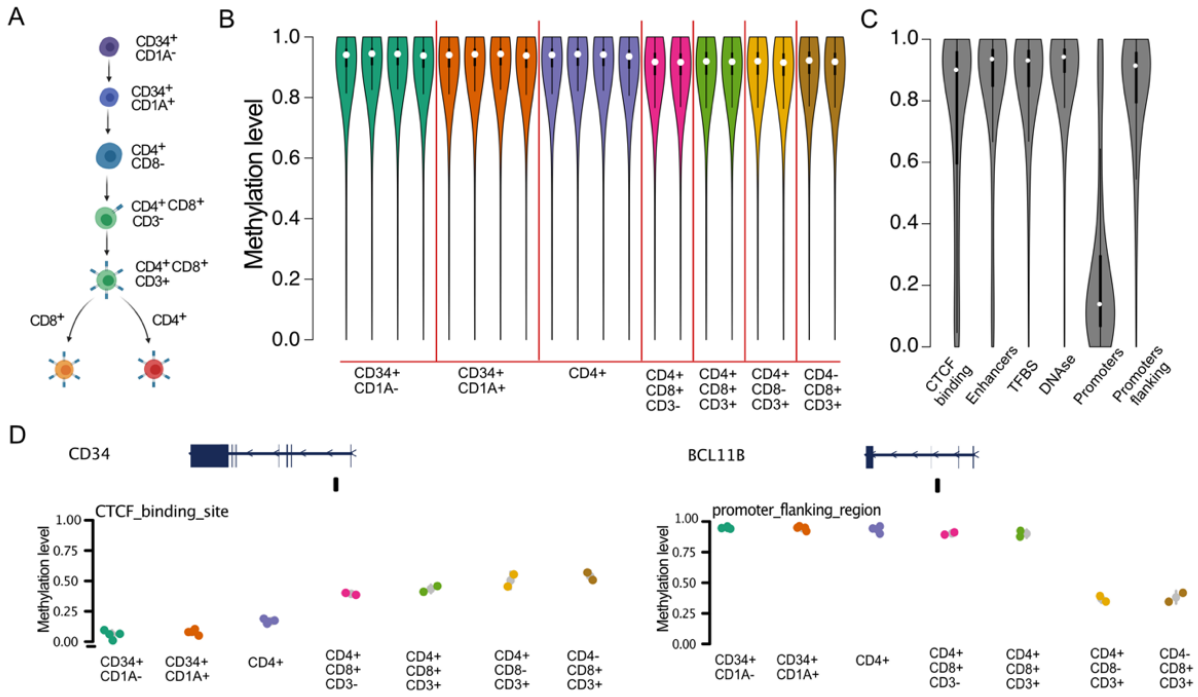

Figure 8. DNAm during human thymopoiesis.

A. *FACS sorted intrathymic cell types. Cell surface receptors are shown along with the cell types organized by maturation hierarchy. Sorting was performed by Aurore Touzart..* ***B.*** *Violin plots of average global DNAm (5-kb windows) levels for all samples color-coded by cell type.* ***C.*** *Violin plot of aggregated DNAm levels for BLUEPRINT regulatory regions.* ***D and E.*** *Example plots of DNAm levels at selected BLUEPRINT loci.*

It is now well established that the mature lymphatic cells arise from the immature progenitor cells undergoing uni-directional differentiation (Till and McCulloch 1980). The epigenetic changes during the differentiation control the gene expression programs critical for cellular identity. We compared each of the intra-thymic cells to the most immature CD34+CD1A- progenitor cells to gauge such differences in mature cells **(Figure 9A)**. De-novo differentially methylated regions (DMRs) ($P < 0.01$, $|meth| > 0.2$) identified by the comparison varied between the cell types but showed a linear trend of loss of methylation along with the differentiation. Interestingly, the number of hypomethylated DMRs post-positive-selection was significantly higher **(Figure 9B)**. To characterize the DMRs, we utilized the LOLA (Locus Overlap Enrichment Analysis) core database consisting of binding sites from over a hundred transcription factors and histone modifications arising from multiple sources (Sheffield and Bock). Enrichment results from LOLA categorized DMRs as overlapping with the TFBS strongly associated with T-cell development. Especially *NOTCH1*, *MYB*, and *RBPJ* transcription factor

binding sites dominated the DMRs, suggesting that the loss of methylation facilitates the binding of TFs critical for T-cell development **(Figure 9C)**.

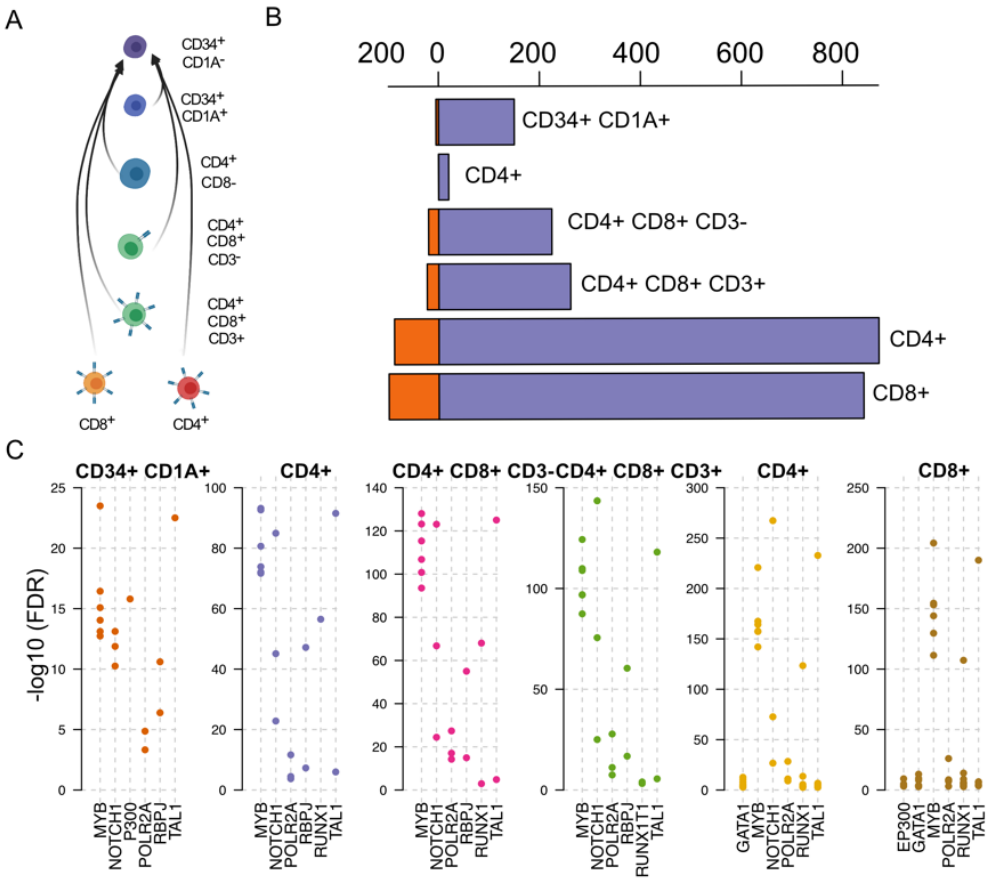

Figure 9. The progressive loss of DNAm during thymopoiesis.

A. *Outline of differential methylation analysis. Each cell type is compared to the most primitive CD34+CD1A- intra-thymic cells.* ***B.*** *Barplot of differentially methylated regions in each cell type (v/s CD34+CD1A-).* ***C.*** *Transcription factor binding sites that are enriched within the DMRs. Each dot represents a ChIP-seq dataset from LOLA core database. Y-axis threshold of two corresponds to an FDR cutoff of 0.05.*

Overall, our WGBS results highlight the progressive loss of DNAm along with the differentiation. The regions undergoing loss of DNAm harbor banding sites for core TFs necessary for proper T-cell development.

3.1.2 Dynamics of regulatory regions during thymopoiesis

Although previous results showed a sequential loss of DNAm compared to the progenitor cells, the relative changes in DNAm at each stage of differentiation were still lacking. To identify such changes, once again, we resorted to the known regulatory regions from BLUEPRINT. Restricting the analysis to BLUEPRINT regions, we compared every mature cell

type to its predecessor cell type - referred to as step-1 (CD34+CD1A- v/s CD34+CD1A+), step-2 (CD4+ISP v/s CD34+CD1A+), step-3 (CD4+CD8+CD3- v/s CD4+ISP), step-4 (CD4+CD8+CD3+ v/s CD4+CD8+CD3-), step-4a (CD4+ v/s CD4+CD8+CD3+) and step4b (CD8+ v/s CD4+CD8+CD3+) **(Figure 10A)**. Similar to the previous analysis, the analysis showed that each differentiation step is characterized by loss of methylation, and the mature terminal cells from step-4a and step-4b contained the greatest number of changes **(Figure 10A; right panel)**. Differentially methylated BLUEPRINT regions at each step contained characteristic loci associating with the genes responsible for stage transitions. For example, a promoter flanking region of *RAG1* – a gene critical for TCR rearrangement – showed massive loss of methylation at step-3, during which the beta-selection occurs (Yannoutsos et al. 2001) **(Figure 10C)**. Similarly, *ZBTB7B* responsible for CD4+ commitment shows specific hypomethylation during step-4a (Wildt et al. 2007) **(Figure 10C)**.

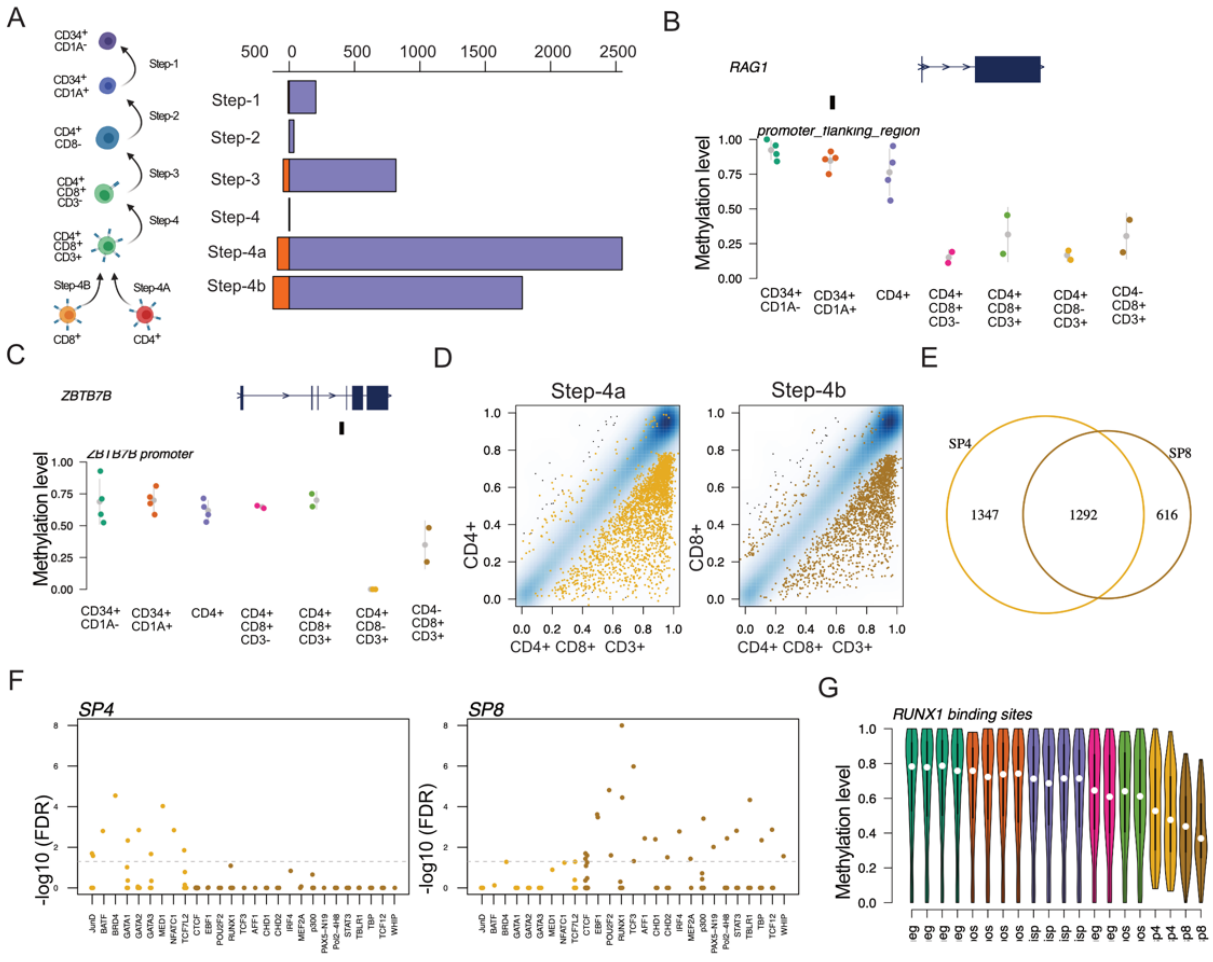

Figure 10. Stage-specific DNAm changes during T-cell differentiation.

A. *Schematic representation of differential methylation analysis. Each cell type is compared with the immediate upstream progenitor. The bar plot shows the number of differentially methylated BLUEPRINT regulatory regions.* ***B and C.*** *Example plots of DNAm levels at selected BLUEPRINT loci RAG1 (B) and ZBTB7B (C) respectively.* ***D.*** *Scatter plot of differentially methylated BLUEPRINT regulatory regions at step-4a and step-4b, respectively. Significant regions (FDR < 0.05 and |meth| > 0.2) are colored in brown.* ***E.*** *Venn diagram of differentially methylated loci from panel D.* ***F.*** *Transcription factor binding sites that are enriched within the CD4+ or CD8+ specific loci. Each dot represents a ChIP-seq dataset from the LOLA core database.* ***G.*** *Distribution of DNAm among RUNX1 binding sites across all thymic cell types.*

Final stages of T-cell development involve the differentiation of DP TCR-high cells to either single MHC class-2 restricted CD4+ or MHC class-1 restricted CD8+ cells (step-4a and step-4b). Commitment towards either cell type involves lineage-specific changes necessary to express cell-type-specific genes and silencing of the others. To characterize such differentially methylated CD4+ or CD8+ specific regions, we further analyzed step-4a and step-4b in detail. Both step-4a and step-4b contained massive methylation loss across thousands of BLUEPRINT regulatory regions **(Figure 10D)**. Overlapping the results further revealed many common

regulatory elements in addition to CD4+ and CD8+ specific regulatory regions **(Figure 10E)**. We next resorted to LOLA functional enrichment tool which uses a core database of TFBS from hundreds of publicly available datasets along with open chromatin and histone marks (Sheffield and Bock 2016). Characterizing the CD4 and CD8 specific by LOLA enrichment analysis showed a plethora of preferentially enriched transcription factors **(Figure 10F)**. For example, GATA family TFs were significantly found within the CD4+ specific hypomethylated regions, whereas *EBF1* and *RUNX1* were mainly found among CD8+ areas **(Figure 10F)**. To further validate the results, we obtained *RUNX1* binding sites from the CUTLL1 T-ALL cell line. Aggregated DNAm levels of the *RUNX1* peaks showed hypomethylation among CD8+ cells compared to CD4+ cells suggesting the critical role of RUNX1 in CD8+ fate decision.

Overall, our results show a loss of DNAm across regulatory regions necessary for T-cell differentiation and the CD4+ and CD8+ specific regulatory regions with binding sites for TFs associated with the fate decision.

3.1.3 Defining thymic developmental associated genomic regions

DNAm changes are strongly associated with organ development and embryogenesis across mammalian species. For example, the absolute necessity of deposition of de-novo, or the erasure of existing methyl groups at distinct stages of embryogenesis, guarantees the proper mammalian development from fertilization to birth (Greenberg and Bourc'his 2019). Similarly, DNAm changes govern the bone marrow hematopoiesis and, the changes can computationally reconstruct the entire hematopoietic system. Using DNAm changes, Farlik et al. characterized a core set of regulatory regions defining the hematopoiesis and predicting the specific cell type (Farlik et al. 2016). To validate if the defined hematopoietic territories apply to thymopoiesis, we obtained the publicly made available datasets of hematopoiesis and measured the dynamics of DNAm (Farlik et al. 2016). As expected, the hematopoietic regulatory regions were variable across all 16 distinct cell types, including bone marrow-derived progenitors and peripheral blood-derived mature myeloid and lymphoid cells **(Figure 11A)**. However, the same genomic regions showed no changes during the thymopoiesis, and the intrathymic cell types showed constant hypermethylation **(Figure 11B)**. These observations suggest that although thymopoiesis begins with the thymic progenitor cells arriving from bone marrow, they undergo distinct DNAm changes and are specific to thymic cell types.

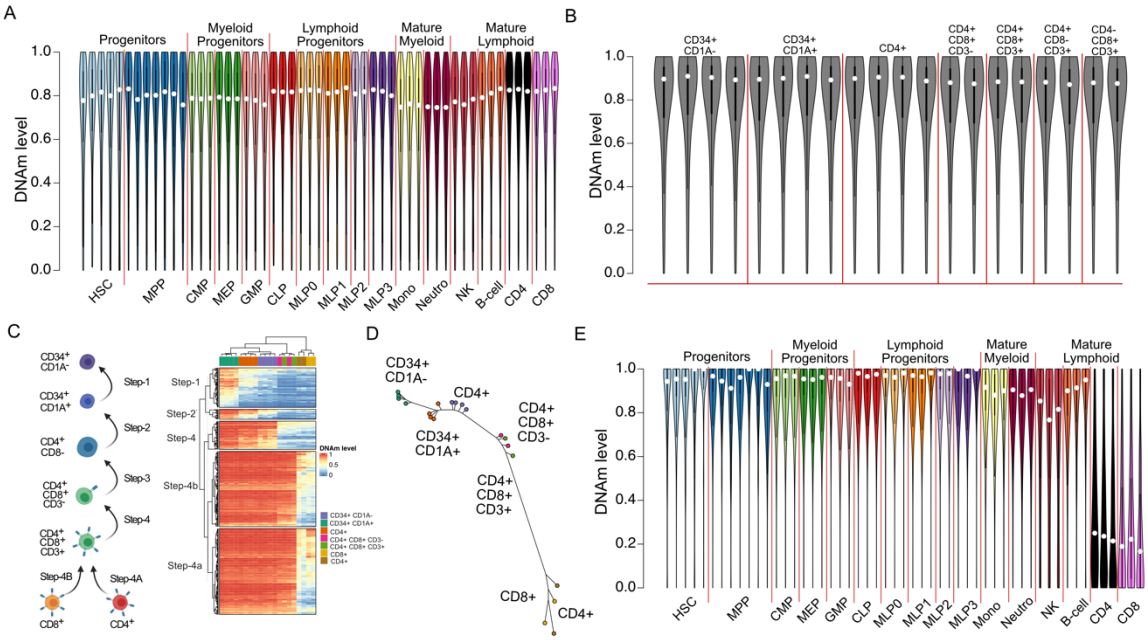

Figure 11. Developmental associated thymic DMRs.

A. *DNAm dynamics of hematopoiesis associated epigenetic signature regions.* ***B.*** *DNAm of hematopoiesis associated epigenetic signature regions in intrathymic cell types.* ***C.*** *Schematic representation of differential methylation analysis. Each cell type is compared with the immediate upstream progenitor. The heatmap shows DNAm levels of thymic DMRs at each step of differentiation (N = 381).* ***D.*** *Phylogenetic tree constructed from aggregated DNAm signals across thymic DMRs. E. DNAm of thymic DMRs in hematopoietic cell types. DMR: Differentially methylated region*

To address the missing gap, we performed a de-novo DMR analysis and identified genomic regions specific to each step of thymic differentiation. Like previous results, terminally differentiated CD4+ and CD8+ cells contained the largest number of DMRs whereas, step-4 contained no significant DMRs. Of note, almost all of the DMRs showed irreversible changes in DNAm, involving hypo-methylation of cell type-specific DMRs coupled with the silencing of the same in progenitor cells **(Figure 11C)**. The mechanism suggests the necessity of silencing genomic regions associated with the pathways that are no longer needed post-cell commitment. Besides, the entire process leaves a trail of *developmental breadcrumbs* sufficient to reconstruct the thymopoiesis **(Figure 11D).**

Moreover, analogous to previous results wherein hematopoietic signature regions are hypermethylated in thymic cell types **(Figure 11B)**, the newly defined thymic signature regions (tDMRs; *N* = 381) showed significant hypermethylation among hematopoietic cells

(Figure 11E). Primarily, BM-derived progenitor cells were significantly hypermethylated, while the mature B-cells and NK-cells showed hypomethylation. Peripheral blood-derived CD4+ and CD8+ cells, however, were strongly hypomethylated for tDMRs. The preferential hypomethylation of tDMRs in PB-derived CD4+ and CD8+ cells further highlight that although the cells are terminally differentiated and migrated to the lymphatic system, the signatures associated with the parental *cell of origin* are epigenetically imprinted and serves as cellular identity.

3.1.4 DNA methylation predicts lymphatic hierarchy

To further validate the specificity of tDMRs, we measured DNAm across six distinct intrathymic cell types using Illumina Infinium EPIC arrays targeting over 850,000 CpG sites across the human genome. The principal component analysis (PCA) analysis using the DNAm levels at hematopoietic signature regions defined by Farlik et al. failed to distinguish thymic cell types **(Figure 12A)** whereas, tDMRs accurately determined the same while maintaining the known hierarchy **(Figure 12B)**.

Next, we compiled an independent collection of cell types originating from hematopoiesis including, BM-derived progenitors and PB-derived mature cell types. The collected data from multiple sources were carefully combined with the in-house thymic arrays resulting in a total of 18 cell types **(Figure 12C, D)** (Jung et al. 2015) (Salas et al. 2018). Similarly, constructing a phylogenetic tree using the aggregated DNAm signals across hematopoietic signature regions and tDMRs, revealed the hierarchical organization of the lymphatic cell types **(Figure 12E)**. All the cell types originating from the lymphoid branch of hematopoiesis were separated from the mature PB-derived myeloid monocytes and neutrophils. Among the tree's lymphoid arm, immature intrathymic cell types were clustered at the top and were closer to the PB-derived mature NK and B-cells. Of note, intrathymic ETP cells retain the potential to de-differentiate into B- and NK-cells and, their placement near the same suggests the shared epigenetic signatures (Luc et al. 2012). Next, intrathymic mature CD4/8 cells were placed closer to the PB-derived CD4/8 cells implying that the terminally differentiated cells have a conserved epigenome regardless of the tissue microenvironment.

Overall, the results validate tDMRs in an independent dataset generated using a different platform.

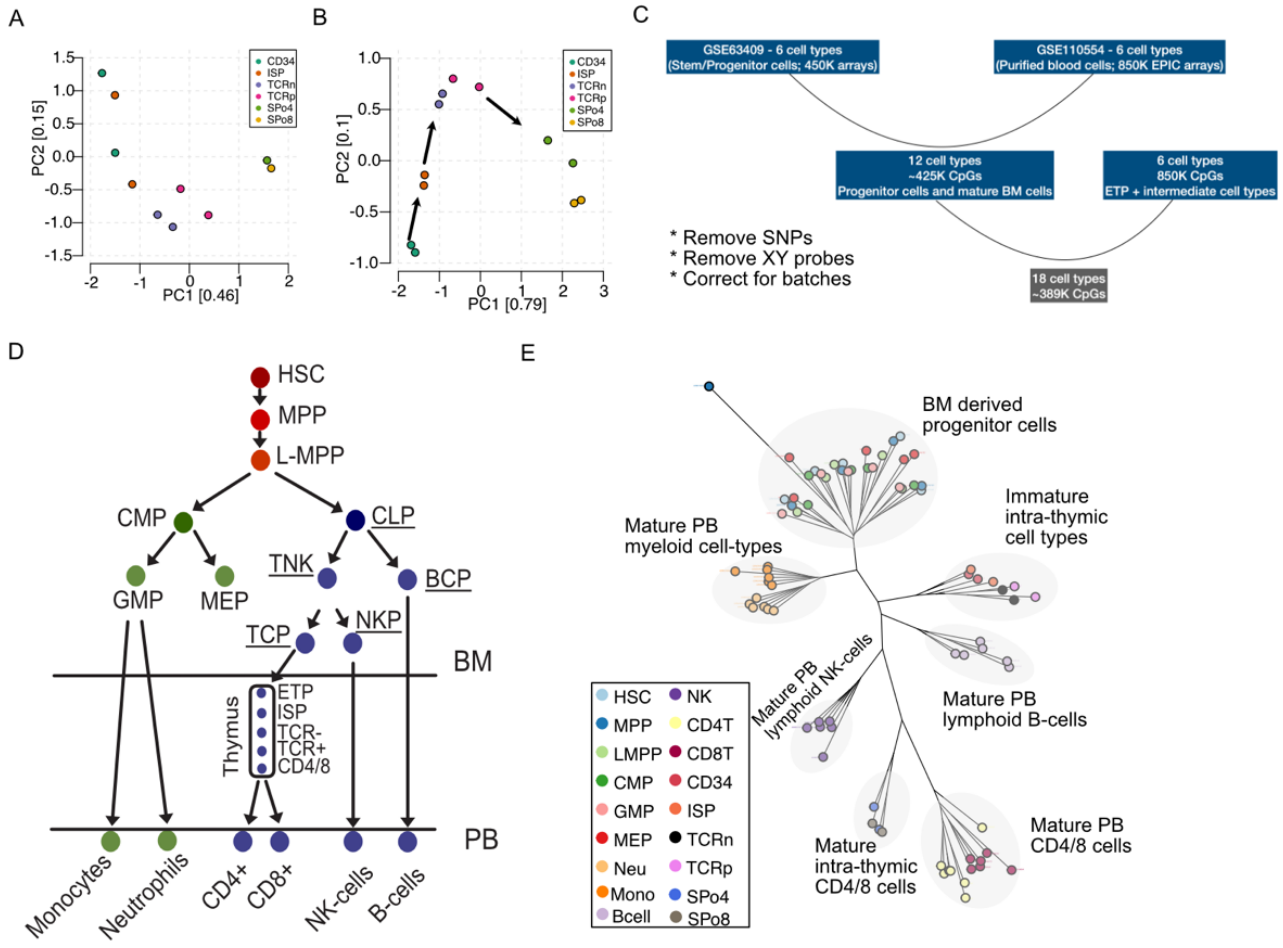

Figure 12. Validation of thymic DMRs in arrays.

***A.** PCA of intrathymic cell types using hematopoiesis associated epigenetic signature regions. **B.** PCA of intrathymic cell types using thymic DMRs. **C**. Overview of array-based public datasets used for validation. Key steps in quality control are mentioned. **D.** Overview of cell types used for validation. **E.** Hierarchical organization of lymphatic cell types predicted using DNAm levels at hematopoiesis associated epigenetic signature regions and thymic DMRs.*

3.1.5 Hypomethylation of regulatory regions is characteristic of thymopoiesis

To further characterize the association of DNAm with the transcriptional program, we collected gene expression data and histone marks associated with the active transcription. Notably, the DNAm dynamics at the promoter regions of T-cell development genes showed a significant inverse correlation with the corresponding gene expression **(Figure 13A)**. Also, genes associated with the tDMRs showed a non-significant trend towards inverse correlations **(Figure 13B)**.

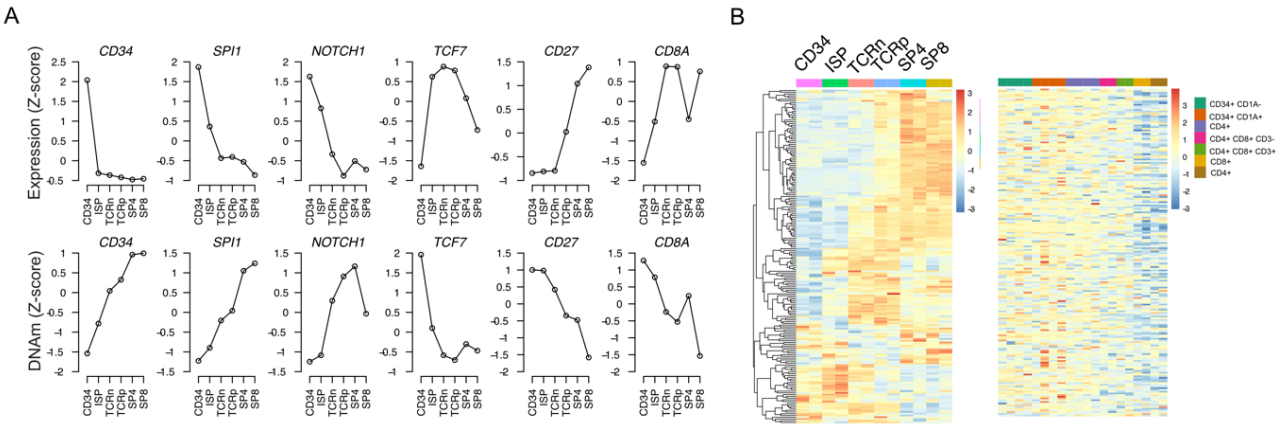

Figure 13. DNAm and gene expression during thymopoiesis.

A. *Transcriptional and promoter DNAm (TSS+/-750bp) dynamics of candidate genes involved in T-cell development.* ***B.*** *Heatmaps of gene expression and promoter DNAm of genes annotated with the thymic DMRs.*

Super enhancers (SE) are genomic regions marked by the hyperacetylation of lysine 27 on histone H3 (H3K27ac) (Pott and Lieb 2015). SEs have been identified in several normal cells and are primarily associated with the genes necessary for cellular identity (Hnisz et al. 2013). For example, genes such as *Oct4*, *Sox2*, and *Nanog* in mouse embryonic cells are all marked by SE (Whyte et al. 2013). Similarly, we utilized H3K27ac ChIP-seq datasets for five of the intrathymic cell types to define thymopoiesis's SE landscape. Results indicated a progressive increase in the number of SEs with the mature single CD8+ cells containing the most significant number **(Figure 14A)**. The same observation of the incremental number of hypomethylated DMRs during T-cell differentiation suggests a coordinated hypomethylation and enhancers' activation. Despite the significant overlap in the SE genes between differentiation steps, several SEs occurred in a cell-type-specific manner. For example, the *ERG* gene contained an SE in CD34+ ETP cells, whereas *RAG1*, a critical gene necessary for TCR rearrangements, harbored an SE during the early cortical stage of T-cell differentiation **(Figure 14B, C)**.

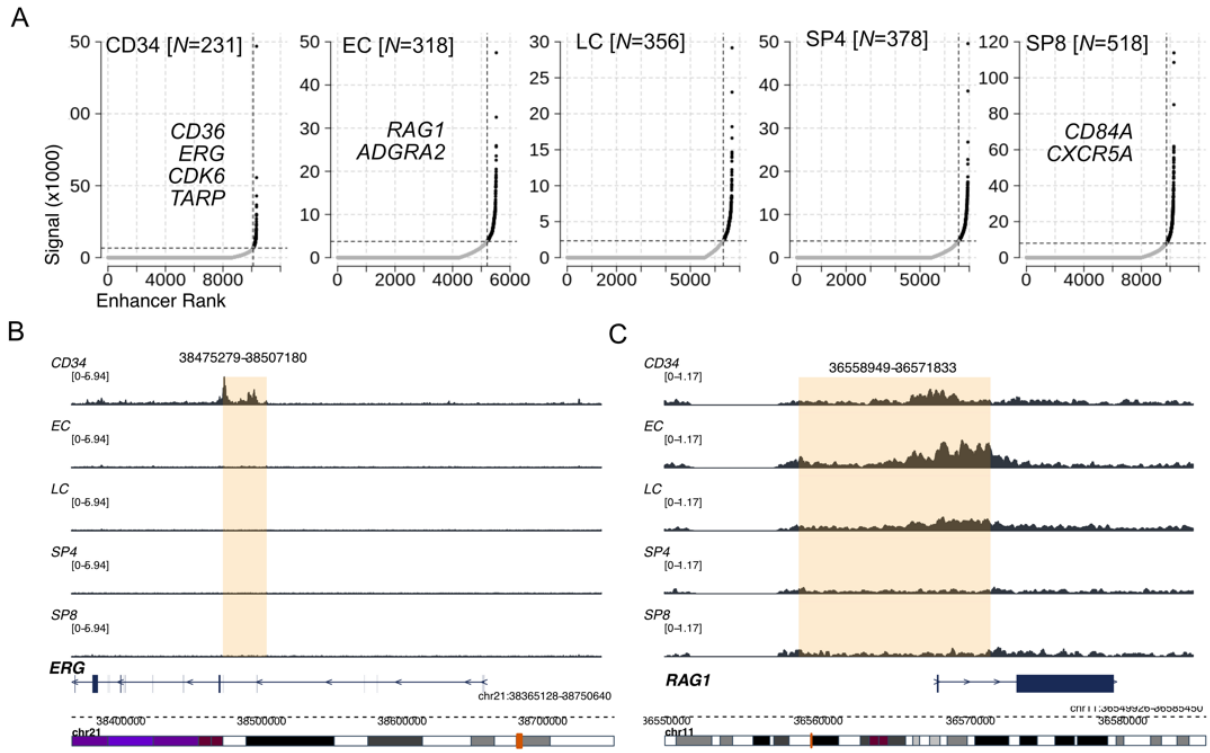

Figure 14. Identification of super-enhancers in intrathymic cell types.

A. *Hockey stick plots of super-enhancers identified in five distinct thymic cell types. Dots in black are super-enhancers, whereas typical enhancers are highlighted in gray. Essential cell-type-specific genes associating with the super-enhancers are highlighted* ***B.*** *CD34+ specific super-enhancer region near the ERG gene.* ***C.*** *RAG1 linked super-enhancer in the early cortical stage of T-cell development.*

In addition to SE analysis, we combined H3K27ac with H3K4me1 – another histone mark associated with the active transcription – and defined three distinct sets of enhancer classes, namely, active enhancers (H3K27ac+ H3K4me1+), poised enhancers (H3K27ac- H3K4me1), and Putative/Primed enhancers (H3K4me1- H3K27ac+) **(Figure 15A)**. Besides, we also compiled H3K4me3 – a promoter mark associated with active transcription. Overall, these histone marks provided a comprehensive status of regulatory regions actively regulating the transcriptional programs. Using the LOLA program, we performed an enrichment analysis to decipher the characteristics of tDMRs. Interestingly, thymic DMRs significantly overlapped with the active promoters (marked by H3K4me3 peaks) than the overall background promoter regions, suggesting that CpG dense promoters' marked by tri-methylation of the histone tail, correlates with the corresponding gene expression **(Figure 15B)**.

Moreover, thymic DMRs were located inside the active and putative enhancers, while poised enhancers showed no enrichment. The results highlight the previous observations wherein

DNAm levels discriminated the active and poised enhancers. Likewise, as expected, thymic DMRs were also found to overlap significantly with SE regions, critical for cellular identity (Bell et al. 2016).

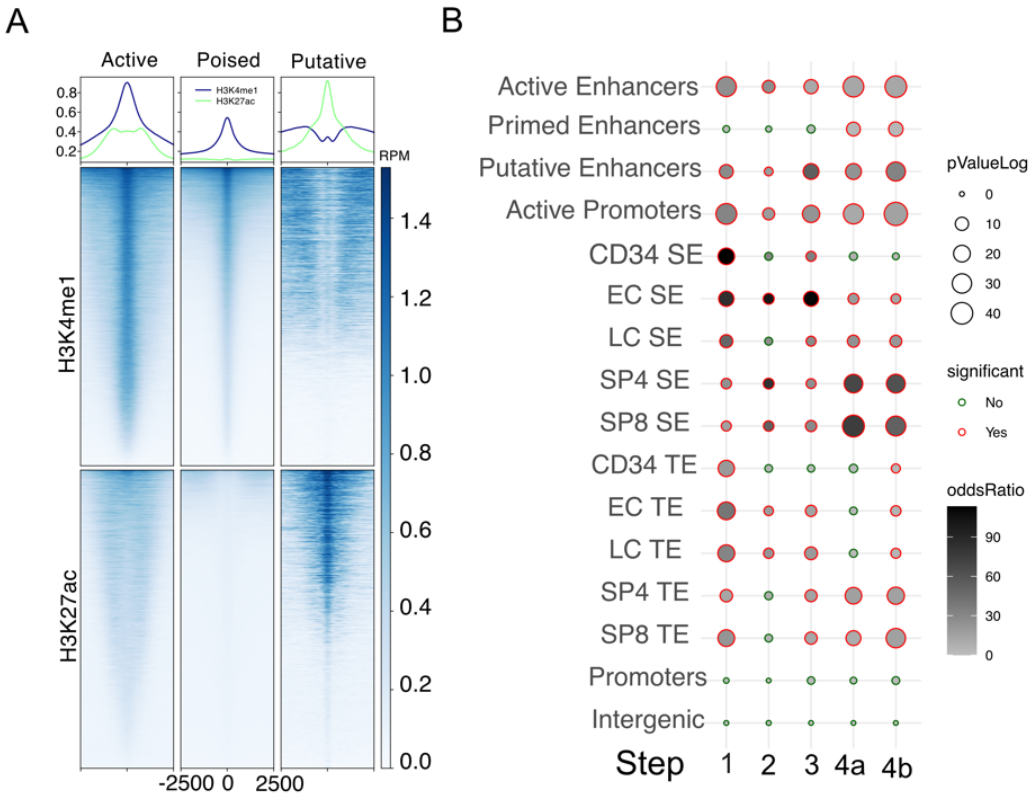

Figure 15. Hypomethylation of regulatory regions

A. *ChIP-seq binding profiles of H3K4me1 and H3K27ac histone marks for three distinct regulatory genomic regions (Active, Poised and Putative enhancers) in CD34+ thymic cells. Regulatory regions are displayed as 5 kb regions centered around the peak. The top line plot shows average signal density, whereas bottom heat maps display ChIP-seq signal for individual peaks. Color gradient reflects the thickness of the ChIP-seq signal.* ***B.*** *Dot plot for the enrichment of thymic DMRs (X-axis) in various regulatory regions (Y-axis) as highlighted in the plot's left side. Dots are color-coded for significance. The size of the dots represents P values in the log10 scale.*

Overall, our results indicate that the thymic DMRs (often hypomethylated) occur among genomic regulatory regions associating with the active promoter marks (H3K4me3) or active enhancer.

3.2 Epigenetic blueprint of adult T-ALL

To understand the role of DNAm in the origin/development of leukemia and to identify clinically relevant subgroups, we collected a cohort of 143 primary young adult T-ALLs from two French ALL cooperative groups (22 from GRALL-2003 and 121 GRALL-2005). All samples underwent assays to measure DNAm of ca. 850,000 CpG sites using Illumina Infinium Methylation EPIC BeadChips (EPICA arrays). As a control, DNAm data for six distinct intrathymic cell types are generated using EPIC arrays. Targeted sequencing of a panel of genes involved in leukemic pathogenesis was also performed. Copy number status for the same genes is obtained using SALSA MLPA P383 T-ALL probe mix (MRC-Holland, Amsterdam, Netherland). Besides, samples were validated for known T-ALL driver events such as overexpression of *TAL1*, *TLX1/3*, and *HOXA* transcription factors. Maturation arrest stages and ETP phenotypes were predicted using surface markers and TCR rearrangements. Other phenotypic and oncogenetic features are obtained as described in (Bergeron et al. 2007; Bond et al. 2016). A complete summary of cohort characteristics is provided in **Table 7**. For a subset of samples, gene expression (*N* = 48) and ChIP-sequencing (*N* = 12) of histone marks associated with the active transcription (H3K27ac, H3K4me1, H3K4me3) are also generated.

3.2.1 The somatic landscape of adult T-ALL

As expected, our results from targeted sequencing included mutations of hallmark genes in expected proportions: *NOTCH1* (66%), *FBXW7* (20%), *CDKN2A* (69%), *PHF6* (38%), *DNM2* (20%), and *BCL11B* (15%) **(Figure 16A)** (Liu et al. 2017). Restricting our analysis to potentially driver genes (mutated in >2% of the cohort) further revealed recurrently altered pathways among adult T-ALL, including Cell cycle (70%), NOTCH signaling (71%), JAK-STAT signaling (38%), and in transcription regulators (50%) **(Figure 16B)**. Most importantly, we observed an unexpectedly high frequency of alterations in epigenetic modulators (58%) and cytosine modifiers involved in the DNAm pathway (18%). Among genes involved in establishing DNA methylation patterns, *DNMT3A* was the most frequently mutated (14%), followed by *TET2* (5%), *IDH2* (4%), *TET3* (3%), and *IDH1* (1%) **(Figure 16C)**.

However, mutations in DNAm pathways have been reported in T-ALL, and their higher frequency in our cohort suggested possible cohort-specific biases. To further validate the results, we re-analyzed an independent publicly available T-ALL cohort (Chen et al. 2018b).

Dividing the cohort into a young adult (> 16 years old) and pediatric (<16 years old) recapitulated a higher frequency of DNAm variants, exclusively in young adult patients **(Figure 16D)**. Most importantly, all of the *IDH2* mutations co-occurred with *DNMT3A* suggesting their potential synergistic role for the underlying leukemogenesis.

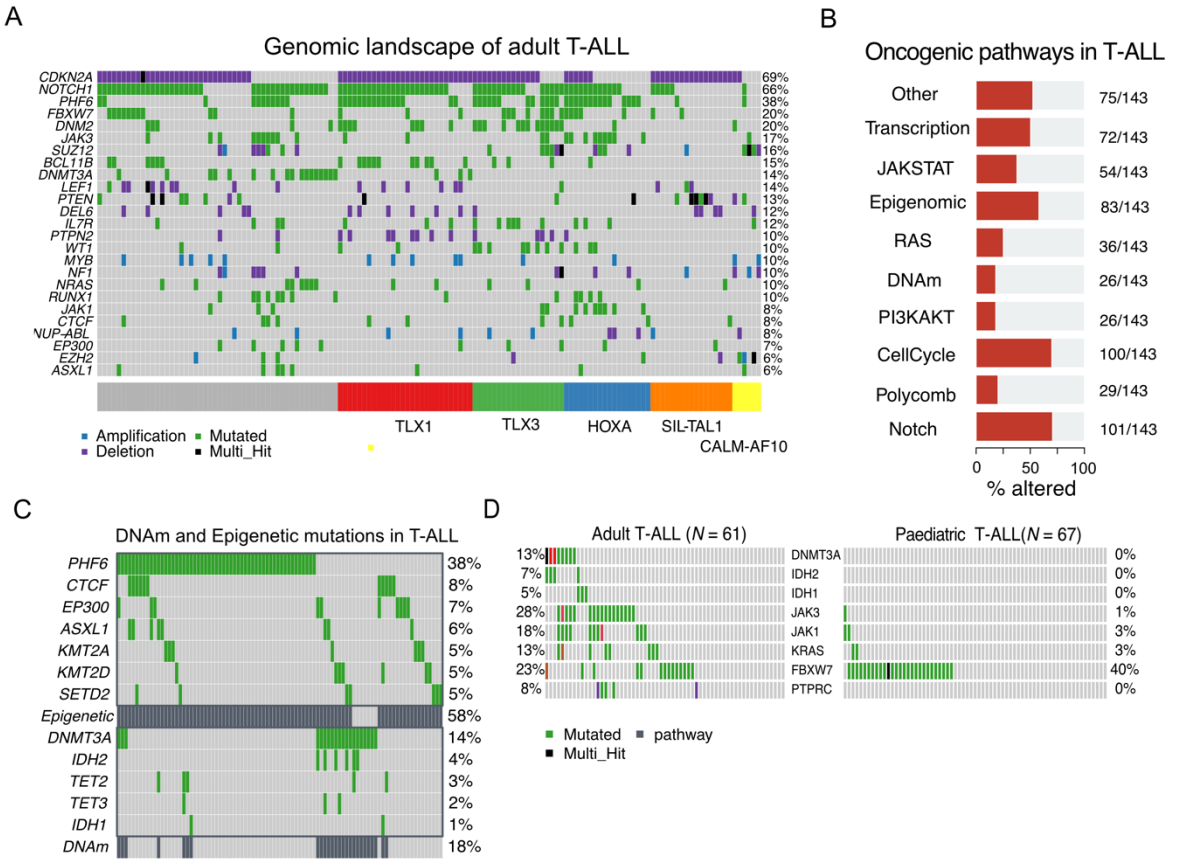

Figure 16. The somatic landscape of adult T-ALL.

***A.** Oncoplot is depicting the mutational status of genes altered in at least 5% of the cohort. The bottom annotation bar shows the deregulated oncogenic transcription factors. **B**. Barplot of recurrently mutated pathways in adult T-ALL. **C**. Mutations in epigenetic regulators and DNAm pathway. DNMT3A mutants co-occur with IDH2. **D.** Significantly differentially mutated genes between adult and pediatric T-ALL. Data obtained from (Chen et al. 2018b).*

In addition to mutational analysis, we used EPIC arrays to characterize copy number variations (CNV) in adult T-ALL. CNV results were then summarized using GISTIC to identify recurrent CNVs **(Figure 17A, B)**. Effects include frequent deletions in the 9p21.3 arm harboring the *CDKN2A* gene known to be deleted in over 70% of the T-ALL. We also observed partial deletions of 15q arm in ca. 40% of the samples associating with the poor clinical outcome (Heerema et al. 2002). Although our results show commonly found CNV patterns in ALL, EPIC arrays' resolution limits the robust CN interpretation.

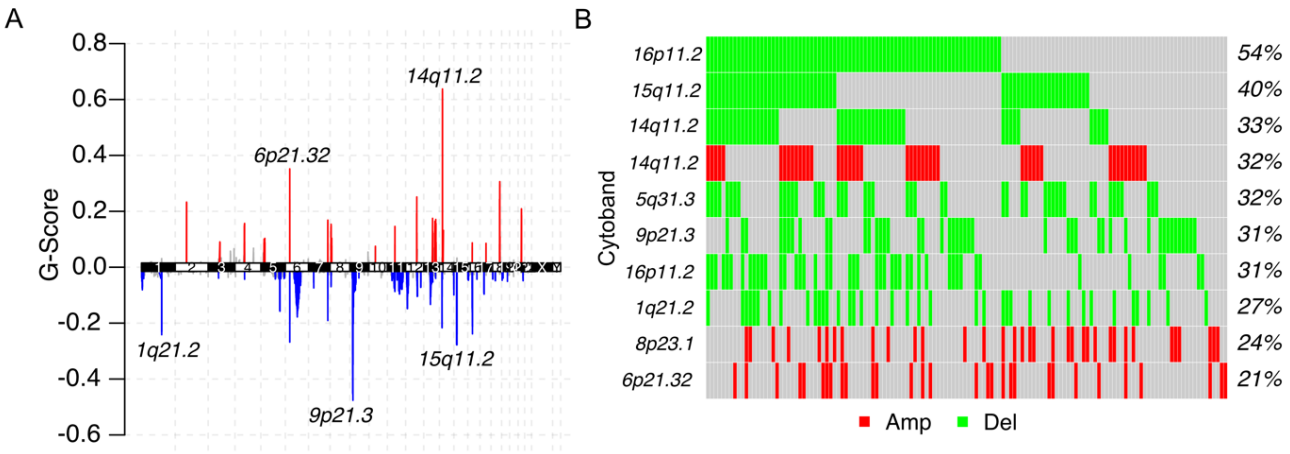

Figure 17. The copy number variations in adult T-ALL.

***A**. Summarized results of recurrent copy number alterations in adult T-ALL. Red bars indicate amplifications, and blue bars indicate deletions. The height of the bars indicates the magnitude of the alterations. Top-5 most altered cytobands are labeled. **B.** Top-10 recurrent copy number deletions and amplifications. Each bar shows a patient sample color-coded for amplification or deletion events.*

3.2.2 DNA methylation identifies distinct T-ALL subtypes

"To better understand the role of DNAm in T-ALL leukemogenesis, we performed genome-wide DNAm analysis using Illumina Infinium Methylation EPIC BeadChip (EPIC arrays) for 143 primary adult T-ALL samples, as well as for six distinct sorted normal thymic T-cell subpopulations."

By utilizing a carefully designed analysis pipeline, including data normalization, removing probes associated with SNPs and those located on CNVs, we performed clustering to identify leukemic subtypes **(Figure 18A)**. Our Non-negative matrix factorization (NMF) based analysis indicated the presence of five distinct clusters irrespective of the number of probes used **(Figure 18B)** (Gaujoux and Seoighe 2010). Moreover, these five clusters were highly stable and showed a high correlation between the results with little to no sample movement as the number of probes varied **(Figure 18C)**.

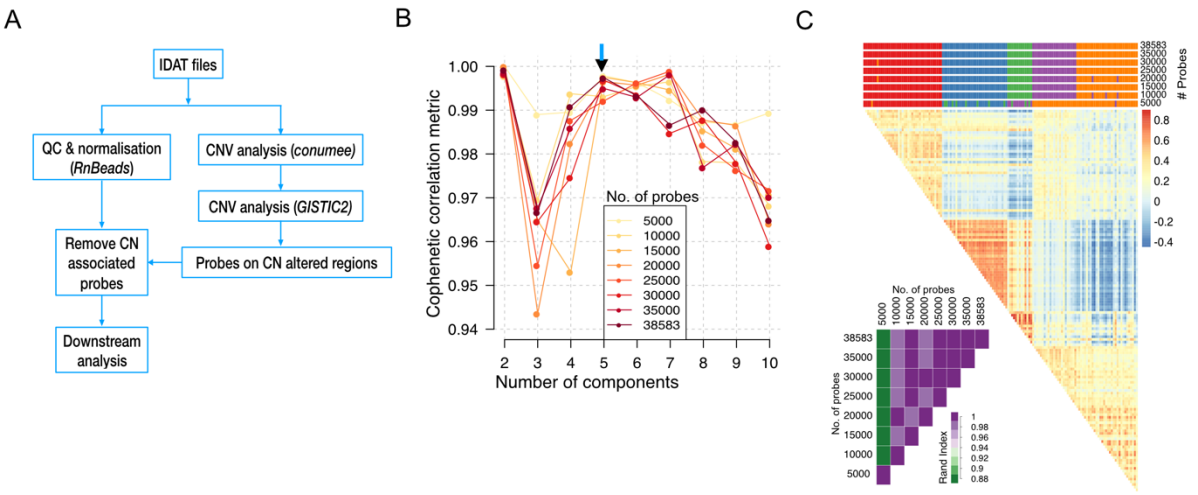

Figure 18. Identification of epigenetic clusters.

***A.** Pre-processing and QC steps involved in analyzing raw Illumina EPIC arrays. R packages used in particular stages are mentioned within parenthesis in italics. **B**. Cophenetic correlation (Y-axis) is measured for a range of values (2...10, X-axis). Optimum value is chosen when the correlation value reaches maximum followed by no-change or decrease in correlation metric. The same step was performed for varying probes as represented by the lines with a color gradient. Arrowhead represents the optimal value chosen for downstream analysis (N = 5). **C.** Cluster stability was measured by re-running clustering with n=5 for the different number of probes as indicated in the top annotation bar. Heatmap represents spearman-correlation coefficient between samples. The inner panel shows pairwise Rand-index values, which means similarity between two clustering results.*

Repeating clustering analysis by removing probes located on chromosomes X and Y showed no significant differences either **(Figure 19A)**. To further test the clustering's robustness, we generated EPIC arrays data for an independent series of 29 adult T-ALL samples (not included in the GRAALL 2003-2005 trial) and repeated the clustering (*N* = 143+29). Once again, our NMF results indicated five components **(Figure 19B)**. Further clustering resulted in the same order for discovery samples even in the presence of an independent validation cohort. Finally, we randomly sampled 80% of the initial cohort (*N* = 114) and repeated the clustering. Rand Index - a similarity score between two clustering results - was measured between original clusters and new clusters generated on subsamples. Repeating the entire process ten times, and results showed high similarity **(Figure 19C)** (Rand index 0.85 – 0.99), thereby confirming the obtained clusters' stability. Overall, the results suggest that the adult T-ALL consists of five distinct subtypes and the conserved methylomes.

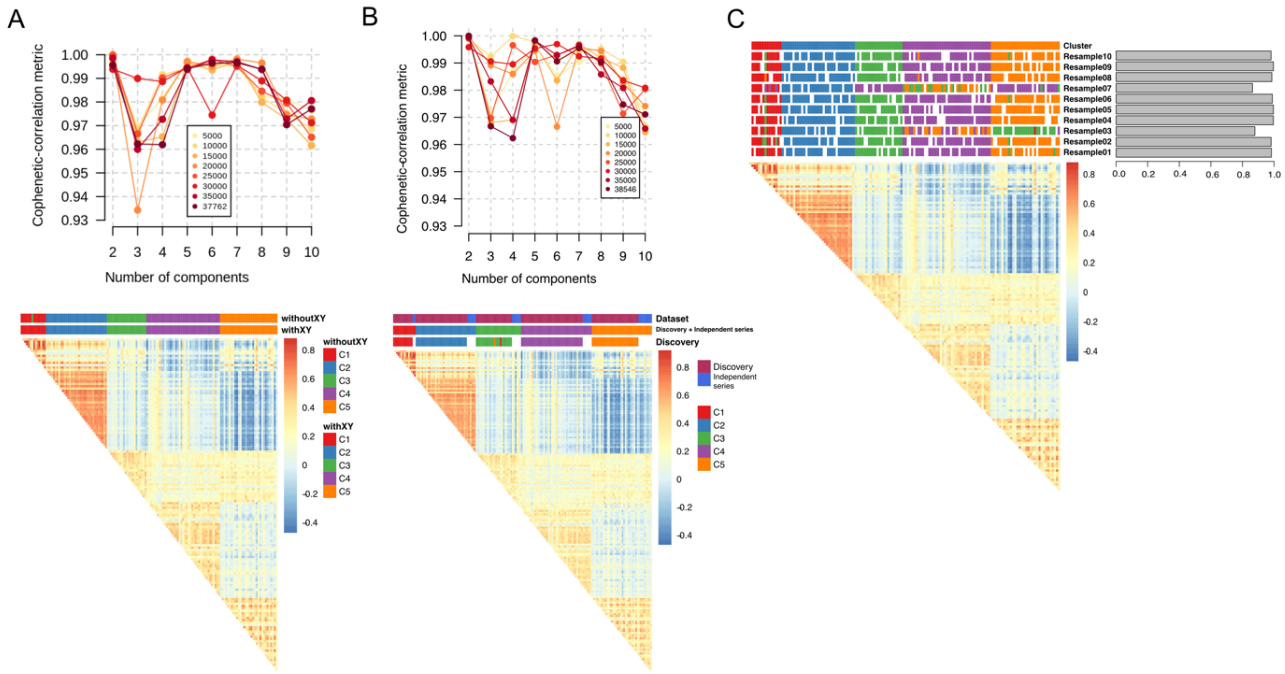

Figure 19. Assessing the stability of epigenetic clusters.

*Cophenetic metric and Spearman correlation heatmaps for clustering results in the presence or absence of XY probes **(A)** or when combined with an independent cohort of 29 samples **(B)**. **C**. Clustering results for the initial cohort (N = 143) and ten subsampled cohorts (N = 114). Right bar plot indicates a similarity score (Rand index) between the original cluster and subsampled clusters.*

Finally, we used the top 5% of most variable probes (*N* = 38583) for all our downstream analyses and defined the five distinct clusters, which all showed significant levels of DNAm **(Figure 20A-C)**. Based on the global DNAm levels, clusters are named - C_1, C_2, C_3, C_4, and C_5, with C_1 displaying the lowest level of DNAm and C_5 the highest. Compared to normal thymic cell types, C_1 and C_2 showed no significant overall differences, whereas C_3, C_4, and C_5 showed hypermethylation **(Figure 20C).**

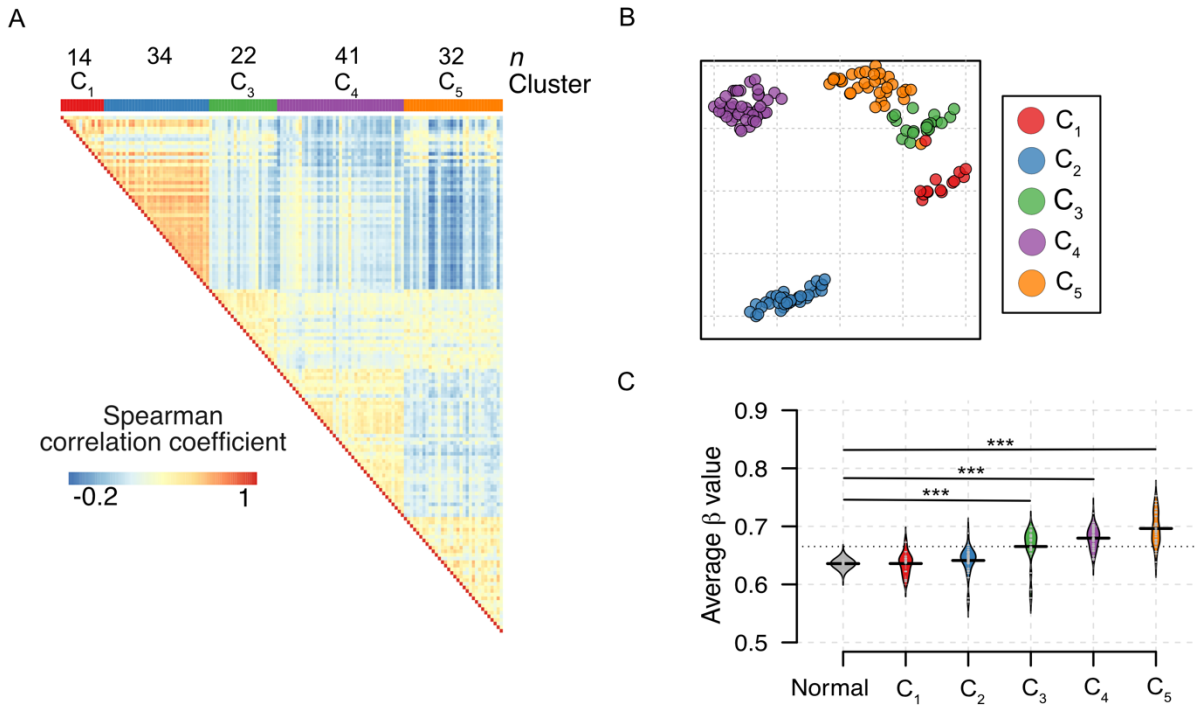

Figure 20. Epigenetic clusters in T-ALL.

A. *Heatmap of spearman correlation coefficient values between all leukemic samples. Top annotation bars indicate cluster sizes (n) and titles, respectively (top to bottom).* ***B.*** *Uniform Manifold Approximation and Projection (UMAP) plot of T-ALL samples color-coded according to the cluster.* ***C.*** *Violin plots depicting genome-wide DNAm values for every group. (*** $P < 0.001$ *two-tailed t-tests for mean differences).*

3.2.3 Characterization of the subtypes

The five identified clusters were significantly associated with signature maturation arrest stages and genetic drivers **(Figure 21A-D)**. Cluster C_1 (n = 14 samples, 9.8%) contained samples with ETP phenotype (7/11; $P < 0.01$) showing immature maturation arrest stages (9/12; $P < 0.01$). Moreover, C_1 samples contained lessened classical T-ALL drivers (*TLX, TAL1*, and *HOXA*). C_2 samples showed maturation arrest at αβ-lineage (28/31; $P < 0.01$) and significant over-representation of *TAL1* oncogene expression (16/33; $P < 0.01$). C_3 and C_4 contained samples with *TLX3* (14/22) and *TLX1* (25/38) overexpression, respectively. Alongside *TLX1* deregulation, a subset of C_4 samples also showed overexpression of *HOXA9* due to chromosomal translocation involving TCR-β locus (7/37; cis). C_3 and C_4 also differed in their maturation arrest stages, with C_3 having samples largely at TCR-γδ whereas, C_4 contained largely of αβ-lineage. In contrast, C_5 – a hypermethylated group – contained samples with *HOXA9* overexpression due to chromosomal translocation involving MLL/MLL10 gene (16/30;

trans). Moreover, similar to C_1, C_5 samples were of ETP phenotype with immature maturation arrest stage.

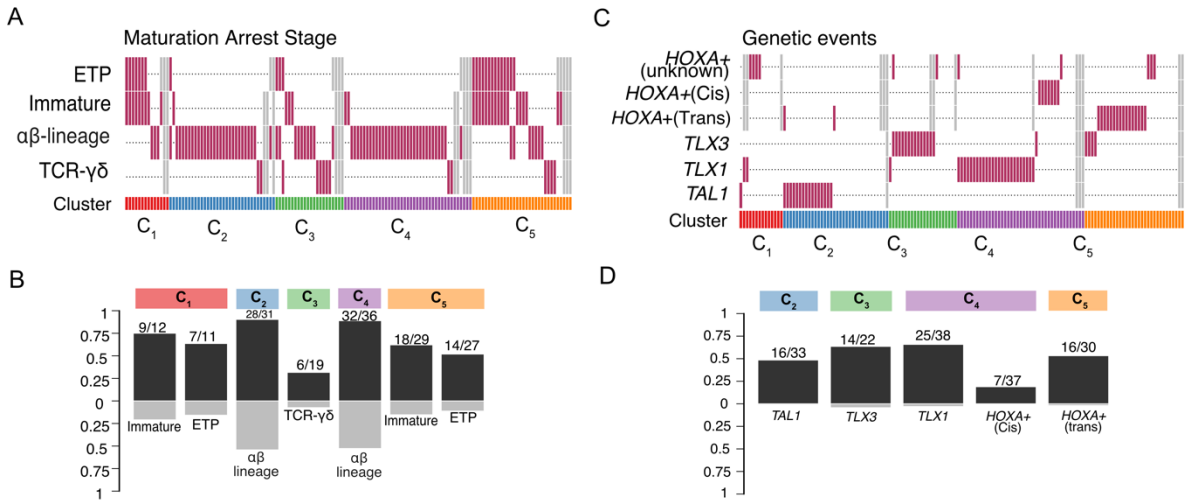

Figure 21. Characterization of epigenetic clusters in T-ALL.

***A**. Plot shows the association between maturation arrest stages (row) and T-ALL samples (columns). Each column represents a sample. The color-coded bottom annotation bar shows corresponding cluster assignments—T-cell maturation stages order rows (from the top most immature to the bottom-most mature). Gray bars indicate data not available. **B.** Bar plots display the significant association between clusters and maturation arrest stages (Fisher's exact test; $P < 0.01$). Bars are annotated with the ratio of samples belonging to a maturation arrest stage (as indicated by the bar title), and the total number of samples within the clusters. Bottom gray bars indicate the number of instances in the background with the highlighted maturation arrest stage. **C.** Heatmap shows the association between genetic events (row) and T-ALL samples (columns). **D.** Bar plots display a significant association between clusters and genetic events (Fisher's exact test; $P < 0.01$).*

In addition to differences in maturation arrest stages and overexpression of oncogenic transcription factors, clusters differed in their somatic mutational profiles. Mainly, genes involved in establishing DNAm – *DNMT3A* and *IDH2* - were significantly found in C_1 (*FDR* < 0.1; Fisher's exact test) **(Figure 22A)**. Moreover, all of the *IDH2* mutations co-occurred with *DNMT3A* making it a characteristic of C_1. Although *DNMT3A* contained a known R882 hotspot, unlike myeloid malignancies, they occurred at a lower frequency (~50% in AML vs. 30% in T-ALL) (Ley et al. 2010). *IDH2* variants, however, primarily occurred at known hotspot R140 **(Figure 22B)**. *PTEN* variants – mainly of the type loss of function INDELS – significantly occurred in C_2 whereas, *WT1* and *STAT5B* occurred in C_3, affecting the PI3K signaling pathway. Similarly, a significant fraction of C_4 samples were of *BCL11B* mutants. C_5 samples mainly contained mutations in PRC2 complex genes – *SUZ12* and *EZH2*.

In conclusion, our data-driven analysis of DNAm profile identifies five distinct groups of adult T-ALL characterized by the maturation arrest stages, overexpression of oncogenic transcription factors, and mutations in signaling pathways.

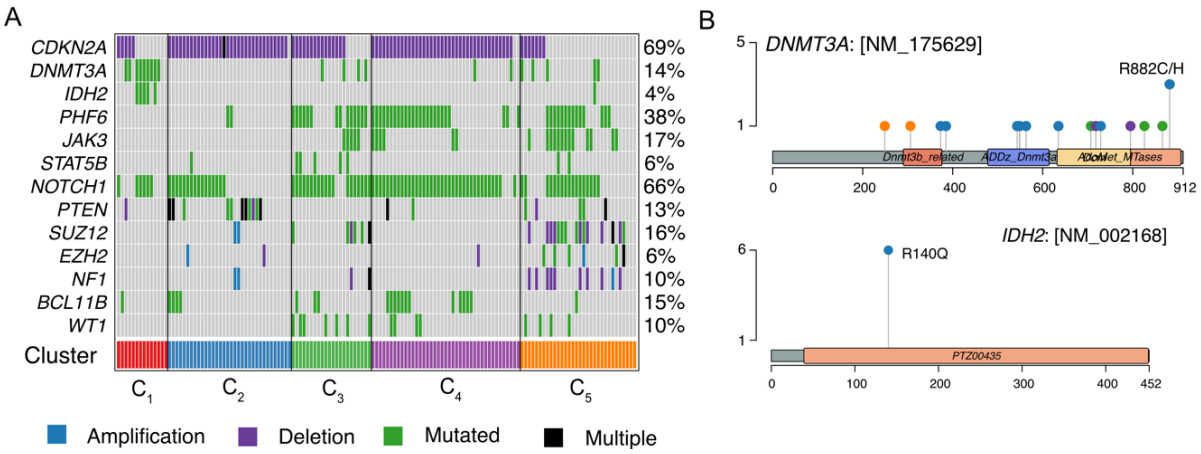

Figure 22. Somatic characteristics of epigenetic clusters

A. *Cluster-specific mutations were found across the cohort (Fisher's exact test; FDR < 0.05).* ***B.*** *Lollipop plot for DNMT3A (top) and IDH2 (bottom). Recurrent R882 and R140 mutational hotspots in DNMT3A IDH2 are named.*

3.2.4 Origins of DNMT3A mutations

In addition to methylation profiles, we carefully analyzed and characterized the distribution of somatic mutations. *IDH2* and *DNMT3A* were frequently altered and, similar to myeloid malignancies we observed co-occurrence of *IDH2* and *DNMT3A* mutations **(Figure 22A)** (Hou et al. 2012). Mutations in *DNMT3A* were largely of gain-of-function type with 30% of them being at R882 hotspot thereby affecting methyl-transferase domain whereas, mutations in *IDH2* occurred at known oncogenic hotspot R140 **(Figure 22B)**.

Furthermore, in contrast to myeloid leukemia, we observed that ~30% of the *DNMT3A* mutations showed high variant allelic frequency (VAF) possibly due to either bi-allelic loss or loss of heterozygosity (uni-parental disomy) **(Figure 23A)**. The same observation was also made in a public cohort of adult T-ALL samples derived from whole exome studies (Chen et al. 2018b). Copy-number analysis showed no deletions/amplifications in genomic loci containing *DNMT3A* (chromosome-2, p23.3), thereby suggesting possible homozygous mutations in T-ALL **(see Figure 17)**. It is possible that small focally deletions or amplifications of *DNMT3A* exist but are below the resolution of the EPIC arrays. High frequency of possible bi-allelic loss of *DNMT3A* events in T-ALL also suggests positive selection pressure for

DNMT3A mutants, thereby leading to rapid progression of the disease. This is further corroborated by survival analysis, whereby we observed *DNMT3A* (C_1) samples showed the worst survival among all the samples **(see Figure 30)**.

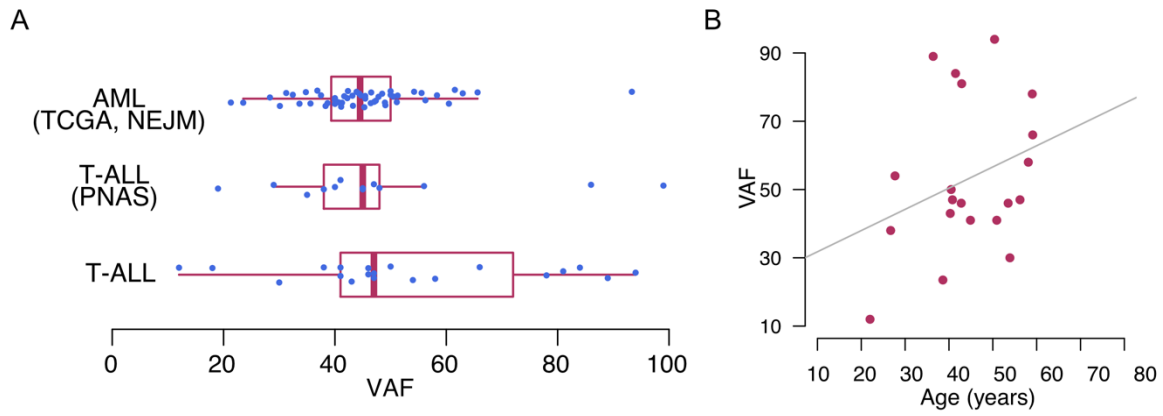

Figure 23. Clonal origins of DNMT3A mutation

A. *Variant Allele Frequency for DNMT3A variants in TCGA AML (Cancer Genome Atlas Research 2013), adult T-ALL from a published cohort (Chen et al. 2018b), and in present cohort.* ***B.*** *A small but non-significant correlation was observed between VAF and age (Pearson cor. 0.29; P-value > 0.05)*

In addition to VAF, we also observed that overall age of the patients belonging to C_1 was significantly higher compared to rest of the cohort (median age 45 years; $P < 0.001$; Wilcoxon rank-sum test). C_1 samples also showed a small but non-significant positive correlation between age and VAF **(Figure 23B)**. Given high age group, high VAF, and possible bi-allelic loss – one likely hypothesis could be that *DNMT3A* mutations have evolved from clonal hematopoiesis, where a dormant *DNMT3A* mutant clone acquired a second hit by means of bi-allelic loss thereby leading to a full-blown T-ALL. Although clonal hematopoiesis is primarily observed in myeloid leukemia, a recent case report has shown clonal hematopoiesis mediated T-cell lymphoma in a 45-year-old patient (Tiacci et al. 2018). These combined observations and evidence from literature urges the need for further in-depth analysis of origin of *DNMT3A* mutations in T-ALL.

3.2.5 DNA methylation changes in regulatory regions

To further understand the heterogeneity among the epigenetic clusters, we generated a genome-wide map of histone modifications associated with the active transcription, namely, H3K27ac and H3K4me3 (12 primary T-ALL; 7 to C_2, 2 to C_4, and 3 to C_5). PCA based on the

active promoter marks (H3K4me3) and enhancer marks (H3K27ac) similarly segregated the samples to that of DNAm. Samples showed clear separation by their cluster assignment, reflecting a conserved epigenome at multiple layers **(Figure 24A, B)**. Besides marking the regulatory regions for enhancers, H3K27ac peaks also harbor a subset called super-enhancers – known to be associated with the cellular identity and maintaining the oncogenic programs (Pott and Lieb 2015). Using H3K27ac, we created super and typical enhancer landscapes for 3 of the 5 clusters for which enough samples were available. The overall number of super-enhancers identified in T-ALL were significantly higher compared to the normal T-cells ($P < 0.001$; t-test) **(Figure 24C)**. SE-associated genes in T-ALL contained genes known to be involved in T-cell maturation and T-ALL leukemogenesis **(Figure 24D)**. In particular, driver genes associated with the clusters (such as *TAL1* in C_2, *TLX1* in C_4, and *HOXA* genes in C_5) had large blocks of super-enhancers encompassing the gene bodies and upstream regions **(Figure 24E)**.

Next, to correlate the changes in DNAm and histone marks, we compared each cluster with the normal thymic subpopulations to identify differentially methylated probes (DMP) (*FDR* < 0.05, |meth change| > 0.2). This analysis revealed varying degrees of DMPs with C_1 (34424 hyper/21478 hypo) and C_2 (48538 hyper/ 27757 hypo) displaying the approximately equal proportion of hyper- and hypo- methylated probes. In contrast C_3 (76873 hyper/ 22272 hypo), C_4 (97297 hyper/ 26728 hypo), and C_5 (101198 hyper/ 14503 hypo) showed large number of hypermethylated probes **(Figure 25A)**. Further genomic and CpG annotations of the DMPs showed hypermethylated DMPs strongly being enriched in promoters and CGIs across all the clusters **(Figure 25B)**.

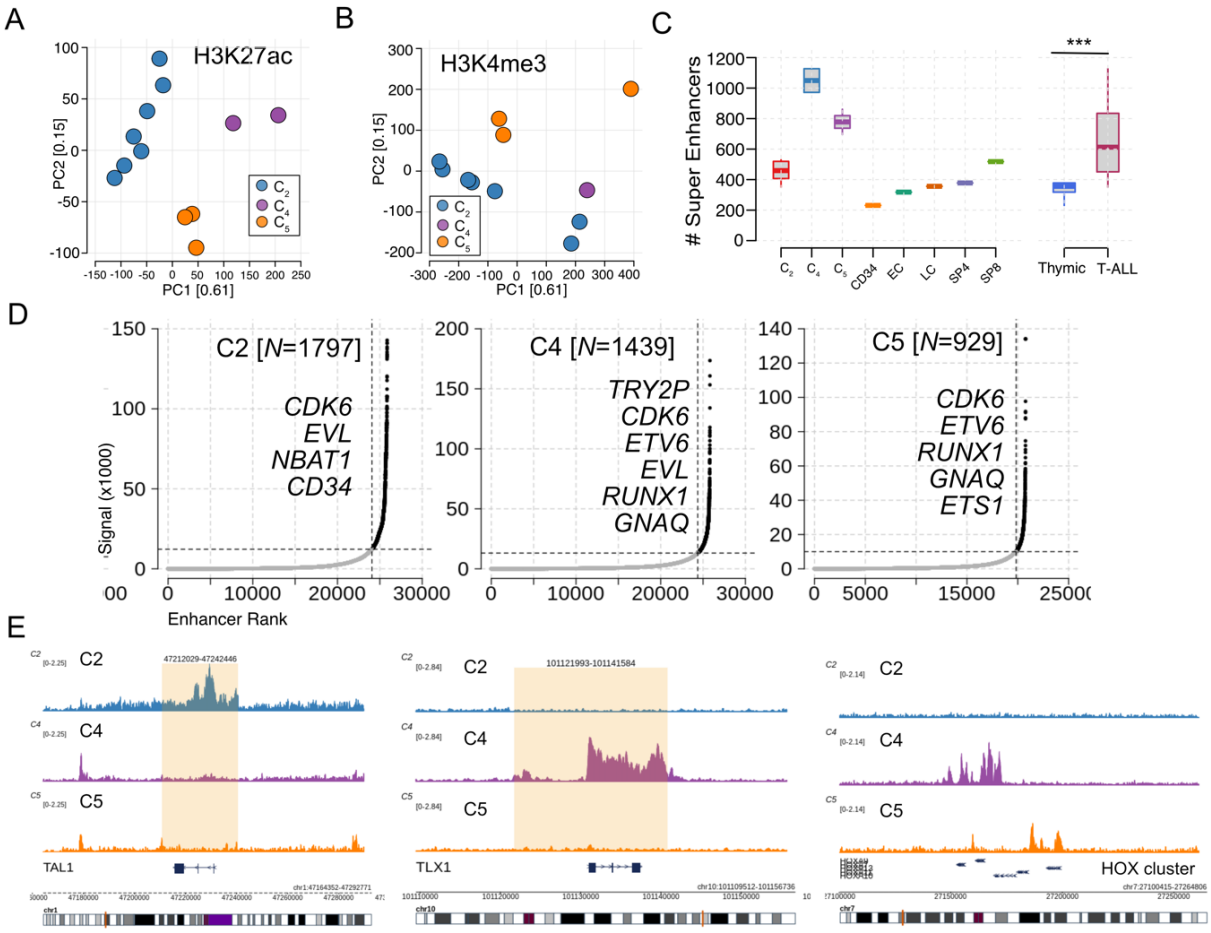

Figure 24. Enhancer landscape of epigenetic clusters.

Principal component analysis of primary T-ALL samples based on active H3K27ac peaks (peaks 2500 bp away from known TSS) ***(A)*** *and H3K4me3 peaks* ***(B)****. Samples are color-coded according to their corresponding cluster.* ***C.*** *Boxplot of the number of super-enhancers identified in each T-ALL sample, groups, and thymic cell types (X-axis) (*** *t-test; P < 0.01).* ***D.*** *Super Enhancers identified for each T-ALL cluster. Key SE associated genes are mentioned in italics. SEs are highlighted in black dots, whereas typical enhancers are in gray* ***E.*** *Examples of cluster-specific track plots display H3K2Ac signals for TAL1 (in C2), TLX1 (in C4), and HOXA (in C5) genes.*

In solid cancers and leukemias, DNAm changes in regulatory regions have been linked to preserving the cellular identity and malignant transformation. (Bell et al. 2016; Benetatos and Vartholomatos 2018). Therefore, we analyzed the dynamics of DNAm changes in regulatory elements of normal T-cell and T-ALL clusters. We used the LOLA to perform the enrichment analysis of DMPs among promoter-associated histone marks (H3K4me3), active gene expression (active, poised, and putative), SE and TE regions **(Figure 25C)** (Sheffield and Bock 2016). Results showed the yin-yang distribution of DMPs with hypermethylated DMPs primarily concentrated in TSS (+/- 1000 bps) and active promoters (H3K4me3 peaks) of genes involved in T-cell development. Hypermethylated DMPs also occurred among poised-

enhancers (H3K4me1+/H3K27ac-) of T-cells. However, hypomethylated DMPs significantly overlapped with enhancer regions derived from T-ALL clusters. Especially, SE and TE regions showed hypomethylation, suggesting an activated downstream expression of genes.

Since we observed the enrichment of hypomethylated DMPs among SE regions, we next performed the motif analysis within 100 bps surrounding the SE-associated DMPs. Results contained motifs characteristic of the corresponding oncogenic drivers **(Figure 25D)**. For example, C_1 had myeloid-like motifs such as CEBP, GATA1/2, PU1, and AP1. C_2 showed the motif enrichment of TAL1 and RUNX, GATA and MYB, which are known to form an autoregulatory loop (Sanda et al. 2012). C_5 being driven by *HOXA9* overexpression contained many HOXA family motifs in a cluster-specific manner.

Overall, our results show the hypomethylation of T-ALL-associated active enhancer regions and hypermethylation of promoters associated with normal T-cell development.

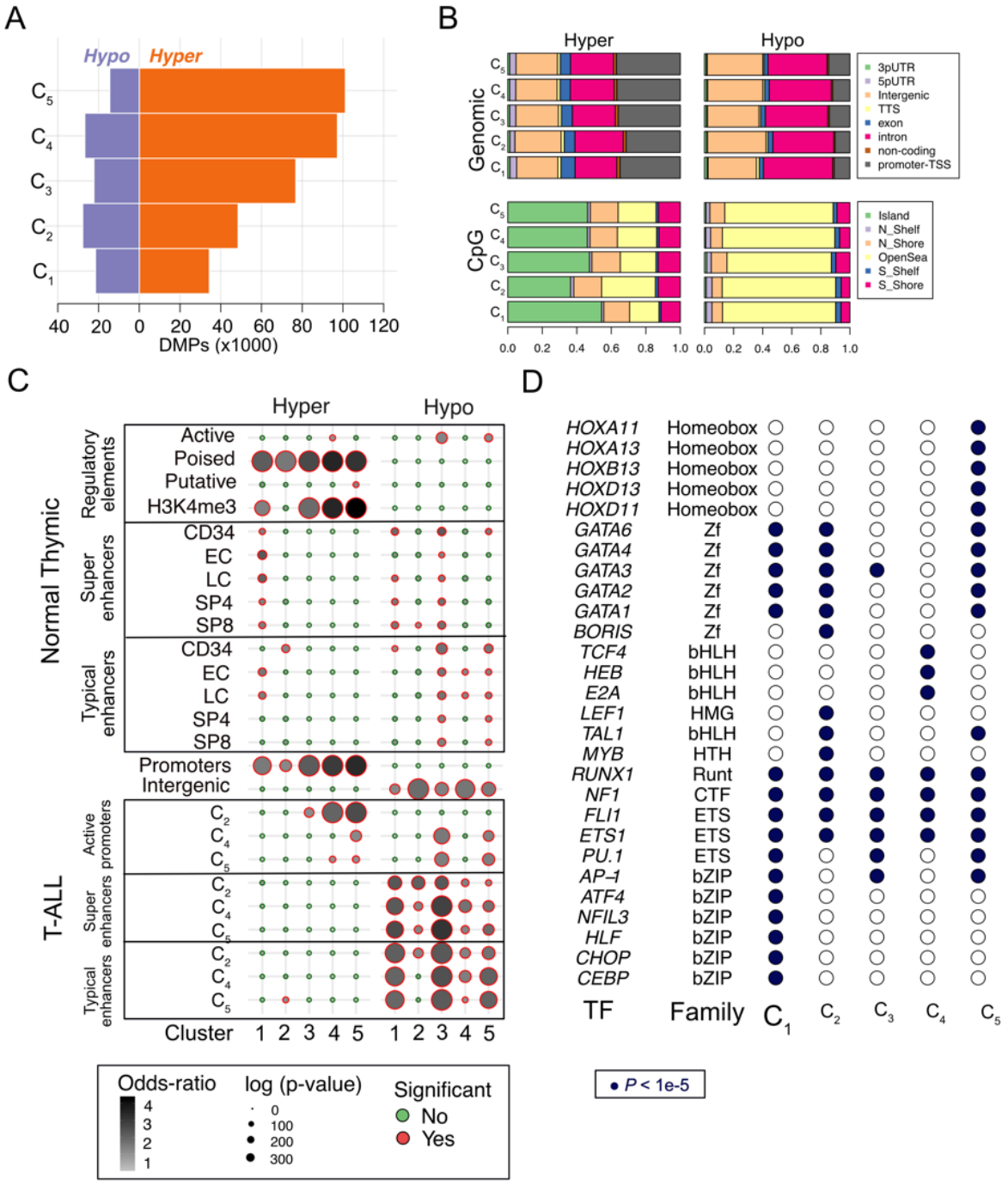

Figure 25. T-ALL enhancers are hypomethylated

A. *Numbers of hyper- and hypomethylated DMPs identified in each T-ALL cluster (C1- C5) compared to normal pooled thymic subpopulations.* ***B.*** *CpG and genomic annotations of hyper and hypo differentially methylated probes placed in each group compared to pooled normal thymic cells.* ***C****. Dot plot for the enrichment of hyper- and hypomethylated DMPs across all 5 clusters (X-axis) in various regulatory regions (Y-axis) as highlighted in the left side of the plot. Dots are color-coded for significance and size of the dots represents P values in log10 scale.* ***D.*** *Disease-associated TF motifs detected among hypomethylated DMPs (+/- 100bp) enriched within T-ALL associated super and typical enhancer regions.*

3.2.6 Integrative analysis of DNA methylation and gene expression

We used RNA-sequencing to produce expression profiles for a total of 48 samples (C_1=4, C_2=13, C_3=6, C_4=7, C_5=14), including four normal total thymus controls, to understand the impact of DNAm on gene expression. Although the samples seemed to cluster according to their methylation cluster in the PCA of the RNA-seq results, the distinction between the different leukemia groups appeared to be much less precise than their methylation signature. Surprisingly, whole thymus samples clustered with C_2 samples in the same way as DNAm findings did **(Figure 26A)**. Also, clusters had a strong expression of their candidate transcription factors **(Figure 26B)**. By comparing each group to total thymus samples, we performed differential gene expression (DGE) analysis. This study showed different degrees of gene expression changes, with C_1 having the lowest number of DE genes and C_5 having the highest number (DEGs). A large number of DE genes were up-regulated, indicating activated gene expression profiles **(Figure 26C)**. We conducted a correlation study between gene expression and the corresponding promoter DNAm levels of all protein-coding genes (*N* = 15,912 genes) across 44 T-ALL samples to determine the genome-wide effect of DNAm on gene expression. A small percentage of genes (*N* = 235) significantly correlated with promoter DNAm levels (FDR < 0.1) **(Figure 26D)**. Multiple studies have shown that DNA methylation levels at promoter regions do not always correlate with gene expression levels (Lister et al. 2009; Challen et al. 2011). However, except for the C_1 cluster, we found a consistent inverse relationship between DE genes and their promoter DNA methylation levels, with up-regulated genes having lower methylation levels than down-regulated genes **(Figure 26E)**.

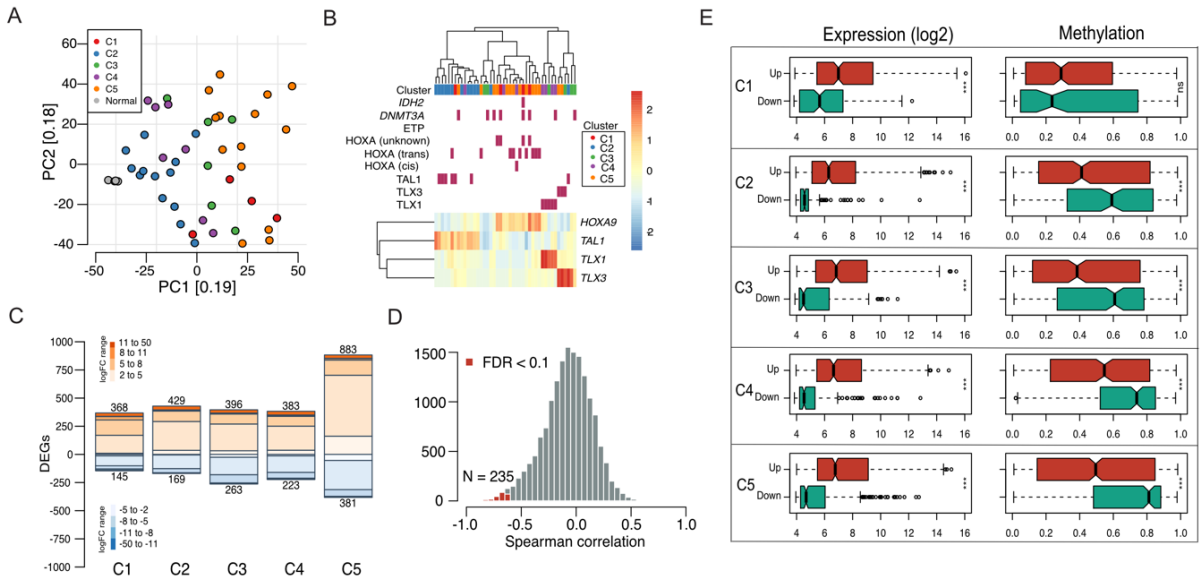

Figure 26. Integrated analysis of DNAm and gene expression.

A**. Principal component analysis of gene expression data for 48 samples (44T-ALL + 4 total thymus). Samples are color-coded for their T-ALL cluster. **B**. Heatmap of known T-ALL oncogenes. Top annotation bars depict validated genetic annotations. **C.** Barplot of the number of differentially expressed genes (DEGs) in each cluster compared to total thymus samples (FDR < 0.1). Up and down-regulated genes are color-coded for their fold change range. **D.** Histogram of Pearson correlation coefficient between promoter* DNAm and gene expressions of all protein-coding genes (N = 15,912 genes) across all tumor samples (N = 44). **E.** Distribution of gene expression (Up and Down-regulated genes from panel C) and their corresponding promoter* DNAm for every cluster. Notches indicate 95% CI around the median. (P < 0.001, t-test for differences in mean.*

**Promoter DNAm is estimated by averaging the beta values from probes within 1200 bp upstream and 800 bp downstream of known TSS.*

"As a result, we identified cluster-specific genes with a strong inverse association between DNAm and gene expression in a robust manner to better represent DNAm clusters. Within promoter regions, we used all of the DMPs (from **Figure 26A**) and looked at their corresponding gene expression **(Figure 27)**. With an inverse association between DNAm and gene expression, careful cataloging of these genes discovered some interesting cluster-specific genes involved in oncogenesis. In C_1 - among hypomethylated and overexpressed genes, we identified the myeloid factors, *AZU1, CSF3*, oncogenic genes *EGFL7* (Papaioannou et al. 2017), *FES* (Zhang et al. 2009), *SLC2A5* (Lai et al. 2020), and *S100A6* (He et al. 2017; Zheng et al. 2017; Chen et al. 2018a). Similarly, C_2 included *CD160* (Lesesve et al. 2015), *CD47* (Pai et al. 2019), *FOSL1* (Jiang et al. 2020), *MAPK8* (Chorzalska et al. 2018; Pyo et al. 2018; Lehmann et al. 2019; Newman et al. 2019) and *MYEOV* (de Almeida et al. 2006; Moreaux et

al. 2010) in C_3, and *CAPG, RGS17* (Bodle et al. 2018; Li and Luo 2018), *FAM83A* (Yu et al. 2020), LGALS(Shih et al. 2019), *NCR1* (Cheminant et al. 2019) and *EMP1* (Aries et al. 2014) in C_5. Contrarily, inactivation of tumor suppressor genes (TSG) by DNA hypermethylation has been suggested as a hallmark of cancers. Therefore, we observed several TSGs that are hypermethylated with decreased expression; C_5 contained the largest number of DEGs with well-known inactivated TSGs such as: *ALS2CL* (Lee et al. 2010), *AMPH* (Yang et al. 2019), *CMTM8* (Zhang et al. 2016), *DEPDC7* (Liao et al. 2017), *HOOK1* (Sun et al. 2017), *MITF* (Vivas-Garcia et al. 2020), *MPPED2* (Gu et al. 2019), *PCDH9* (Lv et al. 2017), *RARRES1* (Roy et al. 2017), *RASEF* (Maat et al. 2008), *RNF180* (Deng et al. 2016), *S100A16* (Zhang et al. 2019) and *SLFN5* (Wan et al. 2019)."

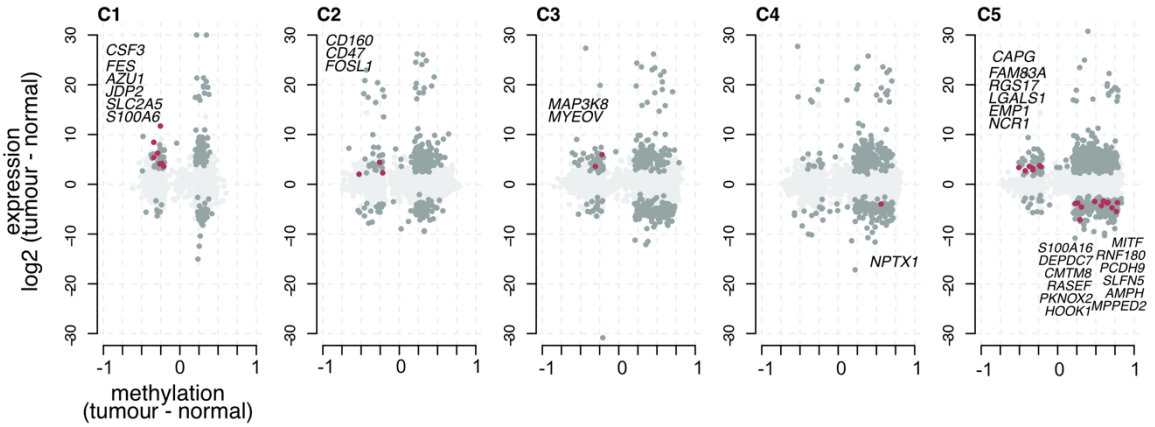

Figure 27: Identification of cluster-specific dysregulated genes.

Scatter plot of gene expression and DNAm differences in DMPs between tumor and normal samples for all 5 clusters (C1-C5). Dark gray color indicates genes significantly differentially expressed (FDR < 0.1; from Figure 16C). Selected genes known to be associated with the T-ALL pathogenesis are highlighted in red and annotated.

In conclusion, our combined analysis of DNAm and gene expression datasets reveal the role of transcriptional and epigenetic modifications in a coordinated way, leading to the identification of several clusters of specific oncogenes and TSGs.

3.2.7 Maturation arrest stages of T-ALL subtypes

Since our WGBS based analysis of intrathymic cell types was able to identify developmental associated genomic regions and reconstruct the thymopoiesis trajectory **(see Figure 11D)**, we asked if the same results can be used to predict the maturation arrest stages of T-ALL clusters.

By utilizing the tDMRs identified from our WGBS analysis, we performed the PCA, which separated the thymic cell types on a two-dimensional plane with cell types organized by their maturation stages **(Figure 28A)**. Phylogenetic trees from the same data reconstructed the thymopoiesis ontogeny **(Figure 28B)**. Next, using this phylogenetic tree as a reference, we projected our entire cohort of T-ALLs which showed the potential maturation arrest stages of T-ALLs **(Figure 28C)**. The combined tree still maintained the T-cell ontogeny even in the presence of T-ALLs suggesting the robustness of tDMRs in preserving the cellular identity **(Figure 28D)**. Moreover, T-ALL samples were hierarchically ordered with the ETP ALLs (C_1 and C_5) occurring during the earlier stages, followed by TLX deregulated clusters (C_3 and C_4) and finally *TAL1* deregulations (C_1). These results are also summarized by the phylogenetic tree constructed from T-ALL groups' aggregated methylation levels **(Figure 28D inset plot)**. Importantly, all 5 clusters were placed midst of ISP and DP TCR stages of T-cell development, indicating that maturation arrest occurs earlier during T-cell differentiation.

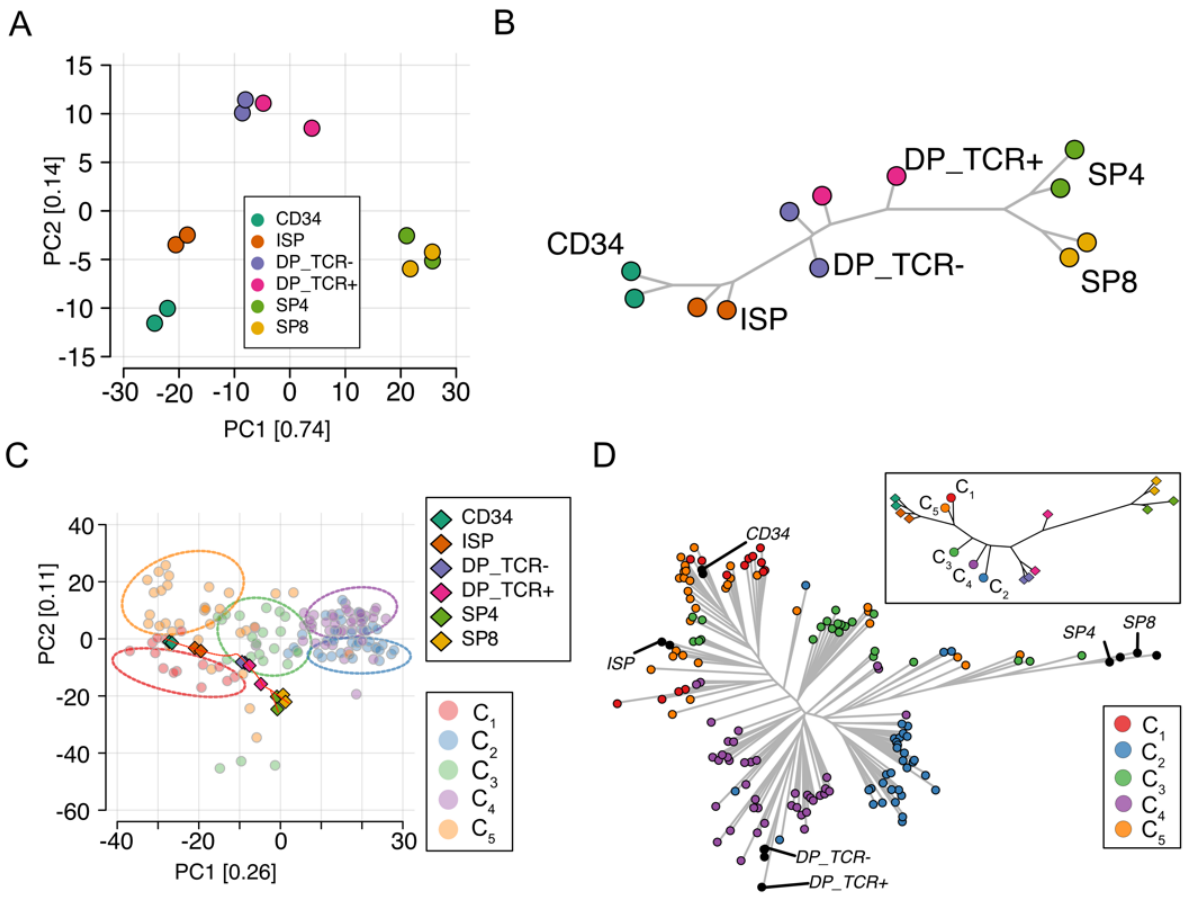

Figure 28. DNA methylation predicts T-cell differentiation and maturation arrest stages of T-ALL clusters.

***A.** Principal Component Analysis (PCA) shows separation and ordering of T-cell subtypes according to different maturation stages (clockwise from bottom left to bottom right). **B.** T-cell developmental phylogenetic tree inferred from tDMPs shows the placement of T-cell subtypes. **C.** PCA of normal T-cell and T-ALL methylomes using tDMPs. Normal T-cell subpopulations (n = 12) are depicted in diamond shapes, and T-ALLs (n = 143) are circles - color-coded according to the cluster (C1 – C5). The known T-cell developmental trajectory starting from CD34 is shown as a red curve overlaid on top of normal T-cell population. **D.** Phylogenetic tree of the entire cohort (n = 143 T-ALLs; n = 12 normal T-cells) constructed using tDMPs shows ordering of T-ALL samples (color-coded according to their corresponding clusters) along T-cell developmental pathway (left to right). Normal T-cells are in thick black circles labeled for cell-types. Inlet panel shows a simplified phylogenetic tree constructed with average DNAm levels of epigenetic clusters along with normal T-cell types (in diamond shapes) shows the order of maturation arrest stages of T-ALL subtypes (in circles) along the T-cell developmental pathway (left to right).*

Of note, we find that global DNAm levels of these clusters do not correlate with the developmental arrest stages, implying a non-linear association of DNAm levels and maturation arrest stages **(Figure 29A)**. Besides, as expected, oncogenic TFs accumulated at precise locations in a hierarchical order **(Figure 29B)**.

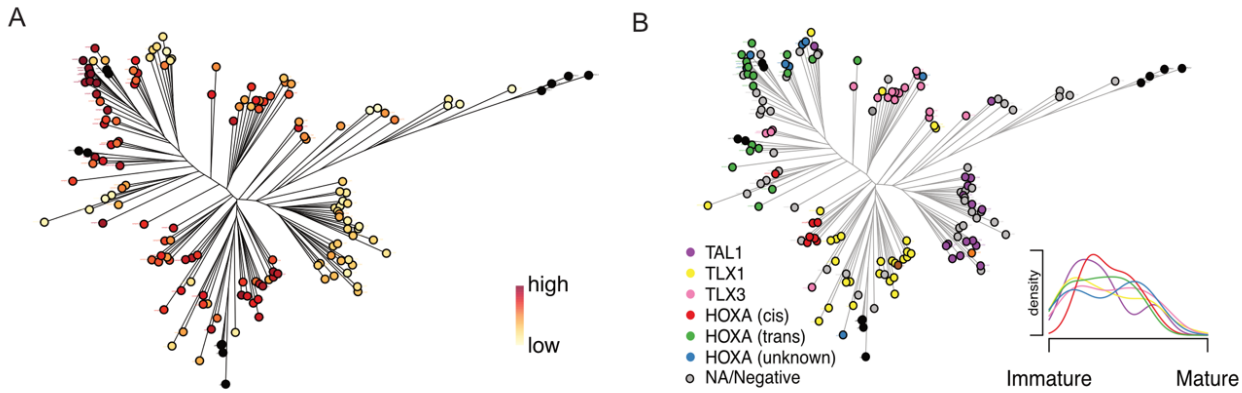

Figure 29. Associating between DNAm levels, TF overexpression with T-ALL ontogeny.

*Phylogenetic tree with samples color-coded for global DNA methylation levels **(A)** and TF deregulation **(B)**. Black dots represent normal thymic cell types. The inset plot shows the density of mutated samples along the trajectory.*

Overall, the results suggest the potential role of DNAm in predicting the normal T-cell developmental trajectory and the order of maturation arrest stages for the five epigenetic T-ALL subgroups.

3.2.8 Epigenetic clusters are associated with the clinical outcome

CIMP-negative T-ALL patients have been shown to be significantly associated with higher cumulative incidence of relapse as compared to CIMP-positive patients suggesting a prognostic relevance of DNAm profiles in T-ALL (Borssen et al. 2016). This was also observed in adult T-ALL (Touzart et al. 2020). Thus, we tested the prognostic relevance of the five DNAm clusters (C_1-C_5) which revealed association with diverse levels of overall survival (OS) and event-free survival (EFS) (P = 0.08 and P = 0.045, respectively) **(Figure 30A, B).** As reported previously, TLX clusters C_3 and C_4 displayed favorable outcomes, whereas hypomethylated C_1 showed unfavorable results (Ferrando et al. 2004).

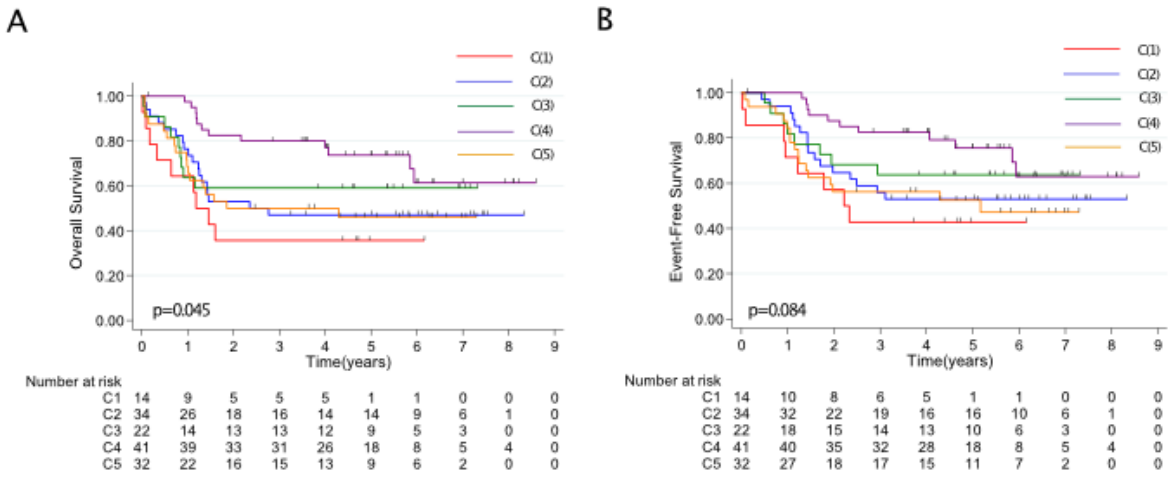

Figure 30. Overall Survival (OS) and Event Free Survival (EFS) of methylation clusters.

A and B. *OS (A) and EFS (B) for five epigenetic groups (C1-C5). P-values are estimated from the log-rank test. Bottom tables show risk table.*

"Based on genome-wide DNAm levels, we classified five clusters into three groups with significantly distinct DNAm levels, $C_{(1+2)}$, $C_{(3+4)}$, and $C_{(5)}$, which displayed hypomethylation, intermediate-hypermethylation, and high-hypermethylation respectively **(Figure 31A, B)**. Patients in the hypomethylated $C_{(1+2)}$ subgroup demonstrated shorter OS (5-year OS probability; 50% [95% CI = 35% to 63%] vs 71.5% [95% CI = 58% to 81%]; P = 0.031) and EFS (5-year EFS probability; 44% [95% CI = 30% to 57%] vs 69% [95% CI = 55% to 79%]; *P* = 0.015) as compared to the intermediate-hypermethylated $C_{(3+4)}$ subgroup. Interestingly, the high-hypermethylated $C_{(5)}$ subgroup displayed distinct clinical outcome and had significantly poorer survival probability as compared to the $C_{(3+4)}$ subgroup (5-year OS probability; 53% [95% CI = 34% to 68%] *vs* 72% [95% CI = 58% to 81%]; *P* = 0.037) and EFS (5-year EFS probability; 46% [95% CI = 28% to 62%] vs 69% [95% CI = 55% to 79%]; P = 0.045). C_5 and $C_{(1+2)}$ patients with ETP-ALL showed similar shorter OS and EFS **(Figure 31C)**."

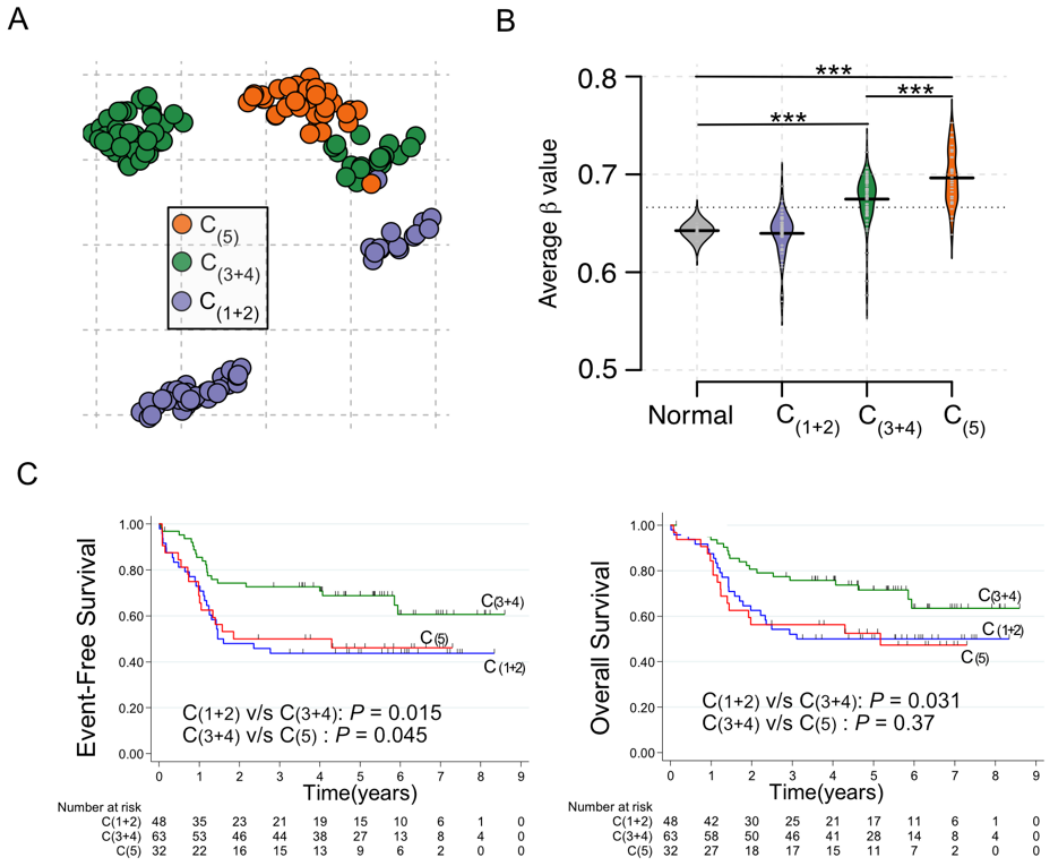

Figure 31. Prognostic impact of DNA methylation.

A. *UMAP of clusters classified into hypo (C_{1+2}), intermediate (C_{3+4}) and hypermethylated (C_5) subgroups.* ***B.*** *Average genome-wide DNA methylation of C_{1+2}, C_{3+4}, and C_5 subgroups (*** $P < 0.001$ two-tailed t-tests for differences in mean)* ***C.*** *Overall Survival (OS) for patients classified into hypomethylated (C_{1+2}), intermediate hypermethylated (C_{3+4}), and hypermethylated (C_5) subgroups. The risk table indicates the number of individuals at risk for a given time point. P-values are derived from log-rank test.* ***D.*** *Event Free Survival (EFS) for patients classified into C_{1+2}, C_{3+4}, and C_5 subgroups.*

"Importantly, it does not seem to be linked to the same clinical parameters in adverse prognosis as $C_{(5)}$ and $C_{(1+2)}$. C5 patients have shown a poor early response with significantly slower D8 prednisone and D15, and higher MRDs (Minimal Residual Disease; 28.1% versus 54.2%, Bone marrow response, 21.9% versus 66.7% and negative MRD induction, and 38.1% versus 78.3%), respectively, unlike $C_{(1+2)}$ patients, which shows a low level of early response with marked slowness in both D8 prednisone and D15 **(Table 4)**."

	$C_{(1+2)}$	$C_{(3+4)}$	$C_{(5)}$	*P*-value†
	N= 48	**N=63**	**N=32**	
Clinical Subsets Analyzed				
The median age in years (range)	26.2 (16.3-59.1)	30.5 (16.4-57.2)	31.4 (18.8-59)	0,3
age >45 years- no./total no. (%)	6/48 (13%)	6/63 (10%)	5/32 (15%)	0,67
Sex ratio, Male/Female- no.	38/10	46/17	23/9	0,7
WBC (G/L), median (range)	69.9 (0.9-604.4)	37 (4.1-645)	31.2 (2.2-241.6)	0,3
CNS involvement- no./total no. (%)	7/48 (15%)	8/63 (13%)	5/32 (16%)	0,9
Early Response				
Prednisone response - no./total no. (%)	26/48 (54%)	44/63 (70%)	9/32 (28%)	**0,001**
Bone marrow response - no./total no. (%)	30/45 (67%)	41/63 (65%)	7/32 (22%)	**<0.001**
MRD (TP1) <10-4 - no./total no. (%)	18/23 (78%)	33/39 (85%)	8/21 (38%)	**0,001**
Complete remission - no./total no. (%)	44/48 (92%)	61/63 (99%)	29/32 (94%)	0,4

WBC (G/L), white blood cells; **CNS**, central nervous system; **MRD** (TP1), post-induction minimal residual disease; †Fisher's exact test Mann-Whitney tests were used where appropriate.

Table 4. Clinical characteristics of the three prognostic subgroups^.

Multivariate survival analysis performed by Guillaume P. Andrieu

In general, our data found a subset of T-ALL with a low primary therapeutic response, which paves the way for unique epigenetic therapeutic plans.

3.2.9 Machine learning models predict risk associated T-ALL subgroups

Considering the stable nature of DNAm and relative ease in measuring, many studies involving solid tumors and leukemias have utilized DNAm to predict disease subtypes and prognostic groups (Figueroa et al. 2010b; Capper et al. 2018). Accordingly, we thought methylation cluster prediction could be an interesting diagnostic and prognostic biomarker in clinical

^ Table copied from the soon-to-be published joint manuscript: Epigenetic blueprint identifies poor outcome and hypomethylating agent-responsive T-ALL subgroup

practice. We first classified our entire cohort into training (60%; *N* = 86) and test (40%; *N* = 57) datasets. Next, we generated Random Forest (RF) models using the training dataset to predict epigenetic and prognostic clusters **(Figure 32A)**. These models were also cross-validated for maximum accuracy using 10-fold cross-validation. Using recursive feature selection, we then derived a minimal set of probes that could assign samples in the test dataset to either of the five epigenetic or prognostic clusters with high accuracy. The first model, which consists of 79 CpG probes (common across 450K and 850K arrays), correctly assigned test samples to their five methylation clusters with an accuracy of 96.29% (52 of 54 were correctly classified) whereas, the second model classified test samples as $C_{(1+2)}$, $C_{(3+4)}$, and $C_{(5)}$ groups with an accuracy of 98.21% using a minimal set of 59 probes **(Figure 32B, C)**. Besides, we built a final model to predict favorable $C_{(3+4)}$ or unfavorable $C_{(1+2+5)}$ risk groups. A predictor of only seven methylation probes predicted the risk group with an accuracy of 91.2% on the test cohort **(Figure 32D)**. We believe that these models could be of clinical relevance in the better assignment of risk groups.

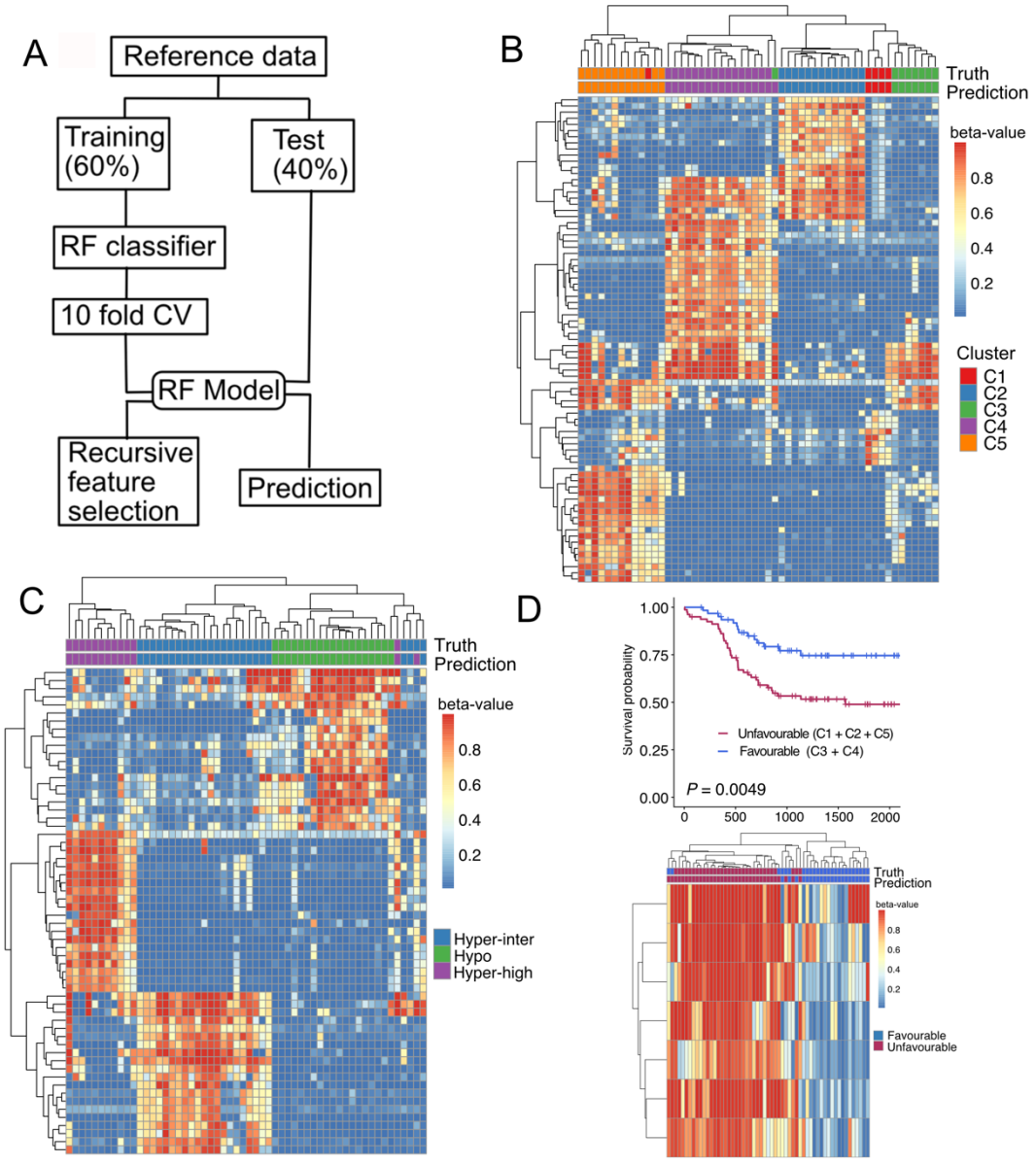

Figure 32. Random Forest (RF) models predict epigenetic clusters and survival.

A. *Overview of steps involved in RF model and feature selection* ***B.*** *Heatmap of most informative probes depicting predicted clusters (C1, C2, C3, C4, and C5) on the test dataset.* ***C.*** *Heatmap of most informative probes displaying the predicted methylation groups (hypomethylated, hyper-intermediate, and hyper-high) on the test dataset.* ***D.*** *OS of favorable and unfavorable subgroups (top panel). Heatmap of 7 CpG signature for predicting prognostic groups, unfavorable (C1, C2, C5) and favorable (C3 and C4) (bottom panel).*

To further validate the RF models' accuracy in predicting the five epigenetic T-ALL subtypes in an independent cohort, we generated EPIC data for a series of samples that were not included in the GRAALL 2003-2005 trial (*N* = 29). This model was able to classify the samples in the validation cohort into 5 clusters which were well correlated with their genetic drivers **(Figure**

33). However, we believe this type of analysis would greatly benefit by training on large cohorts as the model tends to perform better.

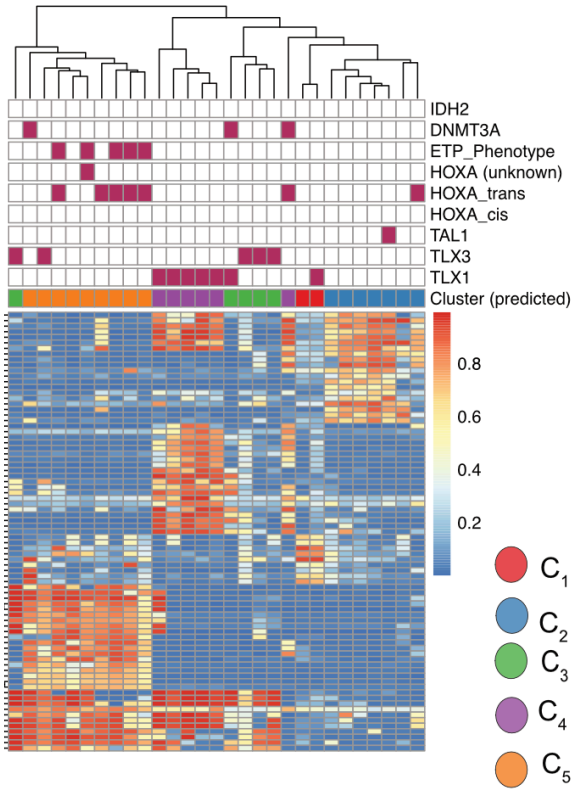

Figure 33. Validation of random forest models in an independent cohort

Heatmap of most informative probes selected by recursive feature selection (N = 79 probes) depicting predicted clusters on the validation dataset (N=29 samples). Top annotation bars indicate key oncogenic events associated with each sample.

4 DISCUSSION

4.1.1 DNA methylation dynamics of thymopoiesis

Using high coverage WGBS of sorted cells isolated from multiple thymi at distinct stages of T-cell differentiation, we describe the high-resolution mapping of DNAm dynamics during human thymopoiesis. Although there have been attempts to achieve the same, they have lacked coverage and are limited to arrays-based assays (Rodriguez et al. 2015). A recent study by BLUEPRINT consortia has performed in-depth multi-omics of epigenetic modifications; however, there still existed an opportunity to address the specific queries (Cieslak et al. 2020).

Our comprehensive analysis reveals a sequential loss of DNAm during thymopoiesis. The loss of DNAm is most prominent post T-cell commitment and is reflected in the number of identified DMRs. Similar results are also shown in B-cell development and some of the myeloid cell maturation (Farlik et al. 2016; Oakes et al. 2016). The systematic loss of methylation observed in multiple branches of hematopoiesis suggests the critical requirement of demethylation in lineage commitment regardless of the terminal cell fate. Besides, methylation losses are prominent among binding sites for master T-cell transcription factors such as *NOTCH1, MYB*, and *RBPJ*. Several studies over the decades have made it clear that the critical mode of action of DNAm involves regulating the binding sites for TFs, and the same seems to be true for thymopoiesis.

Next, our comparative analysis of CD4+ and CD8+ fate decisions across BLUEPRINT regulatory regions shows a significant overlap in the loci that are differentially methylated from the progenitor double-positive cells. Classically, two possible models have been proposed for CD4 or CD8 commitment; an *instructive model* involving MHC class-1 or MHC class-2 ligands whose interactions with CD8 or CD4 decide the final fate; and the *stochastic model* involving a random selection and commitment towards CD4 or CD8 (Kappes et al. 2005). Regardless of the mechanism, DNAm changes required for functional specificity highlight the underlying genomic characteristics. For example, our results show that *RUNX1* binding regions are significantly hypomethylated among the specific areas for CD8+ cell commitment. This observation is corroborated by the study wherein *ZBTB7B* causes differentiation towards CD4 by antagonizing RUNX (Wildt et al. 2007).

By performing a careful de-novo DMR analysis, we reveal a stage-specific genomic region undergoing differential methylation (tDMRs). Significantly, tDMRs show uni-directional and irreversible loss of methylation at each step of differentiation. Similar observations have also been made in the array-based analysis and seem to be thymopoiesis' epigenetic nature (Rodriguez et al. 2015). Interestingly, stage-specific loss of methylation shows hypermethylation in predecessors, whereas the same remains hypomethylated in successors. The observation suggests the progressive epigenetic silencing of pathways no longer needed in the newly differentiated cell type.

Moreover, tDMRs are highly stable and can be used to construct the lineage of thymopoiesis. The functional properties and stability of DMRs have also been shown in hematopoiesis to predict the cell types and lineage tracing (Farlik et al. 2016). Similarly, our combined analysis of hematopoiesis and thymopoiesis in an independent dataset further validates the tDMRs and now serves as an *epigenetic atlas* for T-cell development.

4.1.2 DNA methylation identifies high risk associated T-ALL subgroups

“Epigenetic studies in T-ALL have historically been based on CpG-rich regions (Milani et al. 2010; Borssen et al. 2013; Nordlund et al. 2013; Borssen et al. 2016). The recent extensive study of pediatric T-ALL DNA methylomes showed that CpG islands are hypermethylated with the PRC2 target genes and already present in preleukemic thymocytes (Roels et al. 2020b). Our extensive analysis of 143 T-ALL samples using EPIC arrays has established five distinct T-ALL subgroups with distinct DNAm levels. Four of the five clusters were associated with the known oncogenic TFs whereas, a novel hypomethylated subgroup associating with the poor clinical outcome was identified (Ferrando et al. 2002). These findings show the role of oncogenic developmental events that redefine the underlying methylome in leukemogenesis. The previously described gene expression-based analysis is complemented with features of these DNAm clusters. Extensive overlaps in transcriptional signatures between *TLX1* and *TLX3* T-ALL, for example, have been shown by expression-based studies. However, these two subgroups are robustly distinguished from their DNAm patterns and their phases of methylation-based arrest, indicating a deeper variation in the deregulated pathways. Based on the mode of overexpression, *HOXA9* deregulated samples showed significant differences in DNAm. Those with *HOXA9* deregulation in cis (under the influence of the TCR enhancer) clustered with *TLX1* deregulated samples (C_4) displaying early cortical maturation arrest

stages, whereas those with *HOXA9* deregulation in trans (under the influence of SET-NUP214, MLLT10, or MLL fusions) formed a distinct hypermethylation cluster (C_5), associated with early cortical maturation arrest stages (Bond et al. 2016) . Moreover, PRC2 mutations are frequent in the C_5 subgroup. Immature/ETP ALLs were divided into two groups by DNAm: C_1 hypomethylated *DNMT3A/IDH2*-rich ALLs and C_5 hypermethylated trans-HOXA/PRC2 ALLs, paving the way for alternative therapy. *TAL1* deregulated samples clustered together, indicating a disruption of common downstream pathways, regardless of the mechanism of deregulation (SIL-TAL1, or upstream neo-enhancer (Mansour et al. 2014). *TAL1* is one of the most frequently deregulated driver oncogenes in T-ALL (approximately 30% of cases), either through deletions (SIL-TAL1; 10-20%), oncogenic neoenhancer (20%) or rare V(D)J-mediated translocations [t(1;14); 1-2%) (Bernard et al. 1990). However, such lesions are absent in about 40% of *TAL1*+ cases, indicating that *TAL1* is overexpressed for unknown reasons. Only 16/34 samples in cluster C_2 had either the SIL-TAL1 deletion or the oncogenic neoenhancer, implying that the rest of the samples show TAL1 overexpression due to unknown mechanisms."

"Cluster C_1, enriched for co-occurring *DNMT3A/IDH2* mutants, has no previously identified TF overexpression. However, all *DNMT3A/IDH2* mutated cases had *NOTCH1* mutations, suggesting the synergistic role of the deregulated epigenome and *NOTCH1* signaling in oncogenesis (Grossmann et al. 2013; Kramer et al. 2017). Single mutations in epigenetic factors (*DNMT3A, IDH1, TET2/3*) involved in DNAm, however, were not associated with a specific methylome, raising the question of the role played by such genomic alterations in leukemogenesis. Cluster-specific methylome analysis revealed systematic DMP distribution, with significant hypermethylation of T-cell developmental associated active promoters and poised enhancers and significant hypomethylation of T-ALL-related enhancers. Our findings corroborate previous results of enhancer hypomethylation in solid tumors and leukemia (Bell et al. 2016; Benetatos and Vartholomatos 2018)."

"Our integrated gene expression and DNAm analysis, similar to cluster-specific somatic and epigenetic alterations, highlights the combined effects of DNAm and gene expression in T-ALL. *SLC2A5*, for example, is hypomethylated and overexpressed in C_1, is also overexpressed in a subset of AML and childhood Philadelphia chromosome+ ALL, and linked to a poor prognosis. The encoded protein is a fructose transporter that increases fructose use by leukemic cells and is responsible for fructose uptake by the small intestine (Mansour et al.

2018; Lai et al. 2020). JDP2, a C_1-specific bZIP transcription factor, has been identified as a novel oncogene in ETP T-ALL and has been linked to a poor prognosis. Another gene, *EMP1*, is linked to poor pediatric ALL and confers prednisolone resistance (Aries et al. 2014). It is hypomethylated/overexpressed in C_5, and our multivariate analysis also discovered a link between C_5 and an inadequate early prednisone response."

"The recapitulation of the maturation arrest stages is another important finding of this study. Lymphopoiesis is a complicated process involving the stage-specific expression of several transcription factors and cell surface markers (Koch and Radtke 2011). Our DNAm-based phylogenetic trees accurately captured this ontogeny, demonstrating the importance of DNAm in thymopoiesis regulation. These findings support previous reports in mouse hematopoiesis, which show that DNAm dynamics choreograph myelopoiesis and lymphopoiesis (Ji et al. 2010). Similar reports are also made in human lymphopoiesis, where DNAm dynamics during B cell maturation show a steady progression of changes (Oakes et al. 2016). By combining T-ALL samples, these phylogenetic trees accurately captured the known clinically relevant subgroups, indicating developmental arrest at different stages of thymic maturation. An independent study using open chromatin signals revealed similar results, highlighting the conserved epigenetic marks at multiple layers (Erarslan-Uysal et al. 2020). Overall, our findings show that C_5 and C_1 are characterized by trans overexpression of HOXA and alterations in myeloid-like genes, respectively, occur early, followed by C_3 (enriched in *TLX3* cases) and C_4 (increased in *TLX1* and cis HOXA deregulated instances). Finally, C_2 enriched in *TAL1* deregulated samples occurred with the known late cortical maturation stage arrests. The plasticity induced by DNAm in tumor progression is further highlighted by observing hypomethylated disease enhancers and maturation arrest stages."

Using Random Forest model, we defined a prognostic predictor consisting of only seven differentially methylated probes. This methylation-based predictor could be an interesting tool to refine risk group at the time of diagnosis and it would be important to validate its accuracy in an independent cohort of patients.

Unlike previous transcriptional studies, we were able to identify patient clusters with clinical significance and an unexpected subgroup of hypermethylated patients (C_5) who showed poor

prognosis and could benefit from targeted epigenetic therapies (Soulier et al. 2005; Homminga et al. 2011).

4.1.3 Conclusion

In the current doctoral thesis, a comprehensive analysis of DNA methylation dynamics during human thymopoiesis is performed. Results show that T-cell differentiation is characterized by the gradual loss of methylation, primarily occurring at genomic regions associated with the *NOTCH* and *MYB* binding sites. Using rigorous statistical analyses, regulatory regions associated with the thymopoiesis are established which, serves as an epigenetic atlas for intra-thymic T-cell development. These regions are distinct from BM-derived cell types and conserved across mature peripheral T-lymphocytes.

Besides, the DNAm landscape of a rare cohort of adult T-ALL is studied, resulting in identifying five distinct epigenetic subtypes, including a novel subgroup with co-occurring *DNMT3A/IDH2* mutations (C_1). The subtypes are characterized by specific maturation arrest stages and their oncogenic drivers **(Figure 34)**. Using gene expression data, subtype-specific gene signatures were established. Moreover, integrative analysis with ChIP-seq data revealed the hypomethylation of oncogenic enhancer elements and hypermethylation of T-cell developmental associated regulatory regions. Furthermore, we define the epigenetic *barcode* (random forest models) for each cluster to be used for de-novo subtype prediction of newly diagnosed T-ALLs. Finally, we identify a novel hypermethylated subtype that can be potentially targeted with DNA hypomethylating agents.

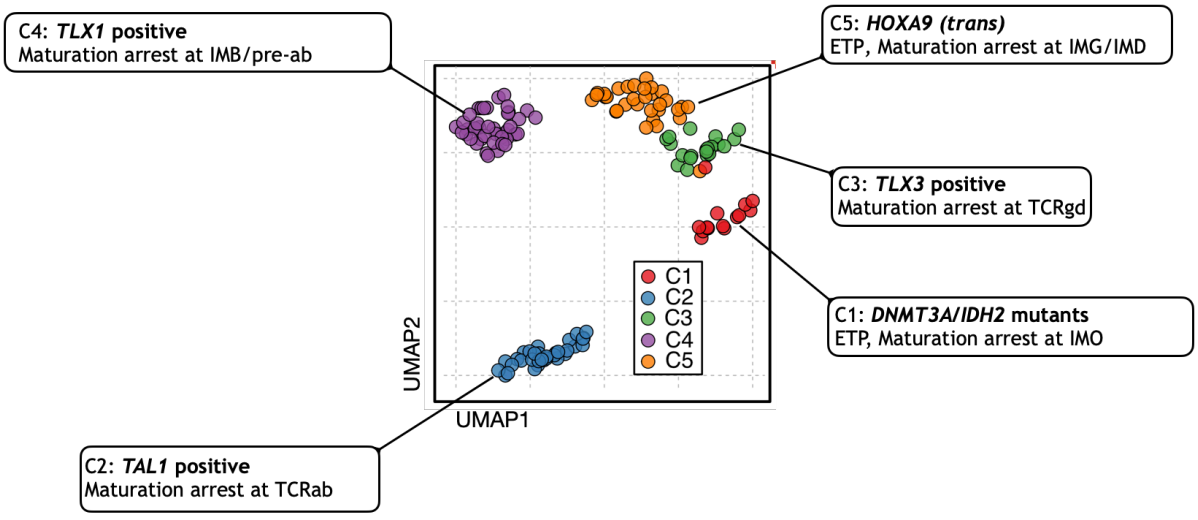

Figure 34. Summary of DNAm based T-ALL subtypes.

Our combined analysis of thymopoiesis and adult T-ALL shows that the DNAm during normal thymopoiesis is linear; however, an oncogenic hit during the early developmental stages results in the differentiation blockade. “The results further highlight two intertwined DNA methylation differences in T-ALL: those preexisting in the cell-of-origin relating to the T-cell differentiation and the T-ALL acquired neoplastic changes.” These overall changes can be used to decipher the ordering and origin of T-ALL subgroups **(Figure 35)**.

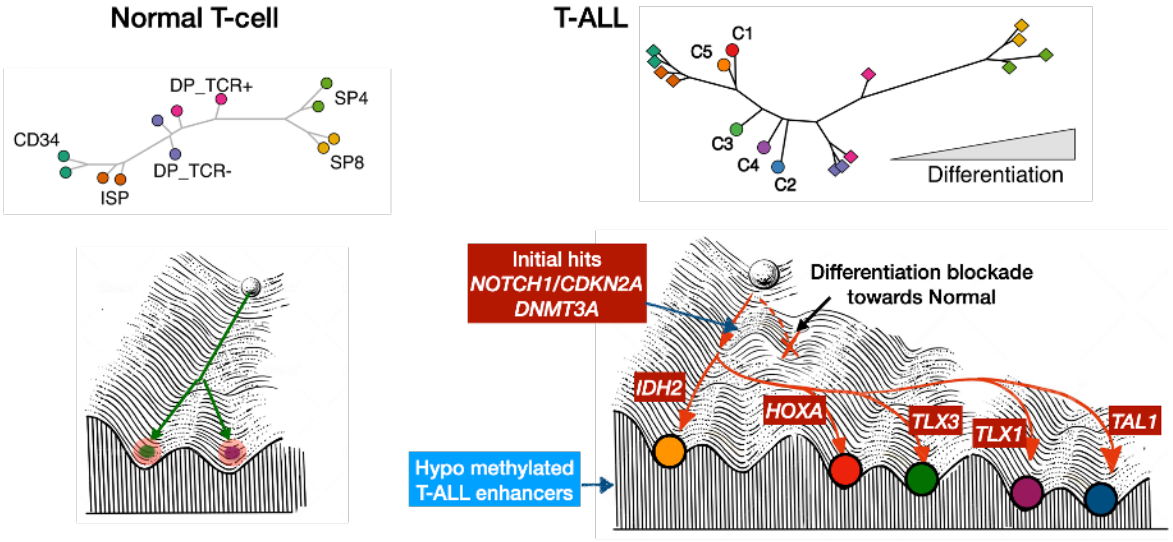

Figure 35. Intertwined DNAm methylomes predict maturation arrest stages of T-ALL.

A visual summary of the DNAm dynamics during normal T-cell development (left) and the oncogenic hits leading to distinct T-ALL subgroups (right). The hypo-methylated pathogenic enhancer elements characterize the T-ALL subgroups. The visualization is inspired from classic Waddington’s ‘ball rolling down the hill’ epigenetic model.

5 MATERIALS AND METHODS

5.1 Intra-thymic and T-ALL samples

5.1.1 Intra-thymic cell types

To study the DNAm dynamics during the thymopoiesis we isolated seven distinct intra-thymic cell types using FACS. Cell-types included rare CD34+ early thymic progenitor cells, TCR negative, TCR positive, mature CD4+ and CD8+ cells **(Table** 5**5)**.

Thymic sub-population	Number of thymi (biological replicates)
CD34+ CD1A-	4
CD34+ CD1A+	4
Immature single CD4+ (4ISP)	4
Early cortical CD4+CD8+CD3-TCR-	2
Late cortical CD4+CD8+CD3+TCR+	2
CD4+CD8-	2
CD4-CD8+	2

Table 5. FACS sorted intra-thymic cell types.

FACS sorting and library preparation performed by Aurore Touzart, Marion Bähr, and Dieter Weichenhan

5.1.2 Adult T-ALL cohort

“A total of 143 adult patients diagnosed with T-ALL (15-60 years old) from two French ALL cooperative groups (*N* = 22 from GRAALL-2003 and *N* = 121 from GRAALL-2005) had their blood or bone marrow analyzed. Between November 2003 and November 2005, the GRAALL-2003 protocol was a multicenter Phase II trial that enrolled 76 adults with T-ALL, 50 of whom had enough diagnostic tumor material for NGS study. Still, only 22 had enough high-quality DNA for the EPIC array (Huguet et al. 2018). The GRAALL-2005 Phase III multicenter randomized trial was very similar to the GRAALL-2003 trial. During induction and late intensification, a randomized evaluation of an intensified sequence of hyper-fractionated cyclophosphamide was added (Maury et al. 2016). In the GRAALL-2005 study, 261 adults with T-ALL were randomized between May 2006 and May 2010, with 185 having diagnostic material available for the NGS study. Despite this, only 121 people had enough high-quality DNA to use in the EPIC array. The availability of high-quality DNA was the primary criterion for inclusion in this study. The 143 patients' survival rates were comparable to those of the remaining 194 T-ALL patients. The study cohort's initial white blood cell count (WBC) was higher, as expected in retrospective studies **(Table 6)**.”

GRAALL (2003/05)	**Study Cohort (*N*=143)**	**Non-investigated (*N*=194)**	***P*-value**
Baseline characteristics			
Male	107	132	0.18
Median Age (range) - Years	29.9 (16.3-59.1)	34.1 (16.8-59.5)	0.01
Median WBC ccount (Range)	40.5 (0.9-645.0)	19.8 (0.9-573.0)	<0.001
CNS involvement – no/total (%)	20/143 (14%)	15/194 (8%)	0.07
Outcome characteristics			
Prednisone response – no/total (%)	79/143 (55%)	126/194 (65%%)	0.09
CR -– no/total (%)	134/143 (94%)	181/194 (93%)	0.99
Allo-SCT - – no/total (%)	53/143 (37%)	48/194 (30%)	0.20
5y-EFS (95%CI)	55% (47-63)	57% (50-64)	0.72
5y-OS (95%CI)	60% (51-68)	67% (60-73)	0.30

(GRAALL-2003/05 trials). ***WBC****, white blood cell count;* ***CNS****, central nervous system;* ***CR****, complete remission;* ***EFS****, event-free survival;* ***OS****, overall survival;* ***CI****, confidence interval;* ***Allo-SCT****, allogeneic stem cell transplantation.*

Table 6. Clinical characteristics and outcome of the study cohort versus non-investigated patients[^].

"More than 80% of the blasts were found in all of the samples. As previously stated, phenotypic and oncogenetic characteristics were obtained (Bergeron et al. 2007; Asnafi et al. 2009; Bond et al. 2016). ETP- ALL is defined as previously described using the classic immunophenotypic criteria: reduced or no expression of CD1a, CD5, and CD8, and positivity for at least one of the following antigens: CD34, CD117, HLA-DR, CD13, CD33, CD11b, or CD65 (Coustan-Smith et al. 2009; Bond et al. 2017). Informed consent was obtained from all patients at enrolment. All trials were conducted per the declaration of Helsinki, approved by local and multicenter research ethical committees. The GRAALL-2003 and -2005 studies were registered at *http://www.clinicaltrials.gov* as *#NCT00222027*[iv] and *#NCT00327678*[v], respectively. A complete summary of the cohort is described in the **Table 7."**

[^] Table copied from the soon-to-be published joint manuscript: Epigenetic blueprint identifies poor outcome and hypomethylating agent-responsive T-ALL subgroup

[iv] https://clinicaltrials.gov/ct2/show/NCT00222027

[v] https://clinicaltrials.gov/ct2/show/NCT00327678

Characteristic	**Value[&]**
Median age at study entry (range) – year	29.9 (16.3-59.1)
Sex ratio (Male/Female) – no.	107/36
T-cell Receptor subsets analyzed – no./total no. (%)	
Immature (IM0. IMδ. IMγ)	33/127 (26)
IMβ/pre-αβ	66/127 (52)
TCRαβ+	14/127 (11)
TCRγδ+	14/127 (11)
Early T-cell precursor (ETP) Immunophenotype – no./total no. (%)	25/125 (20)
High risk patients* – no./total no. (%)	61/143 (43)
Oncogenetic category – no./total no. (%)	
TLX1	28/137 (20)
TLX3	19/137 (14)
SIL-TAL1/TAL1-neoenhancer	17/137 (12)
CALM-AF10	6/137 (4)
None of the above	67/137 (49)
HOXA deregulation[+] – no./total no. (%)	
Cis	7/127 (6)
Trans	17/127 (13)
Unknown	11/127 (9)
Mutations in epigenetic factors – no./total no. (%)	
DNMT3A	20/143 (14)
IDH1	3/143 (2)
IDH2	6/143 (4)
TET2	7/143 (5)
TET3	4/143 (3)
Early response – no./total no. (%)	
Prednisone response	80/143 (56)
Bone marrow response	78/140 (56)
Complete remission	134/143 (94)

[&]Percentages may not total 100 because of rounding up.

*High risk: *NOTCH1/FBXW7*^{WT} OR *NOTCH1*^{mut} + [*KRAS, NRAS, PTEN*]^{mut}

[+]**Cis**: *HOXA* overexpression under the influence of TCRβ enhancer

Trans: *HOXA* overexpression under the result of *SET-NUP214, MLLT10,* or MLL fusion

Table 7: Characteristics of the 143 T-ALL patients[^]

[^] Table copied from the soon-to-be published joint manuscript: Epigenetic blueprint identifies poor outcome and hypomethylating agent-responsive T-ALL subgroup

5.2 Data analysis

5.2.1 WGBS analysis

Raw sequencing reads resulting from WGBS were processed with '*trimmomatic*' to remove adapter sequences and other biases such as random priming (Bolger et al. 2014). Reads were further checked for methylation biases near the 5'/3' ends and trimmed whenever necessary. Final QC passed reads were aligned to hg19 reference genome using *bwa-meth*[vi] 'aligner with default arguments. Post alignment PCR de-duplication was done using the *Picard MarkDuplicates*[vii] tool. Final methylation calling was performed using '*methylctools,'* and results were exported as bedgraph files.

5.2.2 Methrix – a comprehensive suite for DNA methylation analysis

To facilitate the analysis of large bedgraph files from WGBS, we developed a computational framework in the R programming language called *methrix (Mayakonda et al. 2020)*. Methrix allows flexible importing of bedgraph or similar bedgraph like tsv files in a systematic manner. An overview of the package structure and usage is shown in **Figure 36**. At the heart of the package, the function *read_bedgraphs* takes care of file format discrepancies while aggregating them into a single matrix-like object. Another benefit includes adding uncovered CpGs from the reference genome, which results in a homogenous output. Support for a large cohort is made possible by facilitating serialized on-disk arrays, thereby decreasing the memory footprint. Moreover, functions for sub-setting and aggregating over genomic regions are also implemented. *methrix_report* is another process that generates an extensive QC HTML report for the entire cohort.

[vi] https://github.com/brentp/bwa-meth

[vii] https://broadinstitute.github.io/picard/

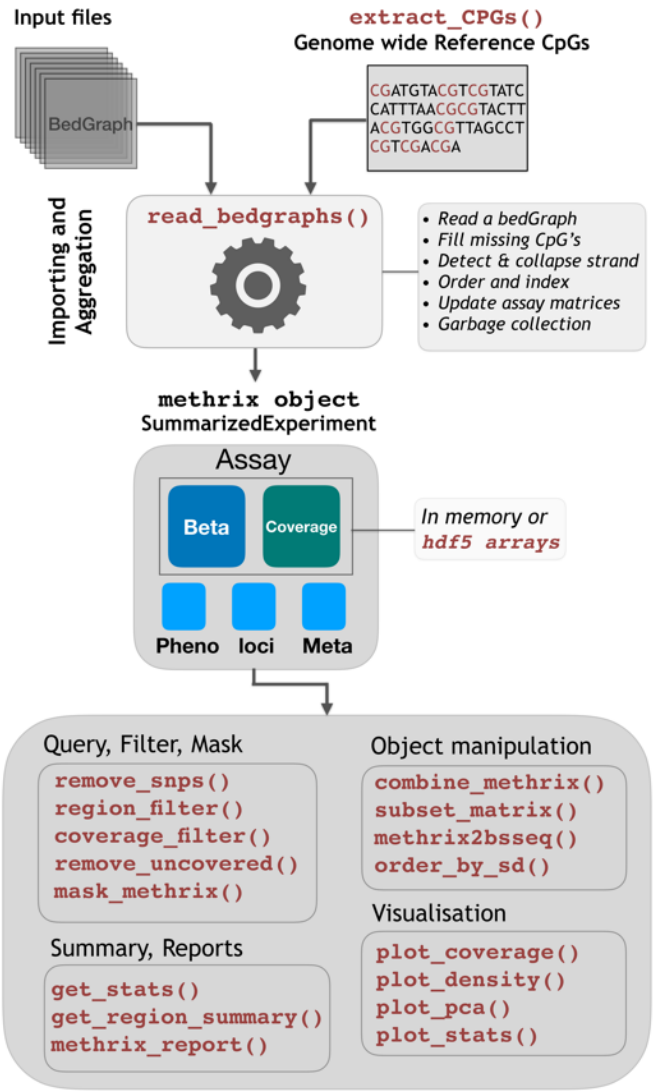

Figure 36. Comprehensive analysis of WGBS with Methrix package.

Data analysis with methrix involves importing bedgraph files along with the reference CpGs using read_bedgraphs(). The resulting Methrix object can be passed several downstream functions for quality control, summarization, and visualization.

Furthermore, *methrix* offers several benefits over the existing tools as shown in the **Table 8** and complements the analysis of WGBS data in an efficient manner. Methrix is made available on Biocodnuctor[viii] and source code is hosted on GitHub[ix].

[viii] https://www.bioconductor.org/packages/release/bioc/html/methrix.html

[ix] https://github.com/CompEpigen/methrix

	methrix	*bsseq*	*methylKit*	*RnBeads*
Supports bedGraph-like files	Yes	No	Yes	Yes
Strand inference and collapsing	Yes	Yes	No	Yes
Filling up of uncovered CpG loci	Yes	No	No	Yes
SNP filtering	Yes	No	No	Partial[^]
Coverage masking	Yes	No	No	No
Extensive interactive html reports	Yes	No	No	Yes
Supports on-disk arrays	Yes[+]	Yes[+]	No	Yes[$]
De novo DMR calling	No	Yes	Yes	No
Number of dependencies*	90	64	103	162

[^] *RnBeads* lacks removal of SNPs based on minor allelic frequencies.
[+] Utilizing *HDF5Array* Bioconductor backend.
[$] Utilizing *ff* CRAN backend.
*Number of package dependencies are as reported on the corresponding Bioconductor landing pages.

Table 8. Comparison of methrix with similar Bioconductor packages[Δ].

Bedgrpah files resulting from WGBS of thymic cell types or for hematopoietic cell types obtained from publicly available database (Farlik et al. 2016), were imported and processed with *methrix* in R. Downstream analysis, such as identifying de-novo differentially methylated regions, was done using the *dmrseq* package (Korthauer et al. 2019). Methylation levels of known regulatory regions associated with the transcriptional regulation were compiled from the Ensemble database[x] and processed with *methrix (Zerbino et al. 2015)*. Differentially

[Δ] Table copied from the published article: Methrix: an R/bioconductor package for systematic aggregation and analysis of bisulfite sequencing data. Mayakonda et al (2020).

[x] http://www.ensembl.org/info/genome/funcgen/regulatory_build.html

methylated Ensemble regulatory regions (FDR < 0.05; |meth| > 20%) were then identified using *limma (Ritchie et al. 2015)*.

5.2.3 EPIC array analysis

IDAT files generated from the T-ALL cohort were processed with the *RnBeads* R package (Assenov et al. 2014). Necessary quality control steps, including removing probes assigned to SNPs and probes located on copy number altered regions, were also removed. Differentially methylated probes were identified using *limma* package using an FDR cut-off of 0.05 and an absolute methylation change of 10%. The batch correction was done using the *Combat* function implemented as a part of the *sva*[xi] package.

5.2.4 Dimensional reduction

"To identify epigenetic clusters, we used Non-negative Matrix Factorization (NMF) on the top 5% of most variable probes (*N* = 38,583) (Gaujoux and Seoighe 2010). NMF decomposes a matrix into two smaller matrices whose product sufficiently recomposes the original matrix. Critical step in NMF is identifying the number of factors. We used a semi-supervised method in which, NMF is run on a range of values and the cophenetic correlation coefficient[xii] (measure of goodness of fit) was determined. An optimal number of clusters was identified for which cophenetic correlation reaches its maximum value (Brunet et al. 2004). Furthermore, to measure the fitness of identified clusters, we repeated the above clustering procedure for varying number of probes (5000 up to 38583). Clusters were stable and reached maximum cophenetic correlation coefficient at *N* = 5. Finally, we used 5 clusters generated by using *N* = 38,583 probes for all downstream analysis. These 5 epigenetic clusters were also robust as measured by rand-index, and samples showed little to no changes in cluster assignments as the number of probes used for clustering increased. Moreover, we randomly sampled 80% of the initial cohort (*N* = 114) and repeated the clustering. Rand Index - a similarity score between two clustering results - was measured between original clusters and new clusters generated on subsamples. Clusters were later visualized using Uniform Manifold Approximation and Projection (UMAP)."

[xi] https://bioconductor.org/packages/release/bioc/html/sva.html

[xii] https://en.wikipedia.org/wiki/Cophenetic_correlation

5.2.5 Gene expression analysis

RNA sequencing for the 44 adult T-ALL samples was performed at the INSERM Paris using the standard library preparation protocol. Fastq files were filtered for quality control and aligned to hg19 human transcriptome using *STAR* aligner (Dobin et al. 2013). Expression level gene counts were generated using the *featurecounts* program. Similarly, counts for intrathymic cell types were obtained from the gene expression omnibus (GEO) (Roels et al. 2020a). Counts were imported into R, and differential expression analysis was done using the *DESeq2* program by accounting batch covariates wherever necessary (*FDR* < 0.1 and logFC > 0.6) (Love et al. 2014).

5.2.6 ChIP-seq analysis

ChIP-sequencing for histone modifications H3K27ac, H3K4me1, and H3K4me3 were performed for 12 T-ALL samples at INSERM Paris using standard library protocol. Fastq files were aligned to the hg19 reference genome using the bwa-mem aligner, and peak calling was done using the MACS2 program (Zhang et al. 2008). Identified peaks were filtered for UCSC blacklisted regions. Different ChIP-seq visualizations such as heatmaps were done using the *deeptools* package (Ramirez et al. 2014). ChIP-seq tracks were generated using the *trackplot*[xiii] R script. Using the dynamics of H3K27ac and H3K4me1 signals, we defined three enhancer classes, namely active, putative, and poised enhancers **(Table 9)**.

	H3K27ac	**H3K4me1**
Active enhancers	+	+
Poised enhancers	-	+
Putative enhancers	+	-
Super Enhancers	+ (12KB)	NA

Table 9. Enhancer classes based on H3K27ac and H3K4me1 marks

[xiii] https://github.com/PoisonAlien/trackplot

Super enhancers are identified using Rank Ordering of Super Enhancers (*ROSE*)[xiv] software by merging consecutive H3K27ac peaks located within 12kb intervals. H3K27ac signals for the merged peaks are ordered by H3K27ac signal intensity, and a mathematical cut-off is estimated based on which the merged enhancers are classified as super or typical enhancers.

5.2.7 Copy number and somatic variant analysis

EPIC arrays were processed with *conumee*[xv] Bioconductor package, which estimates genome-wide copy number variations by merging the signals from methylated and unmethylated probes followed by normalization against the pre-defined controls. Estimated Copy numbers[xvi] are summarized using circular binary segmentation implemented in the *DNAcopy* package (Olshen et al. 2004). Segmentation results from the entire T-ALL cohort (*N* = 143) were summarized using *GISTIC*, which identifies the recurrent copy number alterations across the cohort (Olshen et al. 2004; Mermel et al. 2011). Oncoplots, pathway analysis, and other visualizations for somatic variants were performed with *maftools* Bioconductor package (Mayakonda and Koeffler 2016).

5.2.8 Trajectory analysis

Hematopoiesis and thymopoiesis ontogeny analysis were done using developmental associated EPIC probes or the DMRs identified from WGBS analysis. A data matrix containing aggregated methylation values for regions of interest was generated. The Manhattan distance between the samples was estimated using the *dist* function in R. The distance matrix was used for neighbor-joining and phylogenetic tree construction with the *ape*[xvii] package.

5.2.9 Survival analysis

All statistical analyses are performed in R statistical environment using multitude of software packages (version 4.4)[xviii]. Uni-variate survival analysis was done using *surv*[xix] package with

[xiv] http://younglab.wi.mit.edu/super_enhancer_code.html

[xv] https://bioconductor.org/packages/release/bioc/html/conumee.html

[xvi] Conumee analysis from IDAT files was performed by Dr. Yassen Assenov

[xvii] https://cran.r-project.org/web/packages/ape/index.html

[xviii] https://cran.r-project.org/

[xix] https://cran.r-project.org/web/packages/survival/index.html

log-rank test. “For multivariate survival analysis, since methylation subgroups were strongly associated with maturation arrest and tumor biology, we only considered age, log(WBC), CNS involvement, prednisone response, and D8 bone marrow response as covariates to avoid multicollinearity. Methylation clusters were reduced to a 3-class variable with intermediate methylation clusters ($C_{(3+4)}$) considered as baseline. Covariates finally used in the multivariate cox model were those associated with outcome in univariate analyses ($P < 0.1$).”

6 REFERENCES

Akira S, Uematsu S, Takeuchi O. 2006. Pathogen recognition and innate immunity. *Cell* **124**: 783-801.

Angeloni A, Bogdanovic O. 2019. Enhancer DNA methylation: implications for gene regulation. *Essays Biochem* **63**: 707-715.

Aries IM, Jerchel IS, van den Dungen RE, van den Berk LC, Boer JM, Horstmann MA, Escherich G, Pieters R, den Boer ML. 2014. EMP1, a novel poor prognostic factor in pediatric leukemia regulates prednisolone resistance, cell proliferation, migration and adhesion. *Leukemia* **28**: 1828-1837.

Ashworth TD, Pear WS, Chiang MY, Blacklow SC, Mastio J, Xu L, Kelliher M, Kastner P, Chan S, Aster JC. 2010. Deletion-based mechanisms of Notch1 activation in T-ALL: key roles for RAG recombinase and a conserved internal translational start site in Notch1. *Blood* **116**: 5455-5464.

Asnafi V, Buzyn A, Le Noir S, Baleydier F, Simon A, Beldjord K, Reman O, Witz F, Fagot T, Tavernier E et al. 2009. NOTCH1/FBXW7 mutation identifies a large subgroup with favorable outcome in adult T-cell acute lymphoblastic leukemia (T-ALL): a Group for Research on Adult Acute Lymphoblastic Leukemia (GRAALL) study. *Blood* **113**: 3918-3924.

Assenov Y, Muller F, Lutsik P, Walter J, Lengauer T, Bock C. 2014. Comprehensive analysis of DNA methylation data with RnBeads. *Nat Methods* **11**: 1138-1140.

Baylin SB, Jones PA. 2016. Epigenetic Determinants of Cancer. *Cold Spring Harb Perspect Biol* **8**.

Begley CG, Aplan PD, Davey MP, Nakahara K, Tchorz K, Kurtzberg J, Hershfield MS, Haynes BF, Cohen DI, Waldmann TA et al. 1989. Chromosomal translocation in a human leukemic stem-cell line disrupts the T-cell antigen receptor delta-chain diversity region and results in a previously unreported fusion transcript. *Proc Natl Acad Sci U S A* **86**: 2031-2035.

Bell JJ, Bhandoola A. 2008. The earliest thymic progenitors for T cells possess myeloid lineage potential. *Nature* **452**: 764-767.

Bell RE, Golan T, Sheinboim D, Malcov H, Amar D, Salamon A, Liron T, Gelfman S, Gabet Y, Shamir R et al. 2016. Enhancer methylation dynamics contribute to cancer plasticity and patient mortality. *Genome Res* **26**: 601-611.

Benetatos L, Vartholomatos G. 2018. Enhancer DNA methylation in acute myeloid leukemia and myelodysplastic syndromes. *Cell Mol Life Sci* **75**: 1999-2009.

Bergeron J, Clappier E, Radford I, Buzyn A, Millien C, Soler G, Ballerini P, Thomas X, Soulier J, Dombret H et al. 2007. Prognostic and oncogenic relevance of TLX1/HOX11 expression level in T-ALLs. *Blood* **110**: 2324-2330.

Bernard O, Guglielmi P, Jonveaux P, Cherif D, Gisselbrecht S, Mauchauffe M, Berger R, Larsen CJ, Mathieu-Mahul D. 1990. Two distinct mechanisms for the SCL gene activation in the t(1;14) translocation of T-cell leukemias. *Genes Chromosomes Cancer* **1**: 194-208.

Bernard OA, Busson-LeConiat M, Ballerini P, Mauchauffe M, Della Valle V, Monni R, Nguyen Khac F, Mercher T, Penard-Lacronique V, Pasturaud P et al. 2001. A new recurrent and specific cryptic translocation, t(5;14)(q35;q32), is associated with expression of the Hox11L2 gene in T acute lymphoblastic leukemia. *Leukemia* **15**: 1495-1504.

Bock C, Beerman I, Lien WH, Smith ZD, Gu H, Boyle P, Gnirke A, Fuchs E, Rossi DJ, Meissner A. 2012. DNA methylation dynamics during in vivo differentiation of blood and skin stem cells. *Mol Cell* **47**: 633-647.

Bodle CR, Schamp JH, O'Brien JB, Hayes MP, Wu M, Doorn JA, Roman DL. 2018. Screen Targeting Lung and Prostate Cancer Oncogene Identifies Novel Inhibitors of RGS17 and Problematic Chemical Substructures. *SLAS Discov* **23**: 363-374.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**: 2114-2120.

Bond J, Graux C, Lhermitte L, Lara D, Cluzeau T, Leguay T, Cieslak A, Trinquand A, Pastoret C, Belhocine M et al. 2017. Early Response-Based Therapy Stratification Improves Survival in Adult Early Thymic Precursor Acute Lymphoblastic Leukemia: A Group for Research on Adult Acute Lymphoblastic Leukemia Study. *J Clin Oncol* **35**: 2683-2691.

Bond J, Marchand T, Touzart A, Cieslak A, Trinquand A, Sutton L, Radford-Weiss I, Lhermitte L, Spicuglia S, Dombret H et al. 2016. An early thymic precursor phenotype predicts outcome exclusively in HOXA-overexpressing adult T-cell

acute lymphoblastic leukemia: a Group for Research in Adult Acute Lymphoblastic Leukemia study. *Haematologica* **101**: 732-740.

Bond J, Touzart A, Lepretre S, Graux C, Bargetzi M, Lhermitte L, Hypolite G, Leguay T, Hicheri Y, Guillerm G et al. 2019. DNMT3A mutation is associated with increased age and adverse outcome in adult T-cell acute lymphoblastic leukemia. *Haematologica* **104**: 1617-1625.

Borssen M, Haider Z, Landfors M, Noren-Nystrom U, Schmiegelow K, Asberg AE, Kanerva J, Madsen HO, Marquart H, Heyman M et al. 2016. DNA Methylation Adds Prognostic Value to Minimal Residual Disease Status in Pediatric T-Cell Acute Lymphoblastic Leukemia. *Pediatr Blood Cancer* **63**: 1185-1192.

Borssen M, Palmqvist L, Karrman K, Abrahamsson J, Behrendtz M, Heldrup J, Forestier E, Roos G, Degerman S. 2013. Promoter DNA methylation pattern identifies prognostic subgroups in childhood T-cell acute lymphoblastic leukemia. *PLoS One* **8**: e65373.

Brunet JP, Tamayo P, Golub TR, Mesirov JP. 2004. Metagenes and molecular pattern discovery using matrix factorization. *Proc Natl Acad Sci U S A* **101**: 4164-4169.

Cancer Genome Atlas Research N. 2013. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N Engl J Med* **368**: 2059-2074.

Cancer Genome Atlas Research N Ley TJ Miller C Ding L Raphael BJ Mungall AJ Robertson A Hoadley K Triche TJ, Jr. Laird PW et al. 2013. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N Engl J Med* **368**: 2059-2074.

Capper D Jones DTW Sill M Hovestadt V Schrimpf D Sturm D Koelsche C Sahm F Chavez L Reuss DE et al. 2018. DNA methylation-based classification of central nervous system tumours. *Nature* **555**: 469-474.

Challen GA, Sun D, Jeong M, Luo M, Jelinek J, Berg JS, Bock C, Vasanthakumar A, Gu H, Xi Y et al. 2011. Dnmt3a is essential for hematopoietic stem cell differentiation. *Nat Genet* **44**: 23-31.

Cheminant M, Bruneau J, Malamut G, Sibon D, Guegan N, van Gils T, Cording S, Trinquand A, Verkarre V, Lhermitte L et al. 2019. NKp46 is a diagnostic biomarker and may be a therapeutic target in gastrointestinal T-cell lymphoproliferative diseases: a CELAC study. *Gut* **68**: 1396-1405.

Chen, Luo L, Liang C. 2018a. Aberrant S100A16 expression might be an independent prognostic indicator of unfavorable survival in non-small cell lung adenocarcinoma. *PLoS One* **13**: e0197402.

Chen B, Jiang L, Zhong ML, Li JF, Li BS, Peng LJ, Dai YT, Cui BW, Yan TQ, Zhang WN et al. 2018b. Identification of fusion genes and characterization of transcriptome features in T-cell acute lymphoblastic leukemia. *Proc Natl Acad Sci U S A* **115**: 373-378.

Chen Q, Yang CY, Tsan JT, Xia Y, Ragab AH, Peiper SC, Carroll A, Baer R. 1990. Coding sequences of the tal-1 gene are disrupted by chromosome translocation in human T cell leukemia. *J Exp Med* **172**: 1403-1408.

Chorzalska A, Ahsan N, Rao RSP, Roder K, Yu X, Morgan J, Tepper A, Hines S, Zhang P, Treaba DO et al. 2018. Overexpression of Tpl2 is linked to imatinib resistance and activation of MEK-ERK and NF-kappaB pathways in a model of chronic myeloid leukemia. *Mol Oncol* **12**: 630-647.

Cieslak A, Charbonnier G, Tesio M, Mathieu EL, Belhocine M, Touzart A, Smith C, Hypolite G, Andrieu GP, Martens JHA et al. 2020. Blueprint of human thymopoiesis reveals molecular mechanisms of stage-specific TCR enhancer activation. *J Exp Med* **217**.

Condorelli GL, Facchiano F, Valtieri M, Proietti E, Vitelli L, Lulli V, Huebner K, Peschle C, Croce CM. 1996. T-cell-directed TAL-1 expression induces T-cell malignancies in transgenic mice. *Cancer Res* **56**: 5113-5119.

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

D'Angio M, Valsecchi MG, Testi AM, Conter V, Nunes V, Parasole R, Colombini A, Santoro N, Varotto S, Caniglia M et al. 2015. Clinical features and outcome of SIL/TAL1-positive T-cell acute lymphoblastic leukemia in children and adolescents: a 10-year experience of the AIEOP group. *Haematologica* **100**: e10-13.

de Almeida RA, Heuser T, Blaschke R, Bartram CR, Janssen JW. 2006. Control of MYEOV protein synthesis by upstream open reading frames. *J Biol Chem* **281**: 695-704.

De Keersmaecker K, Ferrando AA. 2011. TLX1-induced T-cell acute lymphoblastic leukemia. *Clin Cancer Res* **17**: 6381-6386.

Della Gatta G, Palomero T, Perez-Garcia A, Ambesi-Impiombato A, Bansal M, Carpenter ZW, De Keersmaecker K, Sole X, Xu L, Paietta E et al. 2012. Reverse engineering of

TLX oncogenic transcriptional networks identifies RUNX1 as tumor suppressor in T-ALL. *Nat Med* **18**: 436-440.

Deng J, Guo J, Guo X, Hou Y, Xie X, Sun C, Zhang R, Yu X, Liang H. 2016. Mediation of the malignant biological characteristics of gastric cancer cells by the methylated CpG islands in RNF180 DNA promoter. *Oncotarget* **7**: 43461-43474.

DiNardo CD, Stein EM, de Botton S, Roboz GJ, Altman JK, Mims AS, Swords R, Collins RH, Mannis GN, Pollyea DA et al. 2018. Durable Remissions with Ivosidenib in IDH1-Mutated Relapsed or Refractory AML. *N Engl J Med* **378**: 2386-2398.

DiNardo CD, Wei AH. 2020. How I treat acute myeloid leukemia in the era of new drugs. *Blood* **135**: 85-96.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**: 15-21.

Erarslan-Uysal B, Kunz JB, Rausch T, Richter-Pechanska P, van Belzen IA, Frismantas V, Bornhauser B, Ordonez-Rueada D, Paulsen M, Benes V et al. 2020. Chromatin accessibility landscape of pediatric T-lymphoblastic leukemia and human T-cell precursors. *EMBO Mol Med* **12**: e12104.

Farlik M, Halbritter F, Muller F, Choudry FA, Ebert P, Klughammer J, Farrow S, Santoro A, Ciaurro V, Mathur A et al. 2016. DNA Methylation Dynamics of Human Hematopoietic Stem Cell Differentiation. *Cell stem cell* **19**: 808-822.

Feinberg AP, Koldobskiy MA, Gondor A. 2016. Epigenetic modulators, modifiers and mediators in cancer aetiology and progression. *Nat Rev Genet* **17**: 284-299.

Ferrando AA, Neuberg DS, Dodge RK, Paietta E, Larson RA, Wiernik PH, Rowe JM, Caligiuri MA, Bloomfield CD, Look AT. 2004. Prognostic importance of TLX1 (HOX11) oncogene expression in adults with T-cell acute lymphoblastic leukaemia. *Lancet* **363**: 535-536.

Ferrando AA, Neuberg DS, Staunton J, Loh ML, Huard C, Raimondi SC, Behm FG, Pui CH, Downing JR, Gilliland DG et al. 2002. Gene expression signatures define novel oncogenic pathways in T cell acute lymphoblastic leukemia. *Cancer Cell* **1**: 75-87.

Figueroa ME, Abdel-Wahab O, Lu C, Ward PS, Patel J, Shih A, Li Y, Bhagwat N, Vasanthakumar A, Fernandez HF et al. 2010a. Leukemic IDH1 and IDH2 mutations result in a hypermethylation phenotype, disrupt TET2 function, and impair hematopoietic differentiation. *Cancer Cell* **18**: 553-567.

Figueroa ME, Lugthart S, Li Y, Erpelinck-Verschueren C, Deng X, Christos PJ, Schifano E, Booth J, van Putten W, Skrabanek L et al. 2010b. DNA methylation signatures identify biologically distinct subtypes in acute myeloid leukemia. *Cancer Cell* **17**: 13-27.

Garcia-Fernandez J. 2005. The genesis and evolution of homeobox gene clusters. *Nat Rev Genet* **6**: 881-892.

Gardin C, Dombret H. 2017. Hypomethylating Agents as a Therapy for AML. *Curr Hematol Malig Rep* **12**: 1-10.

Gaujoux R, Seoighe C. 2010. A flexible R package for nonnegative matrix factorization. *BMC bioinformatics* **11**: 367.

Geenen V. 2017. [History of the thymus: from an "accident of evolution" to the programming of immunological self-tolerance]. *Med Sci (Paris)* **33**: 653-663.

Greenberg MVC, Bourc'his D. 2019. The diverse roles of DNA methylation in mammalian development and disease. *Nat Rev Mol Cell Biol* **20**: 590-607.

Grossmann V, Haferlach C, Weissmann S, Roller A, Schindela S, Poetzinger F, Stadler K, Bellos F, Kern W, Haferlach T et al. 2013. The molecular profile of adult T-cell acute lymphoblastic leukemia: mutations in RUNX1 and DNMT3A are associated with poor prognosis in T-ALL. *Genes Chromosomes Cancer* **52**: 410-422.

Gu S, Lin S, Ye D, Qian S, Jiang D, Zhang X, Li Q, Yang J, Ying X, Li Z et al. 2019. Genome-wide methylation profiling identified novel differentially hypermethylated biomarker MPPED2 in colorectal cancer. *Clin Epigenetics* **11**: 41.

Guo W, Zhu L, Zhu R, Chen Q, Wang Q, Chen JQ. 2019. A four-DNA methylation biomarker is a superior predictor of survival of patients with cutaneous melanoma. *Elife* **8**.

Haider Z, Larsson P, Landfors M, Kohn L, Schmiegelow K, Flaegstad T, Kanerva J, Heyman M, Hultdin M, Degerman S. 2019. An integrated transcriptome analysis in T-cell acute lymphoblastic leukemia links DNA methylation subgroups to dysregulated TAL1 and ANTP homeobox gene expression. *Cancer Med* **8**: 311-324.

He X, Xu X, Khan AQ, Ling W. 2017. High Expression of S100A6 Predicts Unfavorable Prognosis of Lung Squamous Cell Cancer. *Med Sci Monit* **23**: 5011-5017.

Heerema NA, Sather HN, Sensel MG, La MK, Hutchinson RJ, Nachman JB, Reaman GH, Lange BJ, Steinherz PG, Bostrom BC et al. 2002. Abnormalities of chromosome bands 15q13-15 in childhood acute lymphoblastic leukemia. *Cancer* **94**: 1102-1110.

Hnisz D, Abraham BJ, Lee TI, Lau A, Saint-Andre V, Sigova AA, Hoke HA, Young RA. 2013. Super-enhancers in the control of cell identity and disease. *Cell* **155**: 934-947.

Homminga I, Pieters R, Langerak AW, de Rooi JJ, 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.

Hosoya T, Kuroha T, Moriguchi T, Cummings D, Maillard I, Lim KC, Engel JD. 2009. GATA-3 is required for early T lineage progenitor development. *J Exp Med* **206**: 2987-3000.

Hou HA, Kuo YY, Liu CY, Chou WC, Lee MC, Chen CY, Lin LI, Tseng MH, Huang CF, Chiang YC et al. 2012. DNMT3A mutations in acute myeloid leukemia: stability during disease evolution and clinical implications. *Blood* **119**: 559-568.

Huemer F, Melchardt T, Jansko B, Wahida A, Jilg S, Jost PJ, Klieser E, Steiger K, Magnes T, Pleyer L et al. 2019. Durable remissions with venetoclax monotherapy in secondary AML refractory to hypomethylating agents and high expression of BCL-2 and/or BIM. *Eur J Haematol* **102**: 437-441.

Huguet F, Chevret S, Leguay T, Thomas X, Boissel N, Escoffre-Barbe M, Chevallier P, Hunault M, Vey N, Bonmati C et al. 2018. Intensified Therapy of Acute Lymphoblastic Leukemia in Adults: Report of the Randomized GRAALL-2005 Clinical Trial. *J Clin Oncol* **36**: 2514-2523.

Hunger SP, Mullighan CG. 2015. Acute Lymphoblastic Leukemia in Children. *N Engl J Med* **373**: 1541-1552.

Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, Tyson JR, Beggs AD, Dilthey AT, Fiddes IT et al. 2018. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat Biotechnol* **36**: 338-345.

Jaiswal S, Fontanillas P, Flannick J, Manning A, Grauman PV, Mar BG, Lindsley RC, Mermel CH, Burtt N, Chavez A et al. 2014. Age-related clonal hematopoiesis associated with adverse outcomes. *N Engl J Med* **371**: 2488-2498.

Jeong M, Sun D, Luo M, Huang Y, Challen GA, Rodriguez B, Zhang X, Chavez L, Wang H, Hannah R et al. 2014. Large conserved domains of low DNA methylation maintained by Dnmt3a. *Nat Genet* **46**: 17-23.

Ji H, Ehrlich LI, Seita J, Murakami P, Doi A, Lindau P, Lee H, Aryee MJ, Irizarry RA, Kim K et al. 2010. Comprehensive methylome map of lineage commitment from haematopoietic progenitors. *Nature* **467**: 338-342.

Jiang X, Xie H, Dou Y, Yuan J, Zeng D, Xiao S. 2020. Expression and function of FRA1 protein in tumors. *Mol Biol Rep* **47**: 737-752.

Jung N, Dai B, Gentles AJ, Majeti R, Feinberg AP. 2015. An LSC epigenetic signature is largely mutation independent and implicates the HOXA cluster in AML pathogenesis. *Nat Commun* **6**: 8489.

Kappes DJ, He X, He X. 2005. CD4-CD8 lineage commitment: an inside view. *Nat Immunol* **6**: 761-766.

Kastner P, Chan S, Vogel WK, Zhang LJ, Topark-Ngarm A, Golonzhka O, Jost B, Le Gras S, Gross MK, Leid M. 2010. Bcl11b represses a mature T-cell gene expression program in immature CD4(+)CD8(+) thymocytes. *Eur J Immunol* **40**: 2143-2154.

Kennedy MA, Gonzalez-Sarmiento R, Kees UR, Lampert F, Dear N, Boehm T, Rabbitts TH. 1991. HOX11, a homeobox-containing T-cell oncogene on human chromosome 10q24. *Proc Natl Acad Sci U S A* **88**: 8900-8904.

Ko M, Huang Y, Jankowska AM, Pape UJ, Tahiliani M, Bandukwala HS, An J, Lamperti ED, Koh KP, Ganetzky R et al. 2010. Impaired hydroxylation of 5-methylcytosine in myeloid cancers with mutant TET2. *Nature* **468**: 839-843.

Koch U, Radtke F. 2011. Mechanisms of T cell development and transformation. *Annu Rev Cell Dev Biol* **27**: 539-562.

Kondo M, Wagers AJ, Manz MG, Prohaska SS, Scherer DC, Beilhack GF, Shizuru JA, Weissman IL. 2003. Biology of hematopoietic stem cells and progenitors: implications for clinical application. *Annu Rev Immunol* **21**: 759-806.

Konopleva M, Pollyea DA, Potluri J, Chyla B, Hogdal L, Busman T, McKeegan E, Salem AH, Zhu M, Ricker JL et al. 2016. Efficacy and Biological Correlates of Response in a Phase II Study of Venetoclax Monotherapy in Patients with Acute Myelogenous Leukemia. *Cancer Discov* **6**: 1106-1117.

Korthauer K, Chakraborty S, Benjamini Y, Irizarry RA. 2019. Detection and accurate false discovery rate control of differentially methylated regions from whole genome bisulfite sequencing. *Biostatistics* **20**: 367-383.

Kramer AC, Kothari A, Wilson WC, Celik H, Nikitas J, Mallaney C, Ostrander EL, Eultgen E, Martens A, Valentine MC et al. 2017. Dnmt3a regulates T-cell development and suppresses T-ALL transformation. *Leukemia* **31**: 2479-2490.

Kraszewska MD, Dawidowska M, Larmonie NS, Kosmalska M, Sedek L, Szczepaniak M, Grzeszczak W, Langerak AW, Szczepanski T, Witt M et al. 2012. DNA methylation pattern is altered in childhood T-cell acute lymphoblastic leukemia patients as compared with normal thymic subsets: insights into CpG island methylator phenotype in T-ALL. *Leukemia* **26**: 367-371.

Lai B, Lai Y, Zhang Y, Zhou M, Sheng L, OuYang G. 2020. The Solute Carrier Family 2 Genes Are Potential Prognostic Biomarkers in Acute Myeloid Leukemia. *Technol Cancer Res Treat* **19**: 1533033819894308.

Lee DJ, Schonleben F, Banuchi VE, Qiu W, Close LG, Assaad AM, Su GH. 2010. Multiple tumor-suppressor genes on chromosome 3p contribute to head and neck squamous cell carcinoma tumorigenesis. *Cancer Biol Ther* **10**: 689-693.

Lehmann BD, Shaver TM, Johnson DB, Li Z, Gonzalez-Ericsson PI, Sanchez V, Shyr Y, Sanders ME, Pietenpol JA. 2019. Identification of Targetable Recurrent MAP3K8 Rearrangements in Melanomas Lacking Known Driver Mutations. *Mol Cancer Res* **17**: 1842-1853.

Lesesve JF, Tardy S, Frotscher B, Latger-Cannard V, Feugier P, De Carvalho Bittencourt M. 2015. Combination of CD160 and CD200 as a useful tool for differential diagnosis between chronic lymphocytic leukemia and other mature B-cell neoplasms. *Int J Lab Hematol* **37**: 486-494.

Ley TJ, Ding L, Walter MJ, McLellan MD, Lamprecht T, Larson DE, Kandoth C, Payton JE, Baty J, Welch J et al. 2010. DNMT3A mutations in acute myeloid leukemia. *N Engl J Med* **363**: 2424-2433.

Li E, Zhang Y. 2014. DNA methylation in mammals. *Cold Spring Harb Perspect Biol* **6**: a019133.

Li L, Luo HS. 2018. G-Protein Signaling Protein-17 (RGS17) Is Upregulated and Promotes Tumor Growth and Migration in Human Colorectal Carcinoma. *Oncol Res* **26**: 27-35.

Li YC, Wang Y, Li DD, Zhang Y, Zhao TC, Li CF. 2018. Procaine is a specific DNA methylation inhibitor with anti-tumor effect for human gastric cancer. *J Cell Biochem* **119**: 2440-2449.

Liao Z, Wang X, Wang X, Li L, Lin D. 2017. DEPDC7 inhibits cell proliferation, migration and invasion in hepatoma cells. *Oncol Lett* **14**: 7332-7338.

Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z, Ngo QM et al. 2009. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462**: 315-322.

Litzow MR, Ferrando AA. 2015. How I treat T-cell acute lymphoblastic leukemia in adults. *Blood* **126**: 833-841.

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

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

Lowery MA, Burris HA, 3rd, Janku F, Shroff RT, Cleary JM, Azad NS, Goyal L, Maher EA, Gore L, Hollebecque A et al. 2019. Safety and activity of ivosidenib in patients with IDH1-mutant advanced cholangiocarcinoma: a phase 1 study. *Lancet Gastroenterol Hepatol* **4**: 711-720.

Luc S, Luis TC, Boukarabila H, Macaulay IC, Buza-Vidas N, Bouriez-Jones T, Lutteropp M, Woll PS, Loughran SJ, Mead AJ et al. 2012. The earliest thymic T cell progenitors sustain B cell and myeloid lineage potential. *Nat Immunol* **13**: 412-419.

Lv J, Zhu P, Zhang X, Zhang L, Chen X, Lu F, Yu Z, Liu S. 2017. PCDH9 acts as a tumor suppressor inducing tumor cell arrest at G0/G1 phase and is frequently methylated in hepatocellular carcinoma. *Mol Med Rep* **16**: 4475-4482.

Maat W, Beiboer SH, Jager MJ, Luyten GP, Gruis NA, van der Velden PA. 2008. Epigenetic regulation identifies RASEF as a tumor-suppressor gene in uveal melanoma. *Invest Ophthalmol Vis Sci* **49**: 1291-1298.

Mansour MR, Abraham BJ, Anders L, Berezovskaya A, Gutierrez A, Durbin AD, Etchin J, Lawton L, Sallan SE, Silverman LB et al. 2014. Oncogene regulation. An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element. *Science* **346**: 1373-1377.

Mansour MR, He S, Li Z, Lobbardi R, Abraham BJ, Hug C, Rahman S, Leon TE, Kuang YY, Zimmerman MW et al. 2018. JDP2: An oncogenic bZIP transcription factor in T cell acute lymphoblastic leukemia. *J Exp Med* **215**: 1929-1945.

Maury S, Chevret S, Thomas X, Heim D, Leguay T, Huguet F, Chevallier P, Hunault M, Boissel N, Escoffre-Barbe M et al. 2016. Rituximab in B-Lineage Adult Acute Lymphoblastic Leukemia. *N Engl J Med* **375**: 1044-1053.

Mayakonda A, Koeffler HP. 2016. Maftools: Efficient analysis, visualization and summarization of MAF files from large-scale cohort based cancer studies. *bioRxiv* doi:10.1101/052662.

Mayakonda A, Schonung M, Hey J, Batra RN, Feuerstein-Akgoz C, Kohler K, Lipka DB, Sotillo R, Plass C, Lutsik P et al. 2020. Methrix: an R/bioconductor package for systematic aggregation and analysis of bisulfite sequencing data. *Bioinformatics* doi:10.1093/bioinformatics/btaa1048.

Mayle A, Yang L, Rodriguez B, Zhou T, Chang E, Curry CV, Challen GA, Li W, Wheeler D, Rebel VI et al. 2015. Dnmt3a loss predisposes murine hematopoietic stem cells to malignant transformation. *Blood* **125**: 629-638.

McGuire EA, Hockett RD, Pollock KM, Bartholdi MF, O'Brien SJ, Korsmeyer SJ. 1989. The t(11;14)(p15;q11) in a T-cell acute lymphoblastic leukemia cell line activates multiple transcripts, including Ttg-1, a gene encoding a potential zinc finger protein. *Mol Cell Biol* **9**: 2124-2132.

Medeiros BC, Fathi AT, DiNardo CD, Pollyea DA, Chan SM, Swords R. 2017. Isocitrate dehydrogenase mutations in myeloid malignancies. *Leukemia* **31**: 272-281.

Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R, Getz G. 2011. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol* **12**: R41.

Metzeler KH, Walker A, Geyer S, Garzon R, Klisovic RB, Bloomfield CD, Blum W, Marcucci G. 2012. DNMT3A mutations and response to the hypomethylating agent decitabine in acute myeloid leukemia. *Leukemia* **26**: 1106-1107.

Milani L, Lundmark A, Kiialainen A, Nordlund J, Flaegstad T, Forestier E, Heyman M, Jonmundsson G, Kanerva J, Schmiegelow K et al. 2010. DNA methylation for subtype classification and prediction of treatment outcome in patients with childhood acute lymphoblastic leukemia. *Blood* **115**: 1214-1225.

Miller J. 2020. The function of the thymus and its impact on modern medicine. *Science* **369**.

Moran-Crusio K, Reavie L, Shih A, Abdel-Wahab O, Ndiaye-Lobry D, Lobry C, Figueroa ME, Vasanthakumar A, Patel J, Zhao X et al. 2011. Tet2 loss leads to increased

hematopoietic stem cell self-renewal and myeloid transformation. *Cancer Cell* **20**: 11-24.

Moreaux J, Hose D, Bonnefond A, Reme T, Robert N, Goldschmidt H, Klein B. 2010. MYEOV is a prognostic factor in multiple myeloma. *Exp Hematol* **38**: 1189-1198 e1183.

Newman S, Pappo A, Raimondi S, Zhang J, Barnhill R, Bahrami A. 2019. Pathologic Characteristics of Spitz Melanoma With MAP3K8 Fusion or Truncation in a Pediatric Cohort. *Am J Surg Pathol* **43**: 1631-1637.

Nordlund J, Backlin CL, Wahlberg P, Busche S, Berglund EC, Eloranta ML, Flaegstad T, Forestier E, Frost BM, Harila-Saari A et al. 2013. Genome-wide signatures of differential DNA methylation in pediatric acute lymphoblastic leukemia. *Genome Biol* **14**: r105.

Oakes CC, Seifert M, Assenov Y, Gu L, Przekopowitz M, Ruppert AS, Wang Q, Imbusch CD, Serva A, Koser SD et al. 2016. DNA methylation dynamics during B cell maturation underlie a continuum of disease phenotypes in chronic lymphocytic leukemia. *Nat Genet* **48**: 253-264.

Olshen AB, Venkatraman ES, Lucito R, Wigler M. 2004. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5**: 557-572.

Pai S, Bamodu OA, Lin YK, Lin CS, Chu PY, Chien MH, Wang LS, Hsiao M, Yeh CT, Tsai JT. 2019. CD47-SIRPalpha Signaling Induces Epithelial-Mesenchymal Transition and Cancer Stemness and Links to a Poor Prognosis in Patients with Oral Squamous Cell Carcinoma. *Cells* **8**.

Papaioannou D, Shen C, Nicolet D, McNeil B, Bill M, Karunasiri M, Burke MH, Ozer HG, Yilmaz SA, Zitzer N et al. 2017. Prognostic and biological significance of the proangiogenic factor EGFL7 in acute myeloid leukemia. *Proc Natl Acad Sci U S A* **114**: E4641-E4647.

Pearson JC, Lemons D, McGinnis W. 2005. Modulating Hox gene functions during animal body patterning. *Nat Rev Genet* **6**: 893-904.

Pott S, Lieb JD. 2015. What are super-enhancers? *Nat Genet* **47**: 8-12.

Powell-Jones W, Davies P, Wilson DW, Griffiths K. 1976. Proceedings: Specificity of steroid binding by the oestrogen receptor of rat mammary tumours induced by 7,12-dimethylbenz(a)anthracene. *J Endocrinol* **68**: 30P.

Pui CH, Robison LL, Look AT. 2008. Acute lymphoblastic leukaemia. *Lancet* **371**: 1030-1043.

Pyo JS, Park MJ, Kim CN. 2018. TPL2 expression is correlated with distant metastasis and poor prognosis in colorectal cancer. *Hum Pathol* **79**: 50-56.

Radtke F, Wilson A, Stark G, Bauer M, van Meerwijk J, MacDonald HR, Aguet M. 1999. Deficient T cell fate specification in mice with an induced inactivation of Notch1. *Immunity* **10**: 547-558.

Raetz EA, Teachey DT. 2016. T-cell acute lymphoblastic leukemia. *Hematology Am Soc Hematol Educ Program* **2016**: 580-588.

Ramirez F, Dundar F, Diehl S, Gruning BA, Manke T. 2014. deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res* **42**: W187-191.

Rand AC, Jain M, Eizenga JM, Musselman-Brown A, Olsen HE, Akeson M, Paten B. 2017. Mapping DNA methylation with high-throughput nanopore sequencing. *Nat Methods* **14**: 411-413.

Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. 2015. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* **43**: e47.

Rodriguez RM, Suarez-Alvarez B, Mosen-Ansorena D, Garcia-Peydro M, Fuentes P, Garcia-Leon MJ, Gonzalez-Lahera A, Macias-Camara N, Toribio ML, Aransay AM et al. 2015. Regulation of the transcriptional program by DNA methylation during human alphabeta T-cell development. *Nucleic Acids Res* **43**: 760-774.

Roels J, Kuchmiy A, De Decker M, Strubbe S, Lavaert M, Liang KL, Leclercq G, Vandekerckhove B, Van Nieuwerburgh F, Van Vlierberghe P et al. 2020a. Distinct and temporary-restricted epigenetic mechanisms regulate human alphabeta and gammadelta T cell development. *Nat Immunol* **21**: 1280-1292.

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

Rohle D, Popovici-Muller J, Palaskas N, Turcan S, Grommes C, Campos C, Tsoi J, Clark O, Oldrini B, Komisopoulou E et al. 2013. An inhibitor of mutant IDH1 delays growth and promotes differentiation of glioma cells. *Science* **340**: 626-630.

Roy A, Ramalinga M, Kim OJ, Chijioke J, Lynch S, Byers S, Kumar D. 2017. Multiple roles of RARRES1 in prostate cancer: Autophagy induction and angiogenesis inhibition. *PLoS One* **12**: e0180344.

Royer-Pokora B, Loos U, Ludwig WD. 1991. TTG-2, a new gene encoding a cysteine-rich protein with the LIM motif, is overexpressed in acute T-cell leukaemia with the t(11;14)(p13;q11). *Oncogene* **6**: 1887-1893.

Salas LA, Koestler DC, Butler RA, Hansen HM, Wiencke JK, Kelsey KT, Christensen BC. 2018. An optimized library for reference-based deconvolution of whole-blood biospecimens assayed using the Illumina HumanMethylationEPIC BeadArray. *Genome Biol* **19**: 64.

Sanda T, Lawton LN, Barrasa MI, Fan ZP, Kohlhammer H, Gutierrez A, Ma W, Tatarek J, Ahn Y, Kelliher MA et al. 2012. Core transcriptional regulatory circuit controlled by the TAL1 complex in human T cell acute lymphoblastic leukemia. *Cancer Cell* **22**: 209-221.

Sanda T, Leong WZ. 2017. TAL1 as a master oncogenic transcription factor in T-cell acute lymphoblastic leukemia. *Exp Hematol* **53**: 7-15.

Schatz MC. 2017. Nanopore sequencing meets epigenetics. *Nat Methods* **14**: 347-348.

Schmitt C, Ktorza S, Sarun S, Verpilleux MP, Blanc C, Deugnier MA, Dalloul A, Debre P. 1995. CD34-positive early stages of human T-cell differentiation. *Leuk Lymphoma* **17**: 43-50.

Schmitt TM, Zuniga-Pflucker JC. 2002. Induction of T cell development from hematopoietic progenitor cells by delta-like-1 in vitro. *Immunity* **17**: 749-756.

Sheffield NC, Bock C. 2016. LOLA: enrichment analysis for genomic region sets and regulatory elements in R and Bioconductor. *Bioinformatics* **32**: 587-589.

Shih TC, Fan Y, Kiss S, Li X, Deng XN, Liu R, Chen XJ, Carney R, Chen A, Ghosh PM et al. 2019. Galectin-1 inhibition induces cell apoptosis through dual suppression of CXCR4 and Ras pathways in human malignant peripheral nerve sheath tumors. *Neuro Oncol* **21**: 1389-1400.

Simpson JT, Workman RE, Zuzarte PC, David M, Dursi LJ, Timp W. 2017. Detecting DNA cytosine methylation using nanopore sequencing. *Nat Methods* **14**: 407-410.

Soulier J, Clappier E, Cayuela JM, Regnault A, Garcia-Peydro M, Dombret H, Baruchel A, Toribio ML, Sigaux F. 2005. HOXA genes are included in genetic and biologic networks defining human acute T-cell leukemia (T-ALL). *Blood* **106**: 274-286.

Stein EM, DiNardo CD, Pollyea DA, Fathi AT, Roboz GJ, Altman JK, Stone RM, DeAngelo DJ, Levine RL, Flinn IW et al. 2017. Enasidenib in mutant IDH2 relapsed or refractory acute myeloid leukemia. *Blood* **130**: 722-731.

Stock W, Johnson JL, Stone RM, Kolitz JE, Powell BL, Wetzler M, Westervelt P, Marcucci G, DeAngelo DJ, Vardiman JW et al. 2013. Dose intensification of daunorubicin and cytarabine during treatment of adult acute lymphoblastic leukemia: results of Cancer and Leukemia Group B Study 19802. *Cancer* **119**: 90-98.

Sun X, Zhang Q, Chen W, Hu Q, Lou Y, Fu QH, Zhang JY, Chen YW, Ye LY, Wang Y et al. 2017. Hook1 inhibits malignancy and epithelial-mesenchymal transition in hepatocellular carcinoma. *Tumour Biol* **39**: 1010428317711098.

Tiacci E, Venanzi A, Ascani S, Marra A, Cardinali V, Martino G, Codoni V, Schiavoni G, Martelli MP, Falini B. 2018. High-Risk Clonal Hematopoiesis as the Origin of AITL and NPM1-Mutated AML. *N Engl J Med* **379**: 981-984.

Till JE, McCulloch EA. 1980. Hemopoietic stem cell differentiation. *Biochim Biophys Acta* **605**: 431-459.

Touzart A, Boissel N, Belhocine M, Smith C, Graux C, Latiri M, Lhermitte L, Mathieu EL, Huguet F, Lamant L et al. 2020. Low level CpG island promoter methylation predicts a poor outcome in adult T-cell acute lymphoblastic leukemia. *Haematologica* **105**: 1575-1581.

Van Vlierberghe P, Ambesi-Impiombato A, De Keersmaecker K, Hadler M, Paietta E, Tallman MS, Rowe JM, Forne C, Rue M, Ferrando AA. 2013. Prognostic relevance of integrated genetic profiling in adult T-cell acute lymphoblastic leukemia. *Blood* **122**: 74-82.

Vivas-Garcia Y, Falletta P, Liebing J, Louphrasitthiphol P, Feng Y, Chauhan J, Scott DA, Glodde N, Chocarro-Calvo A, Bonham S et al. 2020. Lineage-Restricted Regulation of SCD and Fatty Acid Saturation by MITF Controls Melanoma Phenotypic Plasticity. *Mol Cell* **77**: 120-137 e129.

Wan G, Liu Y, Zhu J, Guo L, Li C, Yang Y, Gu X, Deng LL, Lu C. 2019. SLFN5 suppresses cancer cell migration and invasion by inhibiting MT1-MMP expression via AKT/GSK-3beta/beta-catenin pathway. *Cell Signal* **59**: 1-12.

Watt F, Molloy PL. 1988. Cytosine methylation prevents binding to DNA of a HeLa cell transcription factor required for optimal expression of the adenovirus major late promoter. *Genes Dev* **2**: 1136-1143.

Weinhold B. 2006. Epigenetics: the science of change. *Environ Health Perspect* **114**: A160-167.

Weng AP, Ferrando AA, Lee W, Morris JPt, Silverman LB, Sanchez-Irizarry C, Blacklow SC, Look AT, Aster JC. 2004. Activating mutations of NOTCH1 in human T cell acute lymphoblastic leukemia. *Science* **306**: 269-271.

Whyte WA, Orlando DA, Hnisz D, Abraham BJ, Lin CY, Kagey MH, Rahl PB, Lee TI, Young RA. 2013. Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* **153**: 307-319.

Wildt KF, Sun G, Grueter B, Fischer M, Zamisch M, Ehlers M, Bosselut R. 2007. The transcription factor Zbtb7b promotes CD4 expression by antagonizing Runx-mediated activation of the CD4 silencer. *J Immunol* **179**: 4405-4414.

Yang H, Wan Z, Huang C, Yin H, Song D. 2019. AMPH-1 is a tumor suppressor of lung cancer by inhibiting Ras-Raf-MEK-ERK signal pathway. *Lasers Med Sci* **34**: 473-478.

Yannoutsos N, Wilson P, Yu W, Chen HT, Nussenzweig A, Petrie H, Nussenzweig MC. 2001. The role of recombination activating gene (RAG) reinduction in thymocyte development in vivo. *J Exp Med* **194**: 471-480.

Yen K, Travins J, Wang F, David MD, Artin E, Straley K, Padyana A, Gross S, DeLaBarre B, Tobin E et al. 2017. AG-221, a First-in-Class Therapy Targeting Acute Myeloid Leukemia Harboring Oncogenic IDH2 Mutations. *Cancer Discov* **7**: 478-493.

You MJ, Medeiros LJ, Hsi ED. 2015. T-lymphoblastic leukemia/lymphoma. *Am J Clin Pathol* **144**: 411-422.

Yu J, Hou M, Pei T. 2020. FAM83A Is a Prognosis Signature and Potential Oncogene of Lung Adenocarcinoma. *DNA Cell Biol* **39**: 890-899.

Zerbino DR, Wilder SP, Johnson N, Juettemann T, Flicek PR. 2015. The ensembl regulatory build. *Genome Biol* **16**: 56.

Zhang J, Lu WY, Zhang JM, Lu RQ, Wu LX, Qin YZ, Liu YR, Lai YY, Jiang H, Jiang Q et al. 2019. S100A16 suppresses the growth and survival of leukaemia cells and correlates with relapse and relapse free survival in adults with Philadelphia chromosome-negative B-cell acute lymphoblastic leukaemia. *Br J Haematol* **185**: 836-851.

Zhang S, Pei X, Hu H, Zhang W, Mo X, Song Q, Zhang Y, Xu K, Wang Y, Na Y. 2016. Functional characterization of the tumor suppressor CMTM8 and its association with prognosis in bladder cancer. *Tumour Biol* **37**: 6217-6225.

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W et al. 2008. Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**: R137.

Zhang YL, Ren JH, Guo XN, Zhang JN, Wang Y, Qiao SK, Lin FR. 2009. [Expression of c-fes gene in leukemia cells and its clinical significance]. *Zhongguo Shi Yan Xue Ye Xue Za Zhi* **17**: 1429-1433.

Zheng S, Shen H, Jia Q, Jing C, Lin J, Zhang M, Zhang X, Zhang B, Liu Y. 2017. S100A6 promotes proliferation of intrahepatic cholangiocarcinoma cells via the activation of the p38/MAPK pathway. *Future Oncol* **13**: 2053-2063.

7 CONFERENCE TALKS AND POSTER PRESENTATIONS

7.1 Conference talks

- Methylation based subtype characterization of adult T-ALL. *European Hematology Association annual meeting 25 (EHA25), Frankfurt. (2020-06).*
- Methylation based subtype characterization of adult T-cell Acute Lymphoblastic Leukemia. *Heidelberg Leukemia Network (HeLeNe) annual meeting, Heidelberg. (2019-11)*

7.2 Poster presentations

- Methrix: An R package for efficient processing of bedGraph files from large-scale methylome cohorts. *European Conference on Computational Biology (ECCB) annual meeting-2019. Basel. (2019-07)*
- Methrix: An R package for efficient processing of bedGraph files from large-scale methylome cohorts. *DKFZ research program, Functional and structural genomics annual scientific retreat-2020. Kloster Schöntal. (2020-02)*

8 PEER REVIEWED PUBLICATIONS

1. Aurore Touzart[*], **Anand Mayakonda***, Charlotte Smith, Joschka Hey, Reka Toth, Agata Cieslak[2], Guillaume P. Andrieu, Christine Tran Quang, Mehdi Latiri, Jacques Ghysdael, Salvatore Spicuglia, Hervé Dombret, Norbert Ifrah, Elizabeth Macintyre, Pavlo Lutsik, Nicolas Boissel, Christoph Plass and Vahid Asnafi. *Epigenetic blueprint identifies poor outcome and hypomethylating agent-responsive T-ALL subgroup. Science Translational Medicine. (In Press; 2021)*

2. **Mayakonda A**, Schönung M, Hey J, Batra RN, Feuerstein-Akgoz C, Köhler K, Lipka DB, Sotillo R, Plass C, Lutsik P, Toth R. *Methrix: an R/bioconductor package for systematic aggregation and analysis of bisulfite sequencing data. Bioinformatics.* 2020 Dec 21:btaa1048. doi: 10.1093/bioinformatics/btaa1048. PMID: 33346800.

3. Lutsik P, Baude A, Mancarella D, Öz S, Kühn A, Toth R, Hey J, Toprak UH, Lim J, Nguyen VH, Jiang C, **Mayakonda A**, Hartmann M, Rosemann F, Breuer K, Vonficht D, Grünschläger F, Lee S, Schuhmacher MK, Kusevic D, Jauch A, Weichenhan D, Zustin J, Schlesner M, Haas S, Park JH, Park YJ, Oppermann U, Jeltsch A, Haller F, Fellenberg J, Lindroth AM, Plass C. *Globally altered epigenetic landscape and delayed osteogenic differentiation in H3.3-G34W-mutant giant cell tumor of bone. Nat Commun.* 2020 Oct 27;11(1):5414. doi: 10.1038/s41467-020-18955-y. PMID: 33110075; PMCID: PMC7591516.

[*] Equal contribution

9 SOFTWARE TOOLS DEVLOPED

- **Methrix**: An R package for fast and flexible DNA methylation analysis. Source code available at: https://github.com/CompEpigen/methrix
- **Trackplot**: An R script to generate IGV style locus tracks from bigWig files. Source code available at: https://github.com/PoisonAlien/trackplot
- **Peakseason**: An R package for rapid rapid bigWig file summarization and visualization. Source code available at: https://github.com/PoisonAlien/peakseason

10 ACKNOWLEDGMENTS

I would like to thank the below people for the direct or indirect support, without which this work would not have been possible. I would like to thank,

- Prof. Christoph Plass for providing me the opportunity to carry out the research
- Dr. Aurore Touzart for the supervision and guidance
- Dr. Pavlo Lutsik, for all the valuable suggestions and discussions
- Thesis advisory committee members Prof. Dr. Andreas Kulozik and Prof. Dr. Benedikt Brors for the helpful feedback
- Dr. Odilia Popanda and Dr. Michael Milsom for being part of the thesis defense committee
- Dr. Yassen Assenov for introducing me to the T-ALL project
- All the members of the B370/C010 division

In particular, I am grateful to,

- Joschka and Reka for being the best friends and colleagues I could have ever asked for. Whether it is luncheons, serious discussions, or mindless conversations, it is always fun with you two
- Oliver (Oli), Yoann , and Yunhee for always being the best company
- Alex, Max, Raj, and Sridhar for getting me through the Pandemic
- Justyna, for being a good friend and introducing me to new friends.
- Sequencing and IT core-facility for maintaining the core infrastructure
- *Stack Overflow* – for always being there whenever the R sessions or Python programs crashed
- Entire *CRAN*, *Bioconductor* core team and contributors for such a fantastic resource
- People who made *Conda* package management system
- People who take their time to write random R/Python/Bash scripts to solve trivial problems
- My one-year-old niece Nishita, whose giggles have kept me going