The role of the non-coding genome in pediatric B-cell precursor acute lymphoblastic leukemia

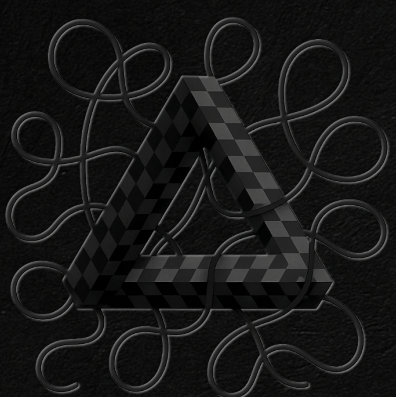
Aydin, Efe

2025

Document Version:
Publisher's PDF, also known as Version of record

The role of the non-coding genome in pediatric B-cell precursor acute lymphoblastic leukemia

EFE AYDIN
DEPARTMENT OF LABORATORY MEDICINE | FACULTY OF MEDICINE | LUND UNIVERSITY



The role of the non-coding genome in pediatric B-cell precursor acute lymphoblastic leukemia

Efe Aydın

LUND
UNIVERSITY

DOCTORAL DISSERTATION

Doctoral dissertation for the degree of Doctor of Philosophy (PhD) at the Faculty of Medicine at Lund University to be publicly defended on Thursday 27th of November 2025 at 13:00 in Lundmarksalen, Astronomicentrum, Sölvegatan 27, Lund.

Faculty opponent
Professor Thomas Milne

University of Oxford, UK

Organization: LUND UNIVERSITY

Document name: Doctoral Dissertation **Date of issue: 27 November 2025**

Author: Efe Aydın **Sponsoring organization:**

Title and subtitle: The role of the non-coding genome in pediatric B-cell precursor acute lymphoblastic leukemia

Abstract:

Acute lymphoblastic leukemia (ALL) is the most common pediatric malignancy. Decades of effort transformed this once incurable disease into one of the greatest success stories in cancer, with overall survival rates now exceeding 90% in developed countries. However, reports from recent trials indicate that conventional therapy has reached its tolerance limit, leaving little room for further intensification. At the same time, chemotherapy-related toxicities motivate ongoing de-intensification efforts for lower-risk groups. Thus, two major challenges dominate ALL research: improving outcomes in high-risk patients and minimizing adverse effects in low-risk patients. Addressing both requires a deeper understanding of leukemia biology.

Non-coding regions constitute ~98% of the human genome, yet they remain far less explored than protein-coding genes. Once dismissed as "junk", they are now recognized—through advances in high-throughput technologies—as essential regulators of spatiotemporal gene expression. In this thesis, we investigated the relevance of non-coding regions in pediatric B-cell precursor (BCP) ALL from a 3D genome perspective. To delineate the cis-regulatory landscapes, we integrated multi-omics data from more than 1,000 primary specimens.

In **Paper I,** we showed that non-coding mutation hotspots can highlight cis-regulatory mechanisms when integrated with epigenomic and transcriptomic data, exemplified by subtype-specific hotspot enrichment near COSMIC genes and the identification of an *NRAS*-associated repressive element. In **Paper II,** we developed BASE, a computational tool that integrates allele-specific copy number states with RNA-seq to resolve allele-specific expression in aneuploid tumors, and applied it to high hyperdiploid ALL to demonstrate the extent of copy-number driven dosage effects alongside additional regulatory mechanisms. In **Paper III**, we generated the largest primary-sample Micro-C dataset in BCP ALL, mapping 3D genome organization at kilobase resolution and demonstrating the impact of somatic events on compartments, topologically associating domains, and regulatory element positioning. To support further research, we developed a web application, enabling access to >25,000 high-resolution loops with regulatory potential. In **Paper IV**, we combined our 3D genome map with whole genome sequencing data from 801 cases to study the interplay between mutagenesis and chromatin organization. We demonstrate that each layer of chromatin organization is tightly associated with both mutagenesis rates and signatures in BCP ALL.

Altogether, our work provides important insights into the role of non-coding regions in BCP ALL and establishes resources that will facilitate future research into leukemia biology.

Key words: acute lymphoblastic leukemia, chromatin organization, non-coding mutations, gene regulation, 3D genome, cis-regulatory elements

Classification system and/or index terms (if any) Supplementary bibliographical information

Language: English **Number of pages**: 118

ISSN and key title: 1652-8220 **ISBN:** 978-91-8021-783-5

Recipient's notes Price Security classification

Signature Date 2025-10-16

The role of the non-coding genome in pediatric B-cell precursor acute lymphoblastic leukemia

Efe Aydın

LUND
UNIVERSITY

Faculty of Medicine

Department of Laboratory Medicine

Division of Clinical Genetics

Lund University

ISBN 978-91-8021-783-5

ISSN 1652-8220

Printed in Sweden by Media-Tryck, Lund University,
Lund, 2025

To Hannah, Berrin and Levent

Table of contents

Original articles

Paper I

Aydın, E., Woodward, E. L., Dushime, G. T., Gunnarsson, R., Lilljebjörn, H., Moura-Castro, L. H., Fioretos, T., Johansson, B., Paulsson, K., & Yang, M. (2025). Discovery of cis-regulatory mechanisms via non-coding mutations in acute lymphoblastic leukemia. *Genes, Chromosomes and Cancer*, **64**(3), e70045.

Paper II

Andersson, J., **Aydın, E.,** Gunnarsson, R., Lilljebjörn, H., Fioretos, T., Johansson, B., Paulsson, K., & Yang, M. (2024). Characterizing the allele-specific gene expression landscape in high hyperdiploid acute lymphoblastic leukemia with BASE. *Scientific Reports*, **14**(1), 23181.

Paper III

Moura-Castro, L. H.*, **Aydın, E.***, Telliam Dushime, G., Woodward, E. L., Castor, A., Gunnarsson, R., Lilljebjörn, H., Olsson-Arvidsson, L., Fioretos, T., Johansson, B., Järås, M., Hagström-Andersson, A. K., Yang, M., & Paulsson, K. (2025). The 3D genome of pediatric B-cell precursor acute lymphoblastic leukemia. *Submitted*. Available as pre-print: https://doi.org/10.1101/2025.07.22.666073

*Shared first authors.

Paper IV

Aydın, E., Moura-Castro, L. H., Gunnarsson, R., Johansson, B., Yang, M., & Paulsson, K. (2025). Interplay between somatic mutagenesis and the 3D genome in pediatric B-cell precursor acute lymphoblastic leukemia. *Manuscript.*

Abbreviations

ALL	Acute lymphoblastic leukemia
AML	Acute myeloid leukemia
Ara-C	Cytarabine
asCN	Allele-specific copy number
ASE	Allele-specific expression
ASP	Asparaginase
ATAC-seq	Assay for transposase-accessible chromatin with high-throughput sequencing
BCP	B-cell precursor
BRE	TFIIB recognition element
CAR	Chimeric antigen receptor
CAGE	Cap analysis of gene expression
CCLE	Cancer Cell Line Encyclopedia
CFU-S	Spleen colony forming unit assay
ChEC-seq	Chromatin endogenous cleavage sequencing
ChIP-seq	Chromatin immunoprecipitation sequencing
CLP	Common lymphoid progenitor
CML	Chronic myeloid leukemia
CMP	Common myeloid progenitor
CNV	Copy number variation
CPE	Canonical promoter element
CRE	Cis-regulatory element
CRISPR	Clustered regularly interspaced short palindromic repeats
CTCF	CCCTC binding factor
DAUN	Daunorubicin
DCE	Downstream core element
DEX	Dexamethasone
DPE	Downstream promoter element
eRNA	Enhancer RNA
ESCs	Embryonic stem cells
E-P	Enhancer-Promoter
FACS	Fluorescence-activated cell sorting
FISH	Fluorescence in situ hybridization

GAM	Genome architecture mapping
GWAS	Genome-wide association study
HAT	Histone acetyltransferase
HDAC	Histone deacetylase
HCTs	Homotypic cluster of transcription factor binding sites
HDM	Histone demethylase
HeH	High-hyperdiploid
HMT	Histone methyltransferase
HR	High risk
HSCs	Hematopoietic stem cells
HSPCs	Hematopoietic stem and progenitor cells
iAMP21	Intrachromosomal amplification of chromosome 21
ICC	International consensus classification
Inr	Initiator
KMT2A-r	*KMT2A*-rearranged
LSCs	Leukemic stem cells
MNase	Micrococcal nuclease
MPP	Multipotent progenitor
MPRA	Massively parallel reporter assay
MRD	Minimal residual disease
MTE	Motif ten element
MTX	Methotrexate
NCI	National Cancer Institute
ncRNA	Non-coding RNA
NGS	Next-generation sequencing
NOS	Not otherwise specified
PBMCs	Peripheral blood mononuclear cells
PCA	Principal component analysis
PCR	Polymerase chain reaction
PRC2	Polycomb repressive complex 2
PRED	Prednisone
PTM	Post-translational modification
PWM	Positional weight matrix
qPCR	Quantitative polymerase chain reaction
RNA-seq	RNA sequencing
RNP	Ribonucleoprotein
SBS	Single base substitution
sgRNA	Single guide RNA
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variant
SPRITE	Split-pool recognition of interactions by tag extension

SR	Standard risk
SVs	Structural variants
SV40	Simian vacuolating virus 40
TAD	Topologically associating domain
TKIs	Tyrosine kinase inhibitors
TSS	Transcription start site
VAF	Variant allele frequency
VCR	Vincristine
WBC	White blood cell
WGS	Whole genome sequencing
XCPE1	X-core promoter element 1
XCPE2	X-core promoter element 2
6-MP	Mercaptopurine

1. Introduction

Every day, approximately 1,000 children receive a cancer diagnosis[1]. Cancer presents with a wide range of types and subgroups, as well as with distinct prognoses and treatment options. The main challenge in finding a common cure lies in the heterogeneity and complexity of the disease. Nevertheless, some shared patterns exist. Cancer cells arise and persist by following a set of organizational principles that grant them evolutionary advantages. These concepts are essential to understand for any cancer researcher, but I did not dedicate a specific section to them in this book, since they are thoroughly discussed in the well-known "Hallmarks of Cancer" review series by Douglas Hanahan and Robert Weinberg[2–4]. Moreover, unifying patterns in cancer are not limited to oncogenesis and some of them have more weight on this work than others. Instead, I would like to highlight three axioms here at the beginning, as they form the foundation of the questions explored throughout the thesis.

Axiom 1: Overall survival rates for cancer drops with metastasis and/or relapse.

Axiom 2: Cancer survivors live with increased long-term risks associated with therapy.

Axiom 3: All cancers ultimately arise from somatic evolution.

While these seem to represent generic knowledge derived from decades of cancer research, each holds unique significance when evaluated through the lens of pediatric oncology.

Starting with Axiom #1, "Overall Survival" and "Event-Free Survival" are standard metrics in medical data analysis, typically measured in time units such as days or years. According to the National Cancer Institute (NCI), the median age of a cancer diagnosis is 67, while life expectancy in the United States is 77.5 years. Needless to say, life is valuable at every age, and the collective human effort to extend our lifespan has been one of civilization's great achievements. Hopefully, through continued global collaboration, we will keep pushing this number upward and help future generations live longer and healthier lives in a sustainable world. Nevertheless, it is clear that the meaning of survival differs greatly between a child and an elderly patient. For example, consider two cancers—Cancer A and B, with a median age of diagnosis at 70 and 5, respectively. If the average event-free survival is five years for both of them, then doubling this rate in Cancer A would have a major impact for the average patient. However, for Cancer B, the same improvement would be far away from restoring a

normal life trajectory. Thus, understanding the biological basis of relapse or metastasis is of paramount importance in childhood cancers.

In the second axiom, our concept of interest is chemotherapy. The term was first used by the prominent German scientist Paul Ehrlich in the early 1900s, while he was investigating the potential of chemicals as a cure against infectious diseases. Naturally, this idea caught up with cancer researchers as well, and the 20th century became a roller coaster for research in finding drugs to cure cancer, with hopes dying and rekindling periodically, almost as orderly as a sine function. Some of the bright moments include the development of the rodent model in 1910s, use of anti-folate compounds in children with acute lymphoblastic leukemia (ALL) and achieving complete remission in 1948, the rise of combination therapies in 1970s, the publication of the human genome in 2001 and the consequent developments on targeted therapies which are ongoing until now. For those who are interested in learning more about this 100+ years old history of drug development in cancer, there is a very nice review published on AACR Centennial Series, which can be found in the references section[5].

Through decades of relentless effort, generations of scientists have built upon each other's discoveries, achieving transformative progress in many cancers—for instance, in ALL, where survival rates have climbed from ~10% to over 90%[6]. However, when it comes to the fight against cancer, even the best-case scenario is not a walk in the park. Cytotoxicity has been a major drawback and a research priority since the early beginnings of drug development. To date, no compound has achieved the perfect balance of sensitivity and specificity—one that effectively eliminates all cancer cells while sparing everything else. Specificity and side-effects exist in a delicate balance, often functioning in inverse proportion within a complex therapeutic equation. This trade-off becomes especially significant in a growing patient, where active developmental programs present a broad range of additional targets vulnerable to collateral damage[7]. It should be noted that considerable progress has been achieved in this field within the last two decades, particularly due to the advance of targeted therapies and improved risk stratification. However, only a small proportion of children benefit from these advances, and more research is needed to identify new molecular targets and biomarkers. As of today, the average childhood cancer survivor lives with an increased risk of numerous long-term complications, including cardiac, pulmonary, hormonal, reproductive, and psychological issues[8].

The third axiom warrants special attention, as it encompasses a large set of ontological elements interacting within an intricate network. Here, it makes sense to follow a reductive approach and examine this complex information through its elemental building blocks. First, we must consider the framework in which everything occurs: the human body. This framework of interest is an open system composed of multiple hierarchical layers, from molecules to cells, cells to tissues, tissues to organs, and organs to systems. On many occasions, the whole exceeds the sum of its parts.

Anyone with a slight notion of interest in biology would be awed by the remarkable complexity present at each of these layers, particularly when exhibited in a higher-order mammal like us. However, I believe one of them, the cell, should have a specific emphasis here, before diving further into the underlying implications of the complex network. The cell, after all, is the smallest unit that can meaningfully be called "cancer". Oncogenes, carcinogenic chemicals, fusion proteins, radiation, viruses and so on may drive malignancy, but none would have function or consequence outside the cellular context. You can insert all known driver mutations into DNA, extract it, and store it in an Eppendorf tube, yet without a living cell, it is nothing more than inert organic matter. In this sense, the cell is kind of similar to the least common multiple in math, as it is the smallest common unit in which all fundamental components—DNA, RNA, proteins, hormones, enzymes—gain functional meaning. Cells proliferate, differentiate, communicate, and die, like an orchestra performing a well-composed symphony operating across multiple levels. This orchestra, however, is significantly more crowded than we would expect and it is scheduled to perform ten quadrillion concerts.

Let's imagine ourselves as the conductor of this ensemble. I think we can all agree that it is not fair to expect every musician to deliver a flawless performance each time. Everyone has off days, one can miss a note, play slightly out of tune, fall out of sync for a moment. In most cases, these small errors would be instantly corrected by the musician, after all, they know what they are doing. Chances are, no one in the audience would even notice these small slips from one member of such a vast orchestra. But what if the errors persist? What if a section of violinists can no longer perform their part correctly? Perhaps they just need to rehearse a little more. We give them feedback, allow them extra-time for rehearsal, and hope they're ready for the next performance. Often, they are. But what if the dissonance does not fade? What if these musicians do not resonate with the orchestra anymore and just play something of their own making? They are simply not themselves anymore. Unfortunately, in these cases, we may have to let them go, for the sake of music.

Readers within the field of molecular biology have likely already grasped the metaphor. The conductor here represents cell cycle regulators and ten quadrillion is an approximation on the number of cell divisions occurring in a human life span[9,10]. Occasionally, these cells make mistakes while replicating their DNA. This is not necessarily a bad thing—after all, without such plasticity, evolution wouldn't exist, and neither would we. Moreover, most of these errors are corrected by a suite of repair mechanisms, equipped with a range of faculties such as halting the cell cycle or inducing apoptosis[11]. Those that slip past the repair machinery do not necessarily pose harm either, as most variations have no impact on the phenotype. Even mutations affecting critical regulators—such as those involved in cell cycle control or apoptotic signaling—are not inherently catastrophic either. Evolution has equipped us with multiple redundant checkpoints to keep the system stable. So, there is not a single conductor in

this orchestra. The human body is not a dictatorship, it is more like a well-established state with very strong checks and balances; it is not easy to take over.

Unfortunately, every once in a while, the stars align for cancer. On some occasions, this is achieved through some external help; a mutagenic compound inhaled from a cigarette, an oncovirus, radiation, or even a chronically poor diet. Sometimes the help comes from within, and internal vulnerabilities play the decisive role. Our germline legacy, which is passed down to us through generations, can quietly predispose us to collapse. After the first malignant blow occurs in a healthy cell, an inherited architecture may immediately become an accomplice, paving the way for cancer to grow. Yet regardless of whether the catalyst comes from our ancestors, our habits, or our environment, one truth remains: the somatic event must occur[12]. Without it, cancer cannot emerge. In Aristotelian terms, it is the efficient cause—the triggering force without which the chain of tumorigenesis cannot begin. At this point, one might wonder why axiom 3 has received such emphasis, and what relevance it holds for the work presented in this thesis. To answer these questions, we need to put back our lens of pediatric oncology.

Most pediatric cancers are characterized by a relatively small mutational burden (**Figure 1**)[13]. This is not counterintuitive, as several mutagenesis patterns are associated with age driven factors such as reactive oxygen species[14]. In fact, mutation accumulation represents one of the central theories in aging research[15], but this is not within our scope. In our context, low mutational burden is a double-edged sword, bringing both challenges and opportunities. The challenge lies in the scarcity of canonical driver mutations. Many children with cancer are ineligible for newly developed targeted therapies simply because their genomes lack the alterations those treatments are designed to exploit[7]. On the other hand, this limitation offers a wide range of opportunities for those who are up for the challenge, as it forces us to shift our perspective. It makes us ask the question "What is causing this dysfunction, if it is not something within the protein coding gene?". To answer this, we need to zoom out from the protein-coding genome to the entire genome, from the one-dimensional sequence of bases to the three-dimensional architecture of chromatin that orchestrates gene regulation. We have to recalibrate our lenses to see beyond Crick's central dogma[16], to see the complex set of systems that is not limited to genetics but also to what is above it, or in plain words, epigenetics. In this way, the biological patterns observed in pediatric cancers lead researchers to the road less traveled, one marked by unknown challenges and transformative potential. Importantly, the insights gained along this path may not only open new avenues for treating these specific diseases but also deepen our understanding of cellular systems more broadly, offering knowledge that extends far beyond pediatric oncology. The work presented here is a product of this path, although I do not claim a paradigm shift on the reader's head. All the scientific output generated here was built on one modest goal: to transfer a small drop from the ocean

of the unknown into the lake of the known, within the context of 3D genome and our disease of interest, pediatric B-cell precursor (BCP) ALL.

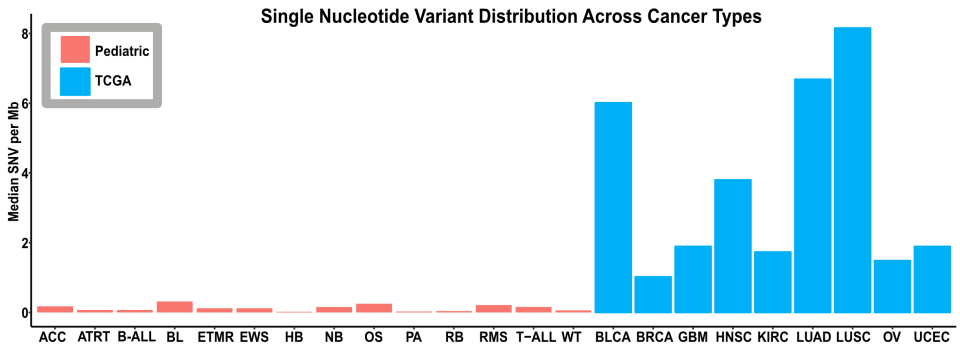


Figure 1 | Barplot showing the single nucleotide variant (SNV) distribution across cancer types. The x-axis represents the following cancer types: Adrenocortical carcinoma (ACC), Atypical teratoid rhabdoid tumor (ATRT), B-cell precursor acute lymphoblastic leukemia (B-ALL), Burkitt's Lymphoma (BL), Embryonal tumor with multilayered rosettes (ETMR), Ewing Sarcoma (EWS), Hepatoblastoma (HP), Neuroblastoma (NB), Osteosarcoma (OS), Pilocytic astrocytoma (PA), Retinoblastoma (RB), Rhabdomyosarcoma (RMS), T-cell ALL (T-ALL), Wilms' Tumor (WT), Bladder Urothelial Carcinoma (BLCA), Breast invasive carcinoma (BRCA), Glioblastoma multiforme (GBM), Head and Neck squamous cell carcinoma (HNSC), Kidney renal clear cell carcinoma (KIRC), Lung adenocarcinoma (LUAD), Lung squamous cell carcinoma (LUSC), Ovarian serous cystadenocarcinoma (OV) and Uterine Corpus Endometrial Carcinoma (UCEC). The y-axis represents the median SNV count per megabase for each of these cancers. Bars are color coded based on the origin of datasets. Red bars represent pediatric cancers. Blue bars represent adult cancers taken from The Cancer Genome Atlas (TCGA). Median values were calculated for the given cancers using the data generated from previously published studies[17,18].

At this point, the three foundational axioms of this thesis are fully introduced. They were first examined through the broad lens of cancer, then through the perspective of pediatric oncology. I highlighted how each of these axioms gains specificity and relevance in this transition. Now, it is time to put on an even higher-resolution lens, one that brings our disease of interest, BCP ALL, into clear focus.

Leukemia is the most common type of cancer in children, accounting for approximately 34% of all pediatric cases[19]. Of these, about 75–80% are ALL, and roughly 80% of ALL cases arise from the B-cell lineage[19,20]. Thus, BCP ALL is, by a wide margin, the most prevalent childhood malignancy (**Figure 2**). Naturally, this high incidence brings both advantages and challenges. On the positive side, BCP ALL and other leukemias have received nearly a century of concentrated financial, institutional, and scientific investment, a level of attention rarely afforded to rarer diseases. As a result, leukemia research has often led the way in cancer biology, serving as the foundation for many major advances in the field. Many breakthroughs in cancer research, such as the identification of first cancer driver translocation (*BCR::ABL1* aka the Philadelphia chromosome)[21], isolation of the first cancer stem cell[22], achieving complete remissions with chimeric antigen receptor (CAR) T cell therapy[23], minimal residual disease

(MRD) based risk stratification[24], and many others were products of leukemia-focused studies. Importantly, the value of scientific discoveries usually extends beyond the discoveries themselves. Each major advance leaves behind a legacy of well-established tools—optimized protocols, validated animal models, and immortalized cell lines—that streamline future discoveries. Yet, as with all mature research fields, certain limitations emerge. After nearly a century of investigation, most of the 'usual suspects' have been extensively studied. It has been more than two decades since the human genome was first sequenced, and the majority of protein-coding regions has since been systematically explored. Only in this year, there are more than 12,000 publications with the keyword leukemia on PubMed. Yet even in light of this enormous body of work, the three axioms remain central, as evidence shows that; 1) Event-free survival rates drop from over >90% to <50% in refractory ALL[25] 2) Chemotherapy remains the standard frontline treatment in children diagnosed with ALL, exposing them to long-term risks[26]. 3) Although certain germline variations were found to increase susceptibility to disease, malignant transformation in ALL occurs through somatic alterations[27].

At this very moment, thousands of scientists around the world are working in their laboratories (and computers) to improve our understanding and treatment of leukemia in ways that relate to the axioms discussed earlier. We no longer live in the age of polymaths; today, deep understanding is achieved through high-level specialization, followed by collaboration between specialists across disciplines.

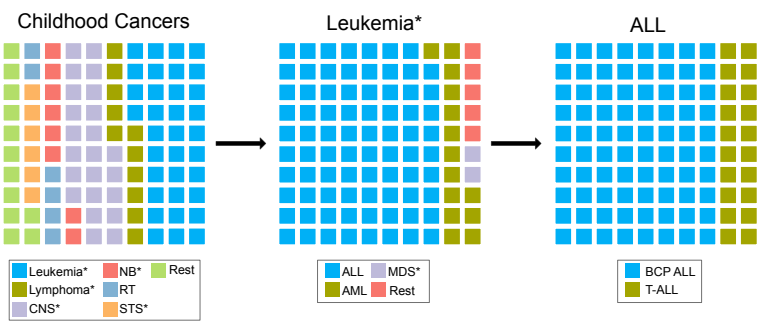


Figure 2 | Waffle charts showing the distribution of childhood cancer types. The first waffle chart represents a general distribution of all childhood cancers. The frequencies and abbreviations are as followed; Leukemias, myeloproliferative and myelodysplastic diseases (34.1%, abbreviated as Leukemia*), Lymphomas and reticuloendothelial neoplasms (11.5%, abbreviated as Lymphoma*), Central nervous system (CNS) and miscellaneous intracranial and intraspinal neoplasms (22.5%, abbreviated as CNS*), Neuroblastoma and other peripheral nervous cell tumors (7.6%, abbreviated as NB*), Renal Tumors (5.6%, abbreviated as RT), Soft tissue and other extraosseous sarcomas (6.1%, abbreviated as STS*). The remaining cancer types (Rest) account for approximately 12.5% of the cases). The second chart represents the subgroups under Leukemia*. These include ALL (78.6%), AML (13.7%), Myelodysplastic syndrome and other myeloproliferative diseases (2.3%, abbreviated as MDS*). The remaining Leukemia* subgroups account for approximately 5.4% of the cases. The final chart represents the distribution of ALL. BCP ALL takes 80% of the cases while T-ALL accounts for 20%. Color codes are represented in legend bars under each chart. Numbers are taken from previously published studies[19,20].

In brief words, the aim of this thesis is to provide new biological insights into regulatory systems that govern BCP ALL from an understudied perspective that could be useful to each of these specialists, whether their focus lies on therapeutics, risk-stratification or fundamental biology. This multi-layered investigation of BCP ALL gene regulation comprises four individual studies. While each explores a distinct question, they are all interconnected by a shared systems-level perspective.

To avoid research bias, embrace the strengths of data-driven science, and effectively integrate existing knowledge of cell biology, we followed a consistent set of methodological principles throughout each study. These guiding principles are key to understanding the structure and rationale of this work. For ease of communication—and perhaps to add a touch of humor—I have chosen to summarize these principles using the framework of diversity, equity, and inclusion, a hot debate topic in politics repurposed as a scientific metaphor.

Diversity: BCP ALL is a highly heterogeneous disease with numerous genetic subgroups[28]. To minimize the risk of bias stemming from inter- or intra-sample variability, we included primary sample data from multiple cohorts, covering diverse backgrounds and all major genetic subtypes throughout the thesis. In total, experimental data generated from more than 1000 primary samples were included in this work.

Equity: Not all regions of the human genome have received equal scientific attention. In the context of BCP ALL, many studies have focused on known protein-coding drivers such as *NRAS*, *KRAS*, *FLT3*, and others, aiming to optimize their clinical utility. Some researchers have extended this search to the full set of protein-coding genes, looking for novel drivers that may have been previously overlooked.

However, protein-coding regions make up less than 2% of the human genome[29]. The remaining 98%, though long dismissed as "junk DNA", is now known to harbor important regulatory elements[30], largely due to advances in multi-omics technologies. Researchers from the ENCODE consortium estimates functional potential for 80% of the genome[31]. Since the seminal 2013 studies demonstrated the driver role of *TERT* promoter mutations in cancer[32,33], there has been a growing interest about their potential roles in carcinogenesis as well. Nevertheless, they are still significantly understudied compared to their protein coding counterparts.

While non-coding regions do not directly encode proteins, many are involved in regulating gene expression through complex and indirect mechanisms that operate in three-dimensional chromatin space. Unlike major coding-region drivers such as *RAS* mutations, the functional impact of non-coding variants can be more subtle, typically exerting fine-tuning effects that are harder to detect with conventional approaches. Moreover, as research into the non-coding genome is still in its early stages, the available

analytical pipelines remain limited and less well established, making systematic investigation particularly challenging.

Using both existing tools and our own code, we have attempted to address this historical imbalance by systematically interrogating the non-coding genome of BCP ALL. All findings presented in this thesis are products of genome-wide analyses designed to capture the behavior of the system as a whole. By developing and applying these pipelines, we aim not only to demonstrate the clinical relevance of the non-coding genome in BCP ALL, but also to encourage further exploration in this field full of potential.

Inclusion: The influential 20th-century psychologist Abraham Maslow, in The Psychology of Science (1966), highlighted the human tendency to treat every problem as a nail when one's only tool is a hammer[34]. This cautionary insight remains highly relevant to contemporary scientific practice, where researchers may rely too heavily on a single perspective or method they are most familiar with. To avoid this pitfall, we deliberately chose not to interpret our findings through the lens of any single technique. Instead, throughout the thesis, we integrated outputs from multiple, complementary approaches.

We combined primary sample data—both generated in our lab and drawn from published sources—across several high-throughput platforms including but not limited to; whole-genome sequencing (WGS), RNA sequencing (RNA-seq), chromatin immunoprecipitation sequencing (ChIP-seq), assay for transposase-accessible chromatin (ATAC-seq), and Micro-C. By incorporating diverse layers of genetic and epigenetic information, we aimed to generate a more comprehensive and inclusive understanding of gene regulatory programs in BCP ALL.

Altogether, we aimed to bring forward a new aspect regarding the regulatory architecture of BCP ALL on each of the four studies presented in the thesis.

Paper I represents a systematic evaluation of non-coding mutations in BCP ALL. This study focused on the genomic windows where mutations were recurrently observed in multiple samples, also known as mutational hotspots. By integrating WGS, RNA-seq, ChIP-seq, ATAC-seq and Micro-C data, we assessed the potential regulatory impact of these hotspots on nearby genes. Consistent with previous studies, we observed a relative depletion of driver mutations compared to those found in coding regions. On the other hand, we found that hotspots co-localizing with active histone marks were associated with differential expression of neighboring genes. Additionally, subtype-specific hotspots were enriched for proximity to genes listed in the COSMIC database. We further characterized a hotspot associated with elevated *NRAS* expression and demonstrated the presence of a repressor element in the corresponding region via dual-luciferase reporter assays and Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR)-based deletion experiments. Altogether, this study highlights the

value of integrating multiple data layers to uncover regulatory signals hidden within the non-coding genome, signals that are not always apparent at first glance.

In **Paper II**, we developed a software tool that integrates WGS and RNA-seq data to quantify the contribution of copy number variation (CNV) in the context of allele-specific expression (ASE). Both CNV and ASE are independently important phenomena in BCP ALL and cancer research in general. Our tool provides researchers a framework to connect these two layers of data and assess the influence of CNVs on ASE across any malignancy of interest. As a proof of concept, we demonstrated an application on high hyperdiploid (HeH; 51-67 chromosomes) ALL, the most prevalent subtype of the disease.

In **Paper III**, we present – to our knowledge – the most comprehensive primary sample driven 3D chromatin map of the BCP ALL genome to date. By optimizing the implementation of Micro-C, a new variant of Hi-C, we generated chromatin interaction maps reaching resolutions up to 1 kb. Through the integration of paired RNA-seq data, we investigated the role of chromatin architecture in BCP ALL gene regulation across multiple scales, including A/B compartments, topologically associating domains (TADs), and cis-regulatory loops. The data generated in this study provide unprecedented insights into BCP ALL gene regulation, such as how alterations in loop strength can drive differential transcriptional programs in the absence of somatic mutations, how molecular subtypes cluster according to chromatin organizational principles, and how alternative promoter interactions are enriched for leukemia associated genes, along with many other findings detailed in the "Present study" section of this book. To facilitate further research in the field, we also developed a web application that enables users to access our loop data through a user-friendly interface, where they can perform gene or location-based queries to receive targeted information on the cis-regulatory architecture of BCP ALL.

Previous studies on cell lines demonstrated that the large-scale chromatin features such as compartmentalization were associated with different mutational rates and mechanisms[35,36]. In **Paper IV**, we built upon this foundation and our own findings from the two preceding studies. We examined the mutational landscape of over 800 BCP ALL samples through the lens of 3D chromatin architecture using the high-resolution chromatin interaction map generated in our lab. Our results reveal that mutation rates and mechanisms are intricately connected to multiple layers of chromatin organization of BCP ALL at all levels, including compartmentalization, TADs and fine-scale chromatin looping. Furthermore, we demonstrate that the mutational landscape undergoes marked changes from a 3D perspective between diagnosis and relapse, revealing a clinically relevant dimension of mutation biology that has previously been underexplored. We showed that the likelihood of mutagenesis within chromatin loops was significantly lower than in the rest of the genome. This association was more pronounced when the anchor overlapped a cis-regulatory element

(CRE), yet even a purely structural anchor was significantly less likely to mutate than a random genomic locus, suggesting that both evolutionary and biochemical constraints are at play. Interestingly, we noticed that the frequency of mutations occurring within loops increases significantly at relapse. Altogether, this study provides a mechanistic explanation for the consistently observed depletion of driver mutations in the non-coding genome and underscores the importance of incorporating spatial chromatin context to uncover clinically relevant regulatory alterations.

In the "Present study" section, the results from all four papers and their relevance to BCP ALL research are discussed in detail. The full manuscripts can be found in the appendix. However, before delving into these studies, it is important to provide background information to help the reader better understand the context and significance of this work. After all, this research did not begin from scratch, and the knowledge we built upon was not received a priori. Without the accumulated experimental outputs generated by our predecessors, none of these projects would have been realized.

Thereby, I will now move forward with a "Background" section to revisit these essential concepts before diving further into the details of the doctoral work. I divided this section into two main parts: the first covers the foundations of fundamental biology underlying our analyses, and the second focuses specifically on BCP ALL. In both subsections, I trace the field from its historical origins to its current state.

This is followed by the "Present study" section, which represents the scientific output of my doctoral work and discusses its contributions to the field. The "Methods" subsection within the "Present study" chapter outlines the experimental and computational techniques used throughout the work. Finally, I conclude with a discussion of future perspectives and open questions in the field.

2. Background

2.1. Molecular biology

2.1.1. Evolution of gene as a concept

According to Thomas Kuhn's theory of scientific progress, science progresses through paradigm shifts, where the revolutionary outputs generate such new avenues to upcoming scientists that were unimaginable before[37]. When we look back at the history of genetics, we can clearly identify some of these pivotal moments, which are worth highlighting here given their continued relevance. For many geneticists, the first paradigm shift began with Gregor Mendel. His meticulous work on the principles of heredity, using pea plants as model organisms, laid the foundations of the field and earned him the title Father of Genetics, rightfully so. Unfortunately, Mendel never witnessed the impact of his work during his lifetime. As Nietzsche suggests, some are born posthumously, and that was certainly the case for Gregor Mendel.

Rediscovery of his work in early 20th century, created a renaissance-like phenomenon in biological sciences. The concepts of 'genetics' and 'gene' were first used in this period, by William Bateson in 1905 and Wilhelm Johannsen in 1909 respectively[38]. A work entitled "The Mechanism of Mendelian Heredity", published by Thomas Hunt Morgan and colleagues in 1915, discussed the effects of mutations on transmitting traits within *Drosophila melanogaster*, using a linear gene-model[39]. During these years, the term 'gene' was an abstract concept used to define this unseen entity which causes the transmission of traits between generations. It was not until 1929 that scientists learnt about the specific physical locations of these trait-carrying entities on the chromosomes. This was achieved through the cytogenetic experiments performed on triploid maize by Barbara McClintock[40].

In the following decades, the spotlight of biological research was shifted to bringing the abstract concept of a gene to a solid foundation by explaining the molecular features behind. This was achieved in 1953 by James Watson and Francis Crick, when the famous duo discovered the double helix structure of DNA, based on the X-ray diffraction data generated by Rosalind Franklin and Maurice Wilkins[41,42]. In the following decade, the genetic code and the flow of information through DNA, RNA and proteins were delineated by the works of Nirenberg[43], Khorana[44], Holley[45] and

colleagues, which brought them the Nobel prize in physiology in 1968. These developments brought the abstract concept of gene towards a more solid definition. By this time, a general definition of gene was that it was a code within a nucleic acid that is used to create a functional product.

Another breakthrough happened in 1970s, with the development of sequencing technologies, where many firsts such as the 'first fully sequenced gene' or 'first fully sequenced organism' occurred (the term organism is disputable, since it was a virus)[46,47]. However, the biggest breakthrough that is most relevant to the scope of this book occurred after the completion of the human genome project and the follow up studies performed by the ENCODE consortium[29,31]. Before the arrival of next-generation sequencing (NGS) technologies in the 21st century, scientists were already struck by some of the complexities of the human genome, such as the existence of overlapping genes, splicing events, and long stretches of seemingly nonfunctional DNA often considered "junk". However, these surprises of the genome represented only the tip of the iceberg, compared to some of the unexpected features uncovered in this century.

The first immediate revelation was about the number of protein coding genes. These were not significantly higher in humans when compared to some other organisms that were supposedly 'inferior' than us. Moreover, less than 1.5% of DNA bases were coding for exons, but interestingly, the remaining junk were largely conserved across multiple organisms[29,48,49]. These findings compelled scientists to reconsider the very concept of function, and by extension, to rethink the definition of a gene.

Alongside this shift, the complex world of the regulome came into focus. Unlike the relatively simple regulatory model of the lac operon in *Escherichia coli*[50], this new picture revealed a far more intricate and dispersed regulatory architecture. Regulatory elements were found within the entire genome, with notable local enrichments and depletions[31]. Some of them are transcribed, while some carry out functions without transcription. They can reside both upstream or downstream of their target genes, some span distances of more than 100 kilobases, although they are nonetheless brought in close proximity within a 3D spatial context. They are not considered genes themselves, as they do not directly encode for a functional product. However, it is not possible to understand genes without incorporating the regulatory aspect anymore. Thereby, it is essential to provide a technical focus on each of these elements as they are highly relevant for this work.

In a thoughtful perspective published by Gerstein et al.[38] in 2007, the authors discuss the conceptual shifts in the definition of a gene throughout history and propose updated terminology better suited to contemporary understanding. Their work served as an important reference point for this section, and I encourage readers interested in a more detailed historical walkthrough to consult their article.

2.1.2. Key elements of gene regulation

As of now, it has become clear that the flow of information depicted in the central dogma does not cover the entire story behind the genotype to phenotype transition. In the previous section, I discussed the surprisingly small fraction of the human genome occupied by exons, along with the conservation of long stretches of seemingly nonfunctional DNA across higher organisms, all pointing to alternative functional mechanisms at work. Moreover, our bodies contain hundreds of different cell types, carrying out distinct functions while possessing the same DNA. This complexity is achieved through the diverse regulatory mechanisms we have acquired throughout our evolutionary history.

Regulome is a term that represents all the regulatory elements within a cell. These elements have been known to us for more than 80 years, but we have only recently acquired the ability to study them on a systematic, genome-wide scale, due to the developments in high-throughput technologies.

Before diving into the regulome, let us reconsider the central dogma and the genetic code one more time. At its core, the blueprint of genotype information resides in DNA. DNA is responsible for its replication to carry out the information to the daughter cells. The genetic information essential for the cell is transcribed into the mRNA molecule and mRNA is then used as a template to synthesize the specific protein. This simplified framework has been known for many decades, and remains valid for many cases, although far from complete. Most importantly for us, it does not explain why the same DNA can produce different outputs under different cellular contexts. In this chapter, I integrate the updated knowledge of the regulome into these canonical steps. This expanded view significantly increases the complexity of the model, as the information flow becomes multi-directional, with regulatory factors interacting with each other at every level. Yet despite this complexity, we can now comprehend cellular capacity on a scale our predecessors could scarcely imagine, and it is impossible not to be awed by it.

In the following subsections, I explain the regulome at the levels of DNA and proteins separately. Given the scope of this thesis, I chose not to include a dedicated section on RNA-based gene regulation. I am aware that the non-coding RNA (ncRNA) revolution represents one of the biggest paradigm shifts in cell biology, and just one year ago, the Nobel Prize in Physiology or Medicine was given to Victor Ambros and Gary Ruvkun for their pioneering work in microRNAs.

However, ncRNAs encompass a diverse class of molecules (microRNAs, long ncRNAs, snoRNAs, piwiRNAs, lincRNAs etc.) with distinct mechanisms of action, making their study an extensive field in its own right. In contrast, this thesis focuses on the organization of chromatin in three-dimensional space and its effects on gene regulation, primarily at the pre-transcriptional level. Consequently, CREs and their associated proteins take the center stage in this work and thus require more detailed attention.

2.1.2.1. Regulome on DNA level

Although it is possible for newcomers to join in the following decades, currently we can say that there are four well-established core components of CREs on DNA: promoters, enhancers, silencers and insulators.

Promoters

Gene promoters refer to the genomic locations where the transcription initiation machinery assembles[51]. The region that spans the immediate neighboring sequences of the transcription start site (TSS) is called the core promoter. Thanks to advances in high-throughput technologies in the 21st century, promoter features can now be explored from multiple perspectives. One of the most important breakthroughs in that aspect was the development of cap analysis of gene expression (CAGE), which enabled simultaneous identification of the exact 5' ends of transcripts and their expression levels[52]. ChIP-seq[53] made it possible to investigate promoters from a chromatin perspective, allowing researchers to examine the interplay between promoter sequences and epigenetic modifications, as well as the recruitment patterns of transcription factors. More recently, massively parallel reporter assays (MPRAs) have been developed, allowing researchers to investigate how transcription factor binding sites influence transcriptional output[54–56]. Together, these three key technologies, along with many other complementary approaches, allow us to examine core promoters in terms of their primary sequences, chromatin architecture, and functional activity. If all this information were to be distilled into a single statement defining a universal core promoter, it would be: there is no universal promoter type. As expected, promoter elements operate within a highly complex, multi-layered system that remains far from fully understood. Nevertheless, what we know today is vastly more than we knew two decades ago, and these insights form essential foundations for the work presented in this thesis.

When examining promoter features at the level of their primary sequence, one of the first distinguishing characteristics is the distribution of TSSs within the core promoter region. Based on this, promoters can be broadly classified into two types: narrow promoters, which have a single dominant TSS, and broad promoters, which contain multiple TSSs dispersed over a wider region[57]. Studies show that there is a general tendency of tissue specific genes to leverage narrow promoters while ubiquitously expressed genes were associated more with broad promoters, although this is not a universal rule[58].

Another long-known primary sequence feature is the presence of canonical and non-canonical core promoter elements (CPEs). Canonical CPEs are established sequence motifs, recurrently found in specified distances from TSSs. Most known examples include TATA-box, Initiator (Inr), TFIIB recognition element (BRE), downstream promoter element (DPE), downstream core element (DCE), motif ten element (MTE)

and X-core promoter element 1 and 2 (XCPE1, XCPE2) motifs[59,60]. One of the interesting findings acquired from recent genome-wide studies is that many core promoters lack these motifs, which indicate the existence of undiscovered elements. Moreover, some non-canonical CPEs, which represent intrinsic features of sequences that are overlapping the core promoter regions, such as CpG islands[61] or ATG deserts[62] have also been discovered to be classifiers of certain promoter subgroups. They are observed to direct the transcription initiation machinery assembly in the absence of canonical CPEs, although they can also work in combination. Discovery of new core promoter elements is still an active research area.

In the beginning of this century, some researchers shifted their focus from the primary DNA sequence towards the 3D chromatin, and several recurrent patterns for promoter regions were noticed. One key observation was that active promoters appeared to be depleted of nucleosomes, thereby facilitating access for the transcriptional machinery. This was initially noticed on *Saccharomyces cerevisiae*[63], but similar results were later published for both *Drosophila melanogaster*[64] and human[65] genomes. However, these early studies relied on micrococcal nuclease (MNase) digestion to map nucleosome positions. In 2009, Jin et al.[66] published a seminal paper where they showed that these active promoter regions were actually not nuclease depleted, but covered with dynamic histone variants of H3.3 and H2A.Z, making them vulnerable to MNase digestion. Over the past decade, studies leveraging ChIP-seq have further illuminated the prevalence of post-translational histone modifications at promoters. H3K4me3 and H3K27ac are two well-documented marks enriched at active promoters[59]. Studies carried out in the embryonic stem cells (ESCs) demonstrated an interesting role for the repressive mark H3K27me3 as well. It was shown that many promoters in developmental genes were co-occupied by H3K4me3 and H3K27me3, creating a bivalent state for the promoters and allowing timely activation or repression of genes within the developmental program[67]. Different combinations of histone marks have been studied and there is also a growing interest in the possibility of non-canonical histone marks that are affecting the regulatory functions[68]. Nevertheless, it is crucial to emphasize that the majority of knowledge about the associations between histone marks and promoter activity is fundamentally correlative[69]. As such, one must exercise caution in drawing conclusions that assume causality, which remains incompletely understood.

Another important aspect regarding the promoter architecture is the transcription factors, which are proteins binding to DNA to regulate gene expression. In section 2.1.2.2., they are explored in more detail. Here, I think it is important to highlight the level of complexity they introduce into the promoter machinery. ChIP-seq studies spanning hundreds of transcription factors are currently available, shedding light on the shared or differential binding patterns between cell types. Some promoters tend to have an enrichment of the same factor throughout its sequence with multiple binding sites, which are known as homotypic cluster of transcription factor binding sites

(HCTs)[70]. However, this does not represent a universal aspect regarding function, and more binding do not necessarily mean more transcriptional output[71]. Moreover, transcription factors can act as activators or repressors depending on the context[72]. It is likely that we are still missing a substantial number of factors that influence promoter function across all these dimensions.

Enhancers

In 1981, Banerji and colleagues published a paper in *Cell* that fundamentally changed our understanding of gene regulation[73]. In that study, authors transfected the HeLa cells with rabbit hemoglobin beta 1 gene. Transfection of the gene alone did not lead to a significant expression in the cells. However, when they also inserted a fragment of DNA sequences derived from simian vacuolating virus 40 (SV40) within 1-3 kilobase distance to TSS, they observed around 200-fold 'enhancement' of beta-globin. Moreover, the drastic effect was observed independent of the orientation, as both upstream and downstream transfections enhanced the expression. This was the first work where the scientific community was introduced with this new class of regulatory elements called enhancers. A couple of years later, they also reported the first enhancer in mammals[74].

The next breakthrough arrived in the 2000s with the arrival of high-throughput sequencing methods, which allowed us to investigate enhancers from a genome-wide scale. Currently, we estimate that the human genome contains ~1 million enhancers[75]. Enhancers vastly outnumber protein-coding genes and serve as the principal regulators of the spatiotemporal transcriptional programs essential for cell differentiation and development. Interestingly, despite being the regulators of mRNA output, when it comes to cell type classification, enhancers have been shown to be even stronger predictors than the transcriptome itself[76]. Thereby, interpreting this vast amount of recently generated information is crucial in understanding their roles in health and disease.

ENCODE's pioneering work in the first decade of this century revealed unprecedented insights about the cell-type specific histone modifications on distal enhancers which are crucial for cell-type specific transcriptional programming[31]. The first histone post-translational modification (PTM) found to be associated with enhancers was H3K4me1[77] and later it was shown that in many cases, it can be co-occupied with other histone modifications, such as H3K27ac[78], which affects activity. However, one must be cautious in interpreting these correlations with functional causality. Neither H3K4me1 nor H3K27ac are exclusive to enhancers, nor are they strictly required for enhancer activity[79]. In a study published in 2008, the authors investigated a selection of 39 histone marks associated with enhancers in T cells using a combinatorial approach and none of the marks turned out to be completely predictive single handedly[80]. On the other hand, we also know that enhancer-associated histone marks are not simply a

by-product. For example, knockdown of H2A.Z in mouse ESCs demonstrably alters the nucleosome architecture, which leads to the disruption of self-renewal and differentiation[81]. In another study, it was also shown that the binding of the same transcription factor to distal enhancers generate different regulatory responses based on the presence of H3K4me1[75].

PTMs found on enhancer-associated histone marks should not be viewed as isolated regulators of gene expression, as they exist within a complex crosstalk involving the enhanceosome. For example, EP300, which is frequently localized at enhancers, is a histone acetyltransferase that acetylates lysines on histone tails, resulting in a more open, and thus active, chromatin state[82]. The majority of enhancers are sensitive to EP300 inhibition[83]. Other important factors that are 'almost' universally associated with enhancers include Mediator and BRD4. I emphasize 'almost' because depletion studies have revealed that some subclasses of enhancers can remain unaffected by the loss of these proteins as well[84]. While Section 2.1.2.2 examines the protein aspect in greater detail, the key point here is the heterogeneous and intricate nature of enhancers, which makes it exceedingly difficult to establish universal principles for their annotation.

Another challenge lies in the fast-evolving nature of the enhancer sequences. According to a comparative study performed on 20 mammalian genomes, only around 1% of human-tissue specific enhancers were conserved[85]. However, sequence conservation does not necessarily equate to functional conservation. In fact, enhancers which seem to be species-specific from a primary sequence point, can still act as enhancers when transferred to another species[86]. This phenomenon has important implications on the nature of enhancers, especially on the conservation and flexibility of the transcription factor binding.

Indeed, clusters of transcription factor binding sites are among the first features that come to mind when considering enhancer characteristics. In terms of function, these sites are quite robust, as they can tolerate mutations to certain levels[87]. The order of the transcription factor binding sites seem to be critical, as it can still be conserved after hundreds of million years of independent evolution, and deletions of different parts of enhancers notably yields different results[86]. Techniques such as ChIP-seq have enabled the mapping of transcription factor binding sites across the genome, providing valuable correlations with regulatory elements. Yet, as with other aspects of enhancer biology mentioned above, caution is warranted when interpreting these bindings in terms of functional necessity. Degron-based studies, for instance, have demonstrated that many transcriptional programs remain unaffected after targeted degradation of transcription factors that were reportedly bound to their respective enhancer sites[79,88]. Altogether, these findings show that enhancers operate through remarkably complex arrangements, and achieving a comprehensive understanding of their function will require integrating

all the players that influence enhancer activity. To do this, we should also understand the mechanistic aspect.

Much like everything else discussed so far, the mechanisms underlying enhancer function are also not yet fully understood. However, our knowledge has grown significantly since 1981, and we now have more refined mechanistic models supported by a wide range of experimental evidence.

The early models have suggested a tracking system, where the transcription machinery components were loaded into the enhancer and then followed through the DNA until they reach the promoter[89]. However, this model was only fitting to the initial findings where the enhancers were regulating the nearby genes. Now, we know this not to be the case, and many researchers, including us with our work in **Paper III**, have shown via chromosome conformation capture studies that the majority of enhancer-like elements in fact skip the nearest genes[90]. In some cases, these long-range interactions can span a million base pair distance[91]. The complexity is not only about the distance either. Frequently, we observe a many-to-many relationship between regulatory pairs. An enhancer can target multiple genes, and a promoter can be targeted by multiple enhancers[68].

Currently, the prevailing models that explain enhancer mechanisms benefit from 3D genome biology. According to the loop model, which is widely accepted, genomic regions that are distant to each other in linear sequences come in close proximity in 3D space through chromatin looping. These looping events can be seen through high resolution microscopy or be inferred by chromosome conformation capture technologies including Micro-C, which is an essential part of my work. Moreover, forced enhancer-promoter (E-P) looping was shown to cause transcriptional activation[92], while inserting insulator elements in between E-P pairs have been shown to disrupt the expression of target genes[93], further supporting the loop model from a functional basis. It is still not very clear how enhancers transfer their information towards their cognate promoters. Although the prevailing assumption is direct transfer via looping, alternative experimentally supported suggestions, such as condensate mediated interactions, have also begun to emerge[94]. Considering that inhibition of proteins involved in E–P-associated condensates can, in certain contexts, disrupt the expression of the corresponding gene without affecting the interaction itself, it is possible that enhancers possess separate mechanisms for loop establishment and transcriptional activation within their toolset[95]. Discussing the mechanistic models individually does not fit within the focus of this thesis, but recent reviews about the subject, which I found to be quite informative, can be found within the references[79,94,96–98].

Evaluating regulatory biology through the lens of 3D genome allowed us to understand how genomic sequences that are 100s of kilobase pairs away from each other can pair

up to generate unique transcriptional programs. Furthermore, we now know that chromatin organizes in the nucleus following a set of hierarchical principles. One hallmark of this complex organization is the generation of TADs, which create regions that preferably interact with each other and remain isolated from the rest[99]. However, even these organizational principles are not sufficient to answer the fundamental question of enhancer biology. Out of all the possible genes and enhancer elements, why do they pair up the way they do? What kind of selection process do enhancers follow to target their promoters? This cannot be solely explained by chromatin architecture, as experiments show that swapping active enhancer elements from wild type E-P pairs can significantly drop transcriptional output[100]. This indicates the existence of a biochemical compatibility aspect in play, which has begun to draw attention in enhancer selectivity research[101].

Another prevalent feature of enhancers is the production of non-coding, short-lived transcripts known as enhancer RNA (eRNA). These transcripts not only serve as valuable markers for annotating enhancers, but also provide a quantitative measure of enhancer activity, as several studies have reported strong correlations between eRNA levels and the expression of their target genes[98]. However, similar to histone marks and transcription factor occupancy, the universal role of eRNAs in enhancer function remains unclear.

In summary, enhancers, with their diverse features and complex mechanisms, are distributed throughout the genome and direct spatiotemporal gene expression programs in ways that we do not yet fully understand. Although in terms of broad classification, they can be annotated as inactive, primed, poised or active[79]. Given the uncertainties described above, we did not rely on any single feature in isolation throughout our work when dealing with enhancers. Instead, by employing an integrated approach that separately assessed chromatin interaction networks, epigenetic marks, and RNA outputs, we aimed to identify putative enhancer elements with an increased level of confidence. For example, in **Paper III**, we identified 55,240 chromatin interactions with 1kb resolution by running Micro-C on 35 primary BCP ALL samples, and instead of relying only on the biochemical features derived from external studies, we annotated every anchor with an established TSS on the other end as an enhancer. As expected, these anchors were strongly enriched for known marks such as H3K27ac and H3K4me1, but neither these nor any other single protein marker was present at all anchors.

Silencers

In the simplest sense, silencers are the opposites of enhancers[102]. They represent stretches of DNA which can interact with cognate promoters through great distances via transcription factor binding. However, the purpose of these interactions is to downregulate gene expression. Yet despite this clear parallel, silencers have long been

overshadowed by enhancers. As of today, a quick look on PubMed returns more than 2 million articles containing 'enhancer' as a keyword, while this number is below 150,000 for 'silencer' (Accessed on 23 September 2025, https://pubmed.ncbi.nlm.nih.gov).

Although the definition of silencers versus enhancers may seem crystal clear at first glance, distinguishing between them in practice is far from straightforward. In fact, even the same authors who first described this new class of regulatory elements in 1985 went on to publish another study in 1987 showing that these very sequences could also act as activators[103,104]. This ambiguity has been repeatedly observed in high-throughput studies to various degrees. For example, a silencer element identified in one cell type can act as an enhancer in a different cellular context[105]. Even within the same cells, it has been suggested that a silencer for one gene could serve as an enhancer for another[106]. These CREs overlap in many features, and there is currently no universal criterion to reliably differentiate them. It remains an open question whether they truly represent two distinct classes of regulatory elements at all.

Nevertheless, there has been a growing interest in the identification of silencer marks, and some correlations have begun to emerge. A major focus point has established around H3K27me3 since its role in polycomb group mediated gene silencing was uncovered in a seminal 2002 *Science* paper[107]. Polycomb repressive complex 2 (PRC2) deposits an H3K27me3 mark to the chromatin, and it has been experimentally shown that promoters interacting with these regions exhibit lower transcriptional levels, which can be reversed by deleting the PRC2 binding regions[108]. In a study performed by Huang et al.[109], it was seen that H3K27me3 regions that were negatively correlated with nearby gene expression levels were enriched for repressive transcription factor binding. However, in the same study, it was also shown that around 75% of H3K27me3 marks are co-located with enhancer associated histone marks. The 3D genome map we generated in this thesis through multi-omics integration (**Paper III**) also revealed similar patterns. Specifically, we observed a significant enrichment of H3K27me3 on promoter-interacting loop anchors of silenced genes, yet this was far from a definitive feature, as the presence of the mark was not restricted to silenced genes and frequent co-localization with active enhancer-associated marks was also evident. Therefore, although its functional capacity to act as a repression machinery component is clear, the presence of H3K27me3 itself is not sufficient to classify a genomic region as a silencer. Considering the bivalent nature of transcription factors in isolation and in combination, as well as the limited source of histone modification information available, it may be possible for a silencer code – if such a thing exists – to be revealed through highly complex combinatorial approaches. The striking bivalency of distal CREs serves as a vivid reminder of how delicately gene regulation must be orchestrated through complex, interwoven mechanisms.

Insulators

Despite having distinct characteristics in terms of transcriptional activity or epigenetic marks, all three CREs mentioned so far share a common ability to directly participate in gene regulation. Insulators, on the other hand, stand out as indirect regulators. Rather than engaging promoters themselves, they exert their influence by preventing other CREs from doing so. As a separate class of regulatory elements, they were first introduced to the field via two landmark studies, published back-to-back in 1992[110,111]. Both demonstrated that when these DNA sequences were positioned between an enhancer and a promoter, they effectively blocked transcriptional activation. Notably, this blocking activity was highly position-dependent; when the insulator was moved outside of the E-P pair, the repression disappeared. Interestingly, insulators may also generate a positive effect on transcription, when they are placed around the silenced genes embedded in the heterochromatin. This observation led to the proposal of a "barrier" function for insulator elements, in which they block the spread of repressive heterochromatin[112]. Although our understanding of insulators has grown significantly more complex over the following decades, their functional landscape can still be broadly described by these two fundamental mechanisms. It is important to remember, however, that these mechanisms are not mutually exclusive.

Initial discoveries of insulators were made in the genome of *Drosophila melanogaster*, but similar studies soon followed in vertebrates. In 1999, a seminal paper published in *Cell* identified CCCTC binding factor (CTCF) as the key DNA-binding protein establishing insulator function in vertebrates[113]. This breakthrough was repeatedly validated and further strengthened by advances in chromosome conformation capture technologies. Today, CTCF is well recognized as the principal component of mammalian insulators, with vast majority of TAD borders reportedly harboring CTCF binding sites[114]. This dominance is interesting, considering the diversity of insulator sequences found in *Drosophila melanogaster,* which has a significantly smaller genome[115]. On the other hand, CTCF should not be considered as an isolated insulation factor. Cohesin, which is an essential component of loop extrusion model (explained in more detail in section 2.1.3) functionally associates and frequently colocalizes with CTCF[116]. Moreover, a more recent study employing unbiased CRISPR screens identified MAZ as an additional factor with insulator function in mammals[117]. These findings indicate that there may be other players in mammalian insulator function that are not discovered yet. Whether combinatorial studies will ultimately reveal novel subclasses of insulator elements remains to be seen.

2.1.2.2. Regulation of the 3D genome via proteins

It is now widely recognized that proteins are the fundamental building blocks of life. Unsurprisingly, they also play indispensable roles in chromatin organization and gene regulation. In this section, I highlight some of the key players, as understanding their

functions and mechanisms is central to the work presented in this thesis. Naturally, it is not feasible to cover the entire proteome involved in gene regulation within this volume. Many crucial components of gene regulatory architecture—such as chromatin remodelers, mediator complexes, or the DNA methylation machinery—do not have dedicated sections. This omission is not due to a lack of importance, but rather reflects their limited direct influence on the scope of this thesis. Instead, I have focused on the regulatory players that are directly part of the manuscripts included in this work.

Histone modifications

One of the most transformative periods in chromatin biology was 1970s, when the "beads on a string" model of chromatin organization was first proposed and subsequently supported through a series of pioneering experiments[118–121]. These studies fundamentally changed our understanding of DNA packaging, revealing that the genome is compacted by repeatedly wrapping around core histones—H2A, H2B, H3, and H4—to form nucleosomes. The interest about the connection between histone PTM and gene regulation however, dates before that, to the landmark experiments of Vincent Allfrey and colleagues[122,123]. They were the first to propose that histone acetylation could influence RNA transcription by reducing the positive charge on histones, thereby weakening their interaction with DNA. Today, in the era of NGS and genome-wide chromatin profiling, we can see that Allfrey's hypothesis was remarkably prescient. Moreover, emerging evidence suggests that histone modifications influence not only the transcriptional landscape, but also play critical roles in DNA replication, repair, and recombination. However, given the scope of this thesis, the focus will be concentrated on their roles in gene regulation.

The attention given to histone PTMs by the scientific community increased dramatically after two pioneering publications in 1996. In these studies, Brownell *et al.*[124] and Taunton *et al.*[125] provided the first clear identifications of bona fide histone acetyltransferase (HAT) and histone deacetylase (HDAC) enzymes, which are responsible for the addition and removal of acetyl groups on histones, respectively. These studies fundamentally changed our perspective on chromatin by revealing the dynamic and reversible nature of post-translational histone modifications. This new understanding naturally comes with an immense potential in terms of therapeutics and precision medicine, which would be described later in the book.

As of today, dozens of distinct histone modifications have been characterized, and mechanisms ensuring their reversibility have been identified for most of them. However, readers will primarily encounter PTMs involving lysine acetylation and methylation throughout the thesis, as these are the most well-established modifications associated with CREs. I acknowledge that there is a growing interest and a huge potential in other histone modifications, such as deimination, ADP-ribosylation, and ubiquitylation, but these were not part of the work presented here.

Each type of histone PTM involves distinct biochemical mechanisms, and yet some shared features exist. In a broad sense, a histone PTM can function either through exerting direct structural changes in the chromatin or via interacting with protein complexes that are members of the regulatory landscape. Moreover, all the modifying enzymes that interact with histone tails can be classified as 'writers', 'readers' or 'erasers'[126]. For instance, histone methyltransferases (HMTs) and demethylases (HDMs) serve as the writer and eraser enzymes for methylation, analogous to the roles of HATs and HDACs in acetylation. Further subgroupings may be observed in terms of functionality and genome-wide localization. While acetylations are typically associated with active regulatory elements, functional outcomes of methylations can be more context dependent. This is not counterintuitive, considering that methylation does not alter the charge of the histone protein, and by extension, the electrostatic interactions with DNA[127]. The specific lysine molecule which undergoes trimethylation also seem to be related to certain locations, as H3K9me3 is associated with constitutive heterochromatin while H3K27me3 is more frequent in facultative heterochromatin[127]. Nevertheless, for almost every correlation identified, there are notable exceptions, and no mark is strong enough to single-handedly classify a chromatin feature. These associations therefore do not represent universal laws but instead suggest regulation by more intricate combinatorial mechanisms—an emerging concept often referred to as the "histone code"[128].

Histone PTMs are a hot topic in current research, and it is likely that novel modifications will be discovered in the near future, along with deeper insights into their combinatorial crosstalk. Although the ChIP-seq and epigenetics revolution initially placed PTMs on histone tails at the center stage, there is now growing interest in modifications occurring on the globular core domains as well. Some researchers argue that core modifications may exert even more profound effects, as they can directly induce structural alterations in chromatin[126]. As an example, H3K56ac can directly influence DNA unwrapping, thus changing the entire chromatin landscape, making it accessible for transcription machinery[129]. However, in this thesis, histone PTMs serve primarily as a secondary layer of information. For example, in **Paper I**, we examined whether non-coding mutational hotspots were enriched for specific histone marks and assessed correlations with nearby differentially expressed genes. In **Paper III**, commonly studied histone marks were used as a comparative layer of information to evaluate their presence at chromatin loops identified by our experiments.

Transcription factors

In traditional terminology, transcription factors represent proteins that bind to specific stretches of DNA in regulatory regions and regulate gene expression. While still accurate, this definition is a flattened version of a multi-layered reality, much like describing a 3D object by its 2D shadow.

One of these deeper layers was famously unveiled in 2006, when Kazutoshi Takahashi and Shinya Yamanaka published a landmark study showing that a defined set of transcription factors could reprogram differentiated cells into induced pluripotent stem cells, fundamentally reshaping our understanding of cell identity[130]. Around the same time, the NGS revolution began, driving major advances in our understanding of transcription factor mediated gene regulation. This tightly controlled network — which begins with the extraordinary simplicity of a single cell — enables cells to differentiate, form tissues, respond to changes in the environment, and orchestrate development in a stepwise manner. Yet, despite our growing knowledge of the indispensable roles transcription factors play in this system, we still do not fully grasp how stretches of DNA and vast assemblies of transcription factors find each other at precisely the right moment and place, to cooperate with such remarkable specificity.

In a broad sense, transcription factors can be evaluated from two perspectives: DNA-binding affinity and functional output. At first glance, this distinction may seem artificial, since binding and function are expected to be tightly interconnected. Intuitively, one can assume that stronger binding by an activator should lead to greater activation of a target gene. However, recent studies suggest that this relationship is not universal. In fact, it is becoming increasingly unclear whether distinct classes such as "activators" and "repressors" exist at all.

More than two decades ago, Wasserman and Sandelin introduced the concept of the 'Futility Theorem', to highlight the limitations of using transcription factor binding site predictions to infer functional regulatory activity[131]. Their argument was based on computational approaches that rely on positional weight matrices (PWMs) to estimate relative binding affinities. However, subsequent genome-wide studies with multi-layered experimental data have confirmed that, at least to some extent, the Futility Theorem holds true — not all predicted binding events translate into biological function[132,133].

Regulatory elements and transcription factors exhibit a many-to-many relationship within the regulatory network. A recent demonstration of this phenomenon was published in *Nature* earlier this year[134]. In that study, the authors systematically examined transcription factor mediated gene regulation in yeast using Chromatin Endogenous Cleavage sequencing (ChEC-seq)[135] alongside a suite of functional assays. Despite the relative simplicity of the yeast genome, they found that a typical promoter is bound by a median of 15 distinct transcription factors, while a single transcription factor targets a median of 624 promoters. Strikingly, more than two-thirds of transcription factors functioned as both activators and repressors, and canonical sequence motifs were poor predictors of in vivo binding for most factors. Collectively, these observations reflect the diverse molecular mechanisms available in the gene regulatory network factor toolkit, even in an organism as simple as yeast. Unsurprisingly, this complexity increases by orders of magnitude in higher eukaryotes.

Let's begin dissecting this toolkit by examining the most basic mechanism first: the additive model. This model reflects the intuitive expectation that there is a linear relationship between transcription factor binding and its functional output. For instance, increasing the concentration of NF-κB family of transcription factors has been shown to produce a proportional effect on the expression of its target genes[136]. This outcome is mediated by the independent recruitment of RNA Polymerase II at HCTs, where multiple NF-κB bind[136]. Given the role of NF-κB in immune responses, the additive mechanism can be seen as an efficient strategy for generating graded responses to environmental signals.

Nevertheless, layers of evidence suggest that, in most cases, transcription factors act through cooperative and combinatorial mechanisms, which themselves encompass a variety of sub-modes[137]. This complexity exponentially increases the event space for regulatory function such as the dual functionality mentioned earlier.

One way to classify cooperative mechanisms is through their hierarchical dependencies. In simple cases, there is no such dependency. Two or more transcription factors can interact independently, stabilizing each other's DNA binding. Increased DNA occupancy with partnered transcription factors has been observed in single-molecule imaging studies[138]. An extreme version of this concept is called transcription factor collective, where multiple transcription factors act as a singular entity and removal of one component is enough to destroy the entire functional output[139]. Another example of co-operative protein-protein interaction is 'latent specificity', which represents alterations in DNA binding affinity of a transcription factor after interacting with a co-factor[140].

In more complex scenarios, cooperation depends on stepwise, tightly regulated events. A well-established example of this occurs with pioneering factors, which bind to nucleosomal DNA and recruit chromatin remodelers, consequently allowing access for other transcription factors to bind[141]. Notably, even in this 3D chromatin context, the primary DNA sequence remains relevant, as many chromatin remodelers exhibit preferences for specific DNA motifs[142]. Importantly, in such hierarchies, neither the pioneering factor nor the secondary factors are typically sufficient to drive transcriptional output in isolation; rather, function emerges from their coordinated and sequential interplay. These hierarchical structures are not limited to pioneering factor mediated processes either. For instance, binding of C/EBPα to an enhancer in myeloid cells alters the chromatin, which leads to the priming of another enhancer element, which is essential for the strong transcription of PU.1[143].

Evolutionary conservation of motif orders in certain enhancers was already mentioned in the corresponding section. This is another framework for a class of co-operative binding. For the regulatory networks that depend on this class, shuffling the order can generate different transcriptional outputs even with the same set of available

transcription factors, indicating that certain motif grammar is essential for certain protein-protein interactions. Interferon-β enhancer is a well-known example for this class, where small sequential changes can disrupt the entire regulatory machinery[144].

Altogether, it seems like no universal rule exists to govern transcription factor mediated gene regulation. Fortunately, absence of a universal rule does not mean there are no rules at all. The collective work on understanding this complex nature of transcription factors has not only yielded fundamental biological insights but also had a significant impact on leukemia research and treatment. The latter is thoroughly discussed in Section 2.2. Nevertheless, much remains to be done to fully decipher this regulatory architecture. Given that transcription factors constitute approximately 8% of all human genes[72], every effort that contributes even a small increment of knowledge to this vast and intricate system is a worthwhile endeavor.

Architectural proteins

The human genome spans approximately two meters in length, yet it is compacted within the limited space of the nucleus through a hierarchical series of structural principles (described in detail in Section 2.1.3). It is well established that the spatial organization of the genome influences gene expression, largely through the presence or absence of physical contacts between CREs and their targets. In this section, I focus on architectural proteins, the key molecular mediators that shape chromatin structure and facilitate these regulatory interactions.

In a broad sense, architectural proteins facilitate cis-regulatory interactions through two primary mechanisms[145]. The first is direct mediation, in which they bind to E-P pairs and actively promote their interaction. The second is indirect, involving the compartmentalization of the genome via protein–protein interactions. This organizational framework leads to (i) enhanced interactions within the same looped domain, (ii) spatial proximity between linearly distant regions adjacent to the loop, and (iii) insulation of interactions between neighboring TADs, as further described in Section 2.1.3.

The main architectural proteins in humans are CTCF and the cohesin complex[146]. CTCF binding sites are highly conserved across vertebrates and are notably enriched at TAD boundaries, as well as at loop anchors[99,147]. In **Paper III**, we show that this pattern holds true in pediatric BCP ALL cells, where 62% of detected loops had CTCF binding at both anchors, and 85% had it at least at one anchor. Cohesin on the other hand, do not have apparent sequence preferences, instead, it is recruited to chromatin by a loader complex[148,149]. The functional relationship between these two seemingly unrelated proteins was first hinted by chromatin immunoprecipitation studies, which showed a high degree of co-occupancy between CTCF and cohesin[116]. The loop extrusion model, in which cohesin extrudes loops along the DNA until it is halted by a CTCF-bound site, provided a mechanistic explanation that is still widely accepted[150]. In 2014, the

seminal paper of Rao et al.[151] demonstrated that over 90% of TAD boundaries were anchored by convergently oriented CTCF motifs, and subsequent motif inversion experiments have demonstrated that disrupting this orientation can impact loop formation[152], highlighting the critical role of CTCF motif orientation in genome architecture.

Numerous studies have shown that depletion of either CTCF or cohesin is sufficient to drive significant changes in chromatin architecture[153–155]. For example, CTCF depletion can lead to TAD merging and promote interactions between CREs that are normally insulated from one another[154]. Moreover, the number of CTCF binding sites at domain borders has been shown to influence TAD strength, with stronger borders exhibiting greater resistance to merging[156]. Nevertheless, CTCF-mediated boundary strength offers only a partial explanation for the persistence of certain TADs following CTCF depletion[154]. Given the diverse repertoire of architectural proteins identified in *Drosophila melanogaster*[145], it is reasonable to assume that alternative architectural elements may also exist in mammals. In **Paper IV**, where we highlight differences in mutation protection rates of TAD boundaries between compartments, we also noticed a compartment-level difference in CTCF occupancy, which may be related to this possibility. Indeed, several proteins that co-localize with CTCF have already been implicated in chromatin organization. One such protein is ZNF143, which has been shown to mediate CTCF binding at E–P loops[157]. Another is YY1, which forms loops between active CREs via dimerization, in a manner functionally similar to CTCF[158].

Altogether, architectural proteins facilitate gene expression through structuring the chromatin in different levels. However, neither the functional output nor the mechanism-of-action is necessarily shared throughout the entire class. This has been clearly demonstrated in *Drosophila melanogaster*, where some architectural proteins inhibit E–P interactions while others promote them[159]. Given that distinct subsets of genes are affected by cohesin versus CTCF depletion[160], it is reasonable to infer that non-overlapping regulatory functions exist in higher organisms as well. It is likely that new insights into CTCF, cohesin, and other architectural regulators will continue to emerge in the coming years.

2.1.3. Organization of 3D genome

Since the foundational studies of Carl Rabl and Theodor Boveri in the late 19th and early 20th centuries, it has been hypothesized that chromosomes are not randomly arranged within the nucleus[161]. The mechanistic and functional basis of this seemingly non-random organization, however, could only be gradually elucidated through over a century of subsequent research. Early insights were driven primarily by microscopy-based approaches conducted by classical cytologists. In 1928, Emil Heitz introduced the terms heterochromatin and euchromatin to describe chromosome fragments with

differential staining patterns[162]. Following the discovery of electron microscopy, more information on this dichotomy was unraveled. For example, in 1966, Don W. Fawcett demonstrated that condensed chromatin was frequently positioned against the nuclear periphery[163], a spatial pattern now recognized as a key feature of nuclear architecture. With the advent of more advanced molecular cytogenetic techniques such as fluorescence in situ hybridization (FISH), the concept of chromosome territories—originally proposed by Boveri—was experimentally confirmed[164]. Unfortunately, it is not possible to give credit to all great studies performed in the 20th century in this section, despite their undisputable roles in shaping our understanding of the 3D genome. Even in the work presented in this thesis, traditional cytogenetic methods continue to play a supporting role. However, deeper insights into the hierarchical organization of the genome only became accessible after the NGS revolution.

In 2009, Lieberman-Aiden et al.[165] published the first genome-wide chromatin interaction map in mammals. In this landmark study, the authors established a new paradigm in 3D genome research by introducing the Hi-C method and creating the concept of A/B compartments (to be discussed in more detail shortly) for the first time. Subsequent studies, leveraging increasingly sophisticated genome-wide crosslinking-based assays, uncovered a conserved and hierarchical organization of the genome in higher resolutions, revealing fundamental structural and functional features. More recently, microscopy-based approaches have made a spectacular come-back to the field, offering single-cell resolution that both challenged and complemented insights derived from bulk Hi-C data. In this section, I aim to bring together this diverse body of knowledge accumulated between 2009 and 2025 to present an integrated perspective on our current understanding of 3D genome organization.

The human genome is approximately 2 meters in length and therefore must be compacted in a highly organized manner to fit within the limited space of the nucleus while keeping functionality intact. This compaction can be broadly categorized into four levels of organization, from largest to smallest: chromosome territories, A/B compartments, TADs, and chromatin loops. Interestingly, these organization principles are highly conserved across species, indicating functional roles[166]. Among these, chromosome territories have been recognized the longest, as described above. The remaining levels, however, could only be studied at the genome-wide scale over the past two decades.

The concept of compartments first appeared in the aforementioned 2009 study, where authors performed a principal component analysis (PCA) on the Hi-C interaction matrix, and found out that based on the first eigenvector, the genome can be segregated into two compartments, each associated with distinct epigenomic features[165]. They named these A and B compartments, with the former associated with active histone marks, gene-rich regions, and open chromatin, while the latter represented a condensed state enriched for repressive marks. This spatial segregation has since been consistently

observed across numerous genome-wide chromosome conformation capture studies and is visually recognizable as a checkerboard pattern in heatmaps (**Figure 3A**). Naturally, compartment annotation is not a product of direct observation. It is a computational inference which is influenced by sequencing depth and analytical pipelines. Due to the developments in both fields, resolutions for compartments have also improved. For example, while early studies identified compartments at the multi-megabase scale, in **Paper III**, we detected them at a 500 kb resolution. Although it should be noted that these 500 kb windows do not necessarily represent uniform chromatin behavior, multi-omics integration consistently reveals broad biological patterns that support this segregation. For example, in **Paper III**, through RNA-seq integration, we demonstrate that highly expressed genes are preferentially correlated with A compartments. In **Paper IV**, we found mutation rates in BCP ALL genome to be significantly higher in B compartments, which supports an established phenomenon observed in heterochromatin and late-replicating regions[35,167].

The next organizational layer in 3D chromatin architecture in humans was discovered on a genome-wide scale for the first time in 2012 by Dixon et al.[99], where the term of topological domains was coined. In the same year, a focused study on the mouse X-inactivation center also described these structures. In the focused study however, authors preferred to call these structures 'TADs', which subsequently became the standard[168]. TADs are genomic regions typically ranging from 0.2 to 1 Mb in length, insulated by architectural proteins (most known example being CTCF) found in their borders[151,166,169]. A widely accepted mechanistic explanation for TAD formation is the loop-extrusion model[150]. According to this model, once cohesin complex is loaded onto chromatin, it pulls DNA in both directions until a convergently oriented CTCF bound region is reached. These boundaries act as insulators between neighboring TADs and, by bringing linearly distant regions into spatial proximity, increase the likelihood of regulatory loops forming within a TAD. The resulting insulation pattern can be visualized as adjacent triangular domains in Hi-C/Micro-C heatmaps, as exemplified in **Figure 3B**. The mechanistic underpinnings of these loop-forming events remain an active area of investigation, some aspects of which have already been discussed in Section 2.1.2.2.

Although the term 'hierarchical structure' is widely used to describe these organizational layers, I find this terminology highly problematic from a literal point of view. Current evidence suggests that these structures neither follow a strict top-down chain-of-command nor share a universal set of rules for self-organization. For instance, while the loop extrusion model for TAD formation is experimentally supported through multiple lines of evidence, when it comes to compartmental segregation, phase separation appears to be the major driving force[170]. Moreover, decreasing cohesin occupancy weakens TAD borders, while strengthening compartmentalization[171]. These kinds of discrepancies are not only limited to larger organizational structures either. From one

point of view, one may assume that TADs indeed represent higher order features than regulatory loops in 3D organization, as disruption of their borders can lead to domain merging and enhancer-promoter rewiring[172]. However, this would be a premature conclusion, since studies also show that E-P loops can form independently, prior to TAD establishment[173]. Based on these findings, it is not logical to support the existence of a strict hierarchy. Instead, chromatin appears to function as a dynamic self-organizing system with remarkable complexity, governed by a set of rules (some of which can be universal, while some are specific to individual structural units) that allow for both independent and interdependent behavioral patterns of attraction and repulsion. Nevertheless, despite this complexity, chromatin is still bound by the laws of polymer physics. This makes it possible for computational simulations to recapitulate experimental data to a certain degree, through a set of defined physical and biochemical parameters[155,174–176].

Another essential point to consider when interpreting genome-wide chromatin interaction data from high-throughput methods such as Micro-C or Hi-C is that these outputs reflect the aggregate signal of millions of individual cells. Given the high degree of conservation observed across cell populations in 3D genome features derived from these bulk methods, and the substantial body of experimental evidence supporting their functional roles, one might intuitively assume a stable and homogeneous chromatin structure within a population or at least overlook the potential impact of cell-to-cell heterogeneity on functional outcomes. Recent findings from the single-cell technologies however, provided a significant challenge against this view. Current evidence demonstrates extensive variability and dynamism in 3D genome organization, and only a small fraction of cell units exhibits the organizational patterns found in bulk Hi-C data at a single time point[177–179].

One aspect of this heterogeneity comes from the dynamic nature of chromatin interactions. In the previous sections, many-to-many relationship between transcription factors and their binding regions have already been discussed. Single-cell/single-molecule imaging studies provide an additional layer of complexity to this phenomenon, as they show that for more than 95% of their time, transcription factors are not bound to their cognate target sites[180]. When this is combined with the diverse sets of protein complexes that can form to generate distinct functional outcomes, the chromatin interaction landscape begins to resemble a probabilistic system modulating transcriptional output, rather than a simple binary switch. Additional layers of heterogeneity within a cell population may result from factors such as asynchronous cell cycle stages, variability in regulatory protein levels, and differences in extracellular signaling response[166,181]. Notably, the smallest unit of heterogeneity is not the cell itself either, as allelic differences can also lead to divergent chromatin architectures within the same nucleus. For example, an E-P interaction may be present on one allele but absent on the other. The relevance of this phenomenon within the BCP ALL disease

context is evident across our work. For example, in **Paper I**, where we systematically investigated non-coding mutations in BCP ALL using two diagnostic cohorts, variant allele frequencies indicated heterozygosity for most single nucleotide variants (SNVs) identified in that study. In **Paper II**, we showed how such heterozygous somatic events can influence ASE landscape of HeH ALL. Annotating allele-specific regulatory events from bulk 5C data, however, is a challenging endeavor, as simple haplotype phasing approaches would fail to extract informative reads for the vast majority of identified interactions[182]. Perhaps the best-known example of such allele-specific and probabilistic chromatin organization is XIST-mediated X-chromosome inactivation[168].

Although the apparent contradiction between the evolutionary conservation of these organizational features and the significant cell-to-cell heterogeneity may seem puzzling at first, this complexity becomes less surprising when viewed through the lens of single-cell transcriptomics. Since transcriptional heterogeneity is an established biological reality, it is only logical to expect a corresponding heterogeneity in 3D chromatin organization, given the deep interconnection between gene expression and chromatin structure. In a review summary published in *Science*, the authors discuss the possibility of this cell-type-specific variability itself to be an additional layer of regulation, which itself is also actively regulated[166]. In any case, the existence of a link between heterogeneity and function is clear, which warrants further investigation.

In summary, we are still far from achieving a comprehensive understanding of the 3D genome organization system. Given the intricate, probabilistic, and interdependent features that govern this architecture, future discoveries will likely benefit from advances in single-cell technologies. Importantly, this is not merely a matter of scientific curiosity, as the 3D genome carries significant implications for cancer research. At every layer of 3D genome organization, there is evidence of oncogenic association. For example, canonical translocations in cancers often occur between loci that are found to be in close spatial proximity[183]. Throughout the thesis, we advanced this knowledge on all organizational levels for our disease of interest, BCP ALL. Some exemplary findings include the similar compartmentalization patterns between immature hematopoietic cells and BCP ALL (**Paper III**), subtype specific TAD formations (**Paper III**) and differential mutagenesis patterns found within and outside chromatin loops (**Paper IV**). **Figure 3** provides an example, based on our own data, of how different layers of 3D genome organization are visualized in Micro-C derived heatmaps. The broader implications of these findings for the 3D genome landscape in BCP ALL, along with potential future directions, are discussed in Sections 3 and 4, respectively.

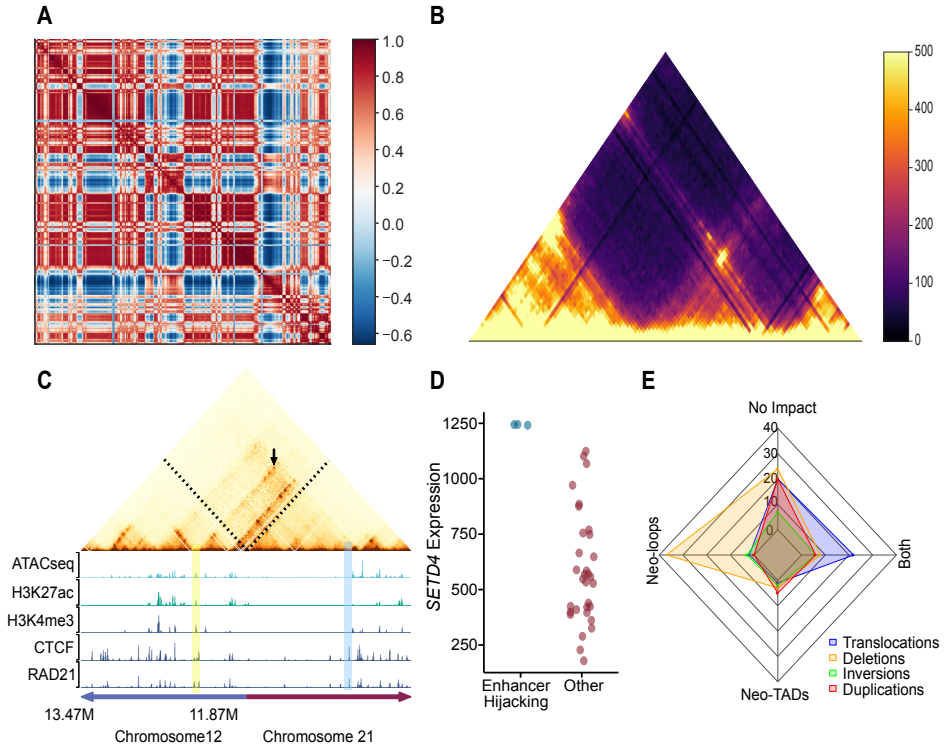


Figure 3 | Studying the 3D Genome with Micro-C. This figure illustrates how chromatin organizational structures can be inferred and visualized from Micro-C data, using the dataset generated for Paper III. (A) Checkerboard patterns are shown from a merged dataset of 14 high-hyperdiploid (HeH) B-cell precursor (BCP) acute lymphoblastic leukemia (ALL) samples. The exemplary region chr1:70–120 Mb demonstrates large-scale A/B compartmentalization, where red and blue squares represent genomic regions that preferentially interact within the same compartment type. (B) Using the same dataset, a triangular heatmap of chr1:91.5–92.5 Mb was generated to illustrate topologically associating domain (TAD) organization. TADs appear as triangles with warmer colors, representing regions with higher interaction frequencies, while the boundaries between them marks zones of insulation from adjacent regions. (C) An example of a cis-regulatory loop resulting from an enhancer-hijacking event caused by the t(12;21) translocation in a sample carrying the *ETV6::RUNX1* fusion. The loop interaction is indicated by an arrow, linking an enhancer from chromosome 12 (yellow) to the *SETD4* promoter on chromosome 21 (blue). (D) RNA-seq data from samples with and without the enhancer-hijacking event show elevated *SETD4* expression in the affected cases. (E) Quantitative summary from Paper III showing how different classes of structural variation (SV) can alter 3D genome organization by inducing neo-loops and neo-TADs. Each SV class is color-coded according to the legend in the lower-right corner.

2.2. Disease of interest: Pediatric ALL

2.2.1. Normal hematopoiesis

Every day, more than 500 billion blood cells are produced within the bone marrow of an average adult[184]. The complex system that tightly regulates this process, both under steady-state and stress conditions for a lifetime, is known as hematopoiesis.

In broad terms, the hematopoietic system relies on two fundamental cell-fate decision mechanisms: self-renewal and differentiation. The former ensures life-long maintenance of the stem cell pool, while the latter enables the formation of diverse cell types that carry out essential functions, such as adaptive immune responses, oxygen transport, and tissue repair. Both properties are indispensable for survival, and inhibition of either one ultimately results in bone marrow failure[185]. Accordingly, a thorough understanding of this system is a prerequisite for elucidating the origins of hematological malignancies and developing effective treatments.

The first clear indications of the fundamental properties of hematopoietic stem cells (HSCs) came from bone marrow transplantation experiments performed in 1950s. In these early studies, researchers showed that transplanted bone marrow cells from healthy donors can repopulate in lethally irradiated mice and increase the expected lifespan[186,187]. In the 1960s, James Till and Ernest McCulloch developed the spleen colony-forming unit assay (CFU-S) and demonstrated that individual bone marrow cells could generate colonies containing multiple blood lineages[188]. Subsequent cytological studies confirmed the stem cell origin of these colonies, providing the first clonal proof to the concept of multipotent progenitors (MPPs)[189]. In the next two decades, researchers used early clonal markers — such as radiation-induced chromosomal aberrations or retroviral integration sites — for bone marrow cells to derive insights regarding the differentiation trajectory of HSCs[190,191]. A wide range of results were observed; some tags were shared by all cell types, while some where restricted to a certain lineage, some seen after multiple generations while some were short lived. These observations supported a top-down hierarchical model in which HSCs sit at the apex with broad differentiation potential, which becomes progressively restricted at each step and is lost entirely in fully mature cells. Following the developments in flow cytometry, which enabled the characterization of bone marrow cells at the single-cell level, Weissman and colleagues were the first to isolate HSCs by using monoclonal antibodies[192]. This breakthrough was soon followed by the identification of additional surface markers that enabled the isolation of downstream progenitor populations, such as common lymphoid progenitors (CLPs)[193] and common myeloid progenitors (CMPs)[194]. Collectively, these advances provided strong experimental support for the traditional hierarchical model of hematopoiesis.

Despite being supported by multiple layers of evidence, "the hierarchical model of hematopoiesis" was still a model, and like all models, it relied on certain assumptions. **Assumption 1**: Each cell type is a discrete unit, formed through a discrete transition mechanism. **Assumption 2:** There is only one differentiation trajectory leading to each mature cell type. **Assumption 3:** Each discrete cell type represents a homogeneous population with identical differentiation properties. Perhaps, these assumptions were due to the limited availability of cell surface markers, which led to the isolation of clearly distinct cell populations while masking intermediate states. However, with the advent of single-cell omics, these assumptions started to break down. Transcriptomic analyses at single-cell resolution revealed a **continuum model** in which cells progress along continuous trajectories rather than through punctuated, discrete transitions[195]. Furthermore, the idea of a preset trajectory from the HSC to the mature cell was also challenged by studies on megakaryocytes, as they were shown to frequently arise directly from HSCs, bypassing the progenitor stages[196]. In addition, neither HSCs nor their downstream progenitors are homogeneous populations. Instead, they display lineage biases that are detectable at both the transcriptional and epigenetic layers, such as chromatin accessibility and DNA methylation patterns[197–199]. That being said, perturbations on certain hematopoietic transcription factors reportedly alter cell-fate decisions, showing that there may still be a discrete layer operating on the depths of the continuum model[195,200].

The origins of lineage bias in hematopoietic stem cells are not yet fully understood[201]. It is likely that both intrinsic factors (e.g., transcriptional and epigenetic programs) and extrinsic cues (e.g., niche signals, systemic influences) contribute to shaping these biases. Importantly, heterogeneity is not confined to the stages of differentiation; considerable variation also exists between individuals[195]. Such inter-individual differences can influence hematopoietic trajectories over a lifetime and may predispose certain individuals to conditions such as clonal hematopoiesis. This concept is particularly relevant to our work, as the biology of leukemic stem cells (LSCs) is deeply rooted in understanding how somatic alterations arise and influence stem cell fate.

Altogether, addressing the complexity of the hematopoietic system requires multi-modal integration of diverse data layers such as chromatin accessibility, 3D genome organization, spatial transcriptomics, histone modifications or RNA m^6A profiling[202]. However, even the most comprehensive molecular analyses may fall short if the bone marrow niche is neglected. Multiple lines of evidence show that cell–cell interactions within the bone marrow microenvironment are essential not only for the maintenance of HSCs but also for initiating their differentiation[203]. Studies indicate that for some cases, perturbations in the bone marrow niche can be sufficient to drive leukemogenesis.[204,205]. As a summary, numerous extrinsic and intrinsic factors can alter transcriptional landscapes without changing the underlying genome sequence. While transcriptomics can reveal the end-state of affected programs, regulatory and 3D

genome analyses are required to uncover the upstream events that shape hematopoietic alterations. Only through a comprehensive understanding of this highly complex network will we be able to explain the underlying nature of these multiple cellular states and identify potential mechanisms that govern the activation or quiescence of HSCs, which may be crucial for treating hematological malignancies.

2.2.2. Etiology and pathogenesis of BCP ALL

As already mentioned in the beginning of this book, pediatric BCP ALL represents a magnificent success story in the fight against cancer, with overall survival rates of an intrinsically lethal disease reaching to over 90% in 5 decades. Unfortunately, we cannot claim the same success when it comes to the etiology of the disease, as the cause of BCP ALL is still not fully understood.

According to Sun Tzu, the principle of victory lies in knowing both ourselves and the enemy. Looking at the tremendous developments in understanding the hematopoietic system, as described in the previous chapter, we might say that the field has made considerable progress on the "self" side. It is only reasonable to assume that increased knowledge of the etiology and pathogenesis of BCP ALL, which represents the 'enemy' in this context, would bring insights into the lost battles, mainly due to relapse. In this section, I summarize the collective insights from this perspective.

In the simplest terms, BCP ALL is the pathogenic consequence of genetic events that interfere with normal B-cell differentiation, leading to the malignant proliferation of immature B-cell progenitors in the bone marrow, peripheral blood, and at extramedullary sites. These differentiation blocks can occur at various stages of B-cell development, as demonstrated by recent single-cell experiments[206]. Given the highly complex and heterogeneous nature of the hematopoietic system, this is not surprising, as such complexity inherently creates a broad event space for "things that may go wrong".

One of the most critical vulnerabilities lies in the precise orchestration of hematopoietic transcription factors. Studies show that ~40% of pediatric BCP ALL cases have mutations in transcription factors essential for lymphoid development[207]. A well-known example is IKZF1, which is pivotal for regulating genes associated with the lymphoid differentiation[208]. Knockdown of this gene or disruption of its DNA-binding domain can be sufficient to completely block B-cell maturation[207,209]. Another example is PAX5, a which is another frequently mutated transcription factor that is crucial for B-cell lineage commitment[210]. Diverse sets of transcription factor mutations can synergistically work with oncogenic signaling mechanism and drive leukemogenesis[210]. A notable example is the JAK/STAT signaling pathway, which confers a substantial proliferative advantage to leukemic cells[207,211]. Studies have shown that the combination

of PAX5 loss-of-function and constitutively active JAK/STAT signaling can produce a highly penetrant disease, consistent with the widely discussed two-hit model[207].

Despite the steady increase in our knowledge of the factors involved in B-cell differentiation since the advent of NGS and single-cell technologies, the historical hallmark of BCP ALL has been chromosomal aberrations and structural variants (SVs), particularly gene fusions. As of today, SVs not only represent essential criteria for therapeutic decisions but also provide important insights regarding the pathogenesis of BCP ALL. The most common, and perhaps the one of the most widely studied, fusion in pediatric BCP ALL is *ETV6::RUNX1*, which occurs in approximately 25% of cases[212]. Multiple lines of evidence indicate a two-hit model for this fusion, in which *ETV6::RUNX1* acts as the initiating event, and subsequent secondary mutations drive progression to overt leukemia[212]. One compelling line of support comes from studies of archived neonatal blood spots, which showed that children diagnosed with *ETV6::RUNX1*-positive BCP ALL can carry the fusion from birth, suggesting a prenatal origin[213,214]. An interesting follow-up came from larger population screenings, which revealed that the frequency of this event is approximately 100 times higher than anticipated from leukemia incidence rates, and the vast majority of individuals carrying the fusion actually do not develop the overt disease[215]. These findings suggest that *ETV6::RUNX1* alone is insufficient for leukemogenesis, and that disease development depends on the probability of acquiring additional "hits" — which may be second, third, fourth, or subsequent events — likely influenced by a wide range of genetic, epigenetic, and environmental factors.

Nevertheless, two-hit model does not represent the entirety of BCP ALL pathogenesis. The most well-known outlier to this framework is the *KMT2A*-rearranged (*KMT2A*-r) leukemia, which is the most common form of ALL occurring in infants, accounting for roughly 75% of cases[216]. Studies in experimental models show that these re-arrangements can be sufficient to induce leukemogenesis on the stem-cell level without the aid of additional events, although without imposing a strict lineage-commitment[217].

Altogether, I believe the examples discussed so far provide sufficient evidence to conclude that BCP ALL is a highly complex and heterogeneous disease with multifaceted pathogenic patterns. Perhaps it should also be discussed to what extent we can encompass such distinct pathogenic mechanisms under the same disease name. I find this similar to the species problem in evolutionary biology — at what point does an organism become a distinct species?

The complex array of pathogenic mechanisms suggests the involvement of an equally complex set of causal agents, which may partly explain our limited understanding of the disease's etiology. This does not mean, however, that the scientific community has failed to produce fruitful ideas; in fact, several hypotheses — some dating back more than a century — remain active today, supported by varying degrees of evidence.

Perhaps the oldest ongoing discussion concerns the possible role of an infectious agent in leukemogenesis. This is not surprising, given that viral-induced leukemogenesis has been firmly established in animal models in the 20th century[218–220]. Although a human equivalent of such a pathogen has not been identified, intriguing epidemiological patterns have led scientists such as Kinlen and Greaves to propose infection-associated causal models involving the indirect effect of viral or microbial agents[221–223]. Examples of these associations include inverse correlations with early-life infections and the amount of time spent in day-care[221].

Of course, one of the most common pitfalls in this area — and one that every scientist should be aware of — is the distinction between causation and association, so results must be interpreted cautiously. On the other hand, support for the infection hypothesis is not limited to epidemiological studies. In experimental settings, inflammatory cytokines have been shown to drive the expansion of leukemic cells[224], and transferring mice with pre-leukemic mutations from a germ-free environment to one containing microbial pathogens can trigger leukemogenesis[225]. Nevertheless, while infectious agents may play a role in certain leukemias, the evidence in humans remains inconclusive at this point. Moreover, the heterogeneity of pathogenic mechanisms in BCP ALL, as discussed earlier, warns us against drawing overly broad conclusions.

Although ALL is largely considered a somatic malignancy, certain genetic predispositions exist, raising the possibility of a genetic susceptibility that has been underappreciated. A widely known example is trisomy 21, which is associated with higher risk of hematological malignancies[226]. The discovery of an aggressive ALL subtype, intrachromosomal amplification of chromosome 21 (iAMP21), provided an important biological context to this risk association[227,228]. Genome wide association studies (GWAS) have yielded valuable insights by identifying a range of single nucleotide polymorphisms (SNPs) associated with disease risk[229–231]. While GWAS findings must be interpreted cautiously due to statistical constraints, the presence of risk-associated SNPs found in transcription factors associated with hematopoietic lineage differentiation — such as *IKZF1*, *ARID5B*, and *GATA3* — supports the existence of a heritable component in ALL etiology. Furthermore, an increased, albeit modest, risk is also observed among siblings, providing additional support for the genetic predisposition hypothesis[232].

Altogether, it is unlikely that a singular cause will be identified in the future that can account for all forms of pediatric ALL. However, integrated studies and expanding datasets have the potential to yield valuable insights, possibly enabling the identification of risk factor combinations in specific patient subsets with greater accuracy. Perhaps, for some very specific groups, the disease may even become preventable. A major challenge lies in the heterogeneous nature of the disease, which makes subset-specific approaches and precise molecular classification the essential components of future studies, which is further discussed in the following section.

2.2.3. Molecular classification of pediatric BCP ALL

The heterogeneous nature of leukemia has long been recognized, and ongoing classification efforts — driven by evolving methodologies — have yielded important insights with direct implications for clinical decision-making. The first widely adopted system was the French–American–British morphological classification, introduced in 1976[233]. This was later followed by composite classifications that integrated both morphological and cytogenetic findings[211]. With the advent of NGS, novel subgroups with cryptic cytogenetic profiles were discovered. As of 2025, more than 20 subgroups of BCP ALL are recognized by International Consensus Classification (ICC)[234].

Although each subtype presents with a distinct genetic profile, their defining features can be broadly classified into five categories: chromosomal aneuploidy, recurrent gene fusions, other structural variations, point mutations, and gene expression–based subtypes. It is important to highlight that these categories are not mutually exclusive and are often interconnected. For example, a distinct gene expression program can also be associated with (or subgrouped by) specific sets of mutations.

The most common chromosomal aneuploidy in BCP ALL is HeH, defined by a modal chromosome number from 51 to 67, and accounting for approximately 30% of pediatric cases[235]. Interestingly, the chromosomal gains are not random, with enrichment found for chromosomes X, 4, 6, 10, 14, 17, 18, and 21[235,236]. Genomic studies revealed patterns of associated secondary events such as RAS pathway mutations[236]. Pediatric cases of HeH ALL are generally associated with a favorable prognosis. Other less common aneuploidies include low hypodiploidy (31–39 chromosomes) and near-haploidy (24–30 chromosomes), which account for approximately 1% and 2% of cases, respectively[27]. Importantly, the reported incidence of hypodiploidy may be underestimated due to masked hypodiploidy, in which a hypodiploid clone undergoes whole-genome duplication, resulting in an apparently hyperdiploid karyotype[237]. This can obscure accurate prognostic assessment and influence therapeutic decision-making.

The first cancer-associated chromosomal translocation was unraveled in 1973, when Janet Rowley demonstrated that the Philadelphia chromosome (which produces *BCR::ABL1*) in chronic myeloid leukemia (CML) resulted from a reciprocal exchange between chromosomes 9 and 22, using quinacrine fluorescence (Q-banding) and Giemsa (G-banding) techniques[238]. *BCR*::*ABL1*-positive leukemia is historically classified as high-risk, however, the prognosis has significantly improved after the discovery of ABL1 tyrosine kinase inhibitors (TKIs)[239]. Although considered as the main driver of CML, this fusion is also seen in 3-5% of pediatric ALL cases[240].

Some of the translocations we know today, on the other hand, can be cryptic to traditional cytogenetic methods. The most common one is *ETV6::RUNX1*, which is seen on approximately 25% of cases and associated with favorable prognosis[241]. Another

hematopoiesis-related transcription factor fusion is *TCF3::PBX1*, which accounts for roughly 5% of cases[241]; in rare instances, *TCF3* can also fuse with *HLF*[242]. It is important to note here that not all fusions are well defined on both ends, in many cases a re-arrangement pattern converging on a singular gene is observed, where one key transcriptional regulator can be fused with multiple partners. Of these, *KMT2A*-r has been known the longest, with more than 80 discovered partners, the most common one being *AFF1*[27]. It accounts for the majority of infant cases and as mentioned earlier, is associated with dismal prognosis. Importantly, many of the more recently recognized rearrangements were discovered or largely characterized through their distinct transcriptional profiles identified in RNA-seq studies. Examples include *DUX4*-[243,244], *MEF2D*-[245], *NUTM1*-[246], and *ZNF384*-rearrangements[247].

Besides translocations, other SVs of varying complexity are also observed in pediatric BCP ALL. A notable example is iAMP21, which typically arises through breakage–fusion–bridge cycles and chromothripsis events affecting chromosome 21[248]. This subtype is more common in older children and is generally classified in the poor-prognosis risk group, although treatment advances have improved outcomes in recent years[249]. There are also other SVs that are frequently observed but on the lower end of the complexity spectrum. Deletions in important transcriptional regulators of lymphoid lineage such as *CDKN2A/B*, *IKZF1* and *PAX5* are frequently seen, with varying degrees in different genetic subtypes[250]. Studies show that secondary lesions identified in established subtypes may have heterogeneous associations with prognosis. For example, the combination of the deletions mentioned above can confer a worse prognosis than an *IKZF1* deletion alone; however, when the cooperating deletion involves *ERG*, this adverse effect is mitigated — underscoring the complexity of leukemia evolution and treatment response[251].

For the vast majority of established subtypes, WGS has also revealed secondary associations at the point-mutation level. Interest in these lesions has been growing because of their potential to refine classification and identify therapeutic targets. One such example, *PAX5* P80R, has met the threshold for significance and distinctiveness and is now formally recognized by the WHO as a separate subtype[234]. Although few point mutations are likely to recur at this level of frequency, it is plausible that combinations of point-mutation–driven subgroups (or sub-subgroups) will be delineated in the future.

High-throughput transcriptomics has revealed several SVs that were otherwise cryptic, as mentioned above. However, this increase in sensitivity reflects only part of RNA-seq's impact on the molecular classification of BCP ALL. Large-cohort studies have identified clusters of cases with transcriptional profiles closely matching those of established SV-defined subtypes but lacking the corresponding genomic alteration. This led to the recognition of *ETV6*::*RUNX1*-like, *BCR*::*ABL1*-like, *KMT2A*-r-like, and *ZNF384*-r-like subtypes, all of which are now formally acknowledged by the WHO

as separate entitites[234]. In **Paper I**, we leveraged this classifier property of RNA-seq to determine the transcriptional placement of cases harboring a non-coding mutational hotspot of interest (**Figure 4**).

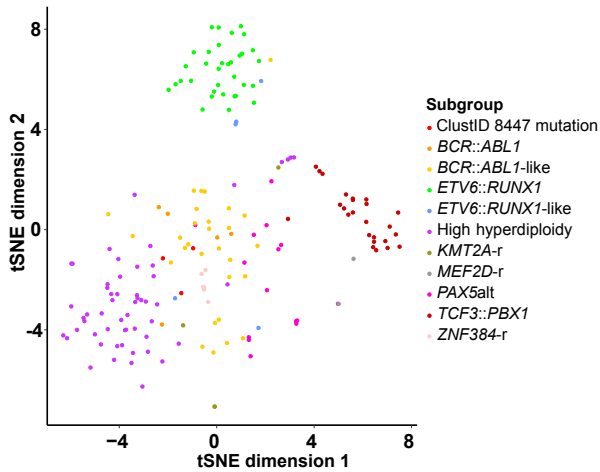


Figure 4 | Transcriptional profile of RNA-seq data from Paper I. t-SNE analysis of RNA-seq data demonstrates that leukemia subtypes form distinct transcriptional clusters. This clustering property was leveraged to explore the transcriptional profiles of cases carrying the recurrent 8447 hotspot mutations described in Paper I.

Altogether, the molecular classification of BCP ALL is an active field of considerable clinical importance. Over recent decades, the proportion of unclassified cases — the so-called "B-other" group — has steadily declined, while provisional entities continue to be proposed. Integrated multi-omics studies now point to the existence of nested subgroups, with more likely to be discovered. For example, the 2023 ICC subdivides the *BCR*::*ABL1*-like subgroup into three categories: ABL-class rearranged, JAK–STAT–activated, and NOS (not otherwise specified)[234]. Such layers of complexity have the potential to transform risk stratification and therapeutic decision-making — a process already underway in current clinical practice. The next section discusses the scope of diagnostics and contemporary treatment approaches.

2.2.4. Treatment

Few areas of oncology have achieved treatment outcomes as remarkable as those seen in pediatric ALL. Over the past century, survival rates have improved from virtually 0% to more than 90% in high-income countries[252]. Notably, this achievement is not primarily due to the discovery of new anticancer drugs, but rather the systematic optimization of dosing schedules and combinations of a relatively small set of existing agents[253]. Designing such regimens has proven no less complex than drug discovery

itself, and today's outcomes represent the culmination of decades of cooperative trials and sustained international effort.

The contemporary therapy for pediatric ALL, widely applied worldwide, stems from the pioneering "Total Therapy" concept introduced by Pinkel in the 1960s and later refined by Riehm et al. in the 1970s[254,255]. Briefly, current protocols are divided into five stages: induction, consolidation, interim maintenance, delayed intensification, and maintenance. In simplest terms, the guiding principle is to escalate chemotherapy to the maximum tolerated intensity in high-risk (HR) patients, while keeping treatment less intensive in standard-risk (SR) cases to minimize long-term toxicity. Naturally, the success of this approach is highly dependent on the accuracy and sensitivity of risk stratification. It is therefore important to highlight the accepted risk factors in today's clinic, as they are highly interconnected to the treatment decisions.

At the time of diagnosis, the first major risk factor is the white blood cell (WBC) count, where less than 50,000/μL is defined as SR, and higher numbers represent the high-risk HR group[256]. Another variable is age: children younger than 1 year or older than 10 years are more prone to worse outcomes[256]. Of course, age itself is not the sole determinant of prognosis, as discussed in the molecular classification section, where infants were noted to be enriched for adverse subgroups such as *KMT2A*-r ALL[257]. Thus, genetic subgroups determined at diagnosis serve as important predictors, i.e. HeH and *ETV6::RUNX1* positivity are both associated with a superior prognosis[253]. Central nervous system or other extramedullary involvement is likewise considered an adverse predictor[253]. These baseline risk factors determine the combination of drugs used in the induction phase, which typically includes dexamethasone (DEX) or prednisone (PRED), vincristine (VCR), asparaginase (ASP or PEG-ASP), and—in HR groups—daunorubicin (DAUN)[252]. The purpose of induction therapy is to reduce leukemic burden to undetectable levels, a goal achieved in the vast majority of cases.

On day 29 of induction therapy, MRD is assessed to evaluate the success of treatment. Since the early 1990s, MRD has been used as a predictive marker of disease progression, and it remains the strongest prognostic factor in pediatric ALL today[24,253]. A common cut-off value for MRD in risk stratification is 0.01%, although this may vary slightly across different study groups[253]. Based on a combination of baseline and post-induction risk factors, consolidation therapy is administered in risk-adapted regimens to target leukemic cells that have survived induction. This phase may include drug combinations such as methotrexate (MTX), mercaptopurine (6-MP), VCR, cyclophosphamide, cytarabine (Ara-C), or PEG-asparaginase, depending on the protocol. Following consolidation, interim maintenance typically consists of MTX, 6-MP, and VCR, aiming to eliminate any remaining leukemic cells. The delayed intensification phase, lasting about eight weeks, resembles a repetition of the first two phases and is administered even in patients without detectable disease, as it significantly reduces relapse risk[258]. Finally, patients proceed to maintenance therapy, the longest but least

intensive phase, lasting up to three years. This stage usually combines 6-MP, MTX, VCR, and corticosteroids (DEX or PRED). Notably, a large part of this phase relies on oral administration, making adherence more difficult to monitor and placing responsibility on the patient and caregivers[253]. Importantly, studies have shown that patients with adherence below 90% during maintenance therapy are up to four times more likely to relapse[259].

This five-phase structure has represented the backbone of ALL therapy for decades. Nevertheless, this does not mean that progress has been limited to optimizing drug combinations. Pediatric ALL has also become a showcase for precision medicine, with several novel approaches already in clinical use. Two critical examples are TKIs and immunotherapy. Before the introduction of TKIs, *BCR::ABL1*-positive ALL was associated with a very poor prognosis, but the incorporation of TKIs into the treatment protocols has drastically improved outcomes[239]. Another major breakthrough has been observed within the field of immunotherapy: blinatumomab (CD19 bispecific T-cell engager), inotuzumab ozogamicin (anti-CD22 antibody-drug conjugate), and CAR-T cell therapy are all approved for clinical use in HR or relapsed patients, with highly promising results[260–262]. Alongside these new modalities, allogeneic hematopoietic stem cell transplantation continues to be an important option for refractory disease[253]. Importantly, the frontiers of treatment research are not limited to the development of novel agents or methods. Another major focus has been on the reduction of therapy intensity in low-risk patients, an effort ongoing for more than two decades, with promising results from multiple clinical trials[263,264]. As a recent example, post-hoc analyses from the Children's Oncology Group AALL0932 trial demonstrated excellent outcomes for low-risk patients receiving one-third of DEX/VCR pulses compared to the previous reference protocol of the group[265]. With the continuous integration of novel biomarkers from multi-omics and single-cell studies, risk stratification is expected to improve further, allowing more patients to receive optimally tailored regimens that minimize long-term cytotoxic side effects.

2.3. Somatic alterations and the rewiring of regulatory architecture

Up to this point, I introduced the fundamental concepts of gene regulatory architecture from a contemporary 3D perspective, and BCP ALL as the disease of interest, both from molecular and translational standpoints. In Section 2.2.2, I discussed genetic predispositions to the disease and underscored their limited penetrance without the accompanying somatic alterations. In the earlier pages of the Introduction, I mentioned how BCP ALL, similar to other pediatric cancers, is characterized by a low mutational burden, and how the limited number of biomarkers restricts the proportion of children

who can benefit from targeted therapies. Probably more times than I should, I have emphasized how the non-coding genome has been overlooked for decades. Altogether, these points create a clear necessity to evaluate somatic alterations in the non-coding genome for their potential functional outcomes, which forms the foundational rationale of this thesis.

In simplest terms, this work explores how the non-coding genome is rewired in BCP ALL, reshaping regulatory architecture to drive leukemogenesis and confer proliferative advantages to leukemia cells. Throughout the Introduction and Background sections, I have highlighted the complexity of this architecture, which partly explains why functional annotation of non-coding variants remains far more challenging than for their protein-coding counterparts. Right before presenting the main findings of my doctoral work, I feel it is appropriate here to provide an overview of the major classes of somatic alterations and how they can rewire gene regulation, supported by brief mechanistic insights and real-world examples. I would also like to emphasize the term regulatory rewiring here. While SVs can certainly cause dosage effects (such as increased transcript output from a gene duplication), these are not the main subject of interest, and thereby not discussed here.

In broad terms, the somatic alterations of interest in this study can be divided into three main categories: SVs — representing larger genomic rearrangements of various kinds —, chromosomal aneuploidies and SNVs. In the molecular classification section, I highlighted how SVs play a crucial role in subtype identification and how the most frequent subtypes are associated with such alterations. Based on their impact on copy number, these events can be further divided into unbalanced and balanced SVs. The former group includes alterations that change the total genomic copy number, such as duplications, deletions, insertions, and unbalanced translocations, while the latter includes variants that do not, such as reciprocal translocations and inversions.

Duplications: As the name suggests, these SVs add genomic material through copying and pasting themselves, which can range from small tandem duplications to large scale copy number gains. Naturally, these alterations are strongly associated with a dosage effect, the extent of which is investigated in **Paper II**. Yet, gene dosage is not the only consequence that can arise from gene duplication, as multiple rewiring patterns that can drastically alter regulatory architecture also exist. For example, in the previous section, I introduced TADs and explained how they are separated from each other by boundaries often surrounded by inversely oriented CTCF binding sites. In cases where a duplication occurs across these boundaries (also known as inter-TAD duplications), it can result in the addition of an adjacent boundary to the original one, which then leads to the formation of a neo-TAD[266]. These neo-TADs, if they include a gene, may have profound effects on its expression, as the entire regulatory landscape of the corresponding locus would be altered. This phenomenon was previously demonstrated

in mice, where the induction of neo-TADs through duplication led to limb malformations[267].

Such TAD-boundary–associated outcomes, on the other hand, represent only a subset of the potential regulatory rewiring mechanisms driven by duplications. In **Paper III**, we observed that only 31% of the duplications lead to neo-TAD formation within our BCP ALL dataset, and this does not mean the rest should be overlooked. As the most basic example, duplication-driven rewiring can also operate through the addition of binding sites for activator proteins, boosting the enhancer-driven transcriptional output of the corresponding genes, which can be viewed as a CRE version of the dosage effect[268]. A well-known leukemia-related example of such an event is the Notch-dependent MYC enhancer, where amplifications are observed in around 5% of T-ALL cases[269]. A similar case in AML occurs at the blood enhancer cluster region, which also regulates MYC, where amplifications were recurrently observed across multiple studies[270–273].

Deletions: From a simple perspective, deletions can be seen as the opposite of duplication events, as they lead to the loss of genomic material. Much like duplications, deletions can exert their effects either by directly causing a dosage reduction of a gene or through regulatory rewiring. Again, similar to duplications, the occurrence of deletions that cross TAD boundaries can lead to significant reorganization of chromatin in 3D space. For example, if a boundary region between two adjacent TADs is completely deleted, this can result in the fusion of these two TADs into a single large neo-TAD, effectively removing the insulation that exists in the wild type[266]. A leukemia-associated example of this mechanism was elegantly demonstrated for the first time in a *Science* paper published in 2016, where CRISPR–Cas9 experiments in HEK cells showed that loss of insulation due to boundary deletions led to the upregulation of two T-ALL proto-oncogenes, *TAL1* and *LMO2*[274]. In BCP ALL, one of the most well-known examples of such a mechanism was published by our lab prior to my arrival, where Yang et al.[275] demonstrated a deletion-mediated enhancer hijacking event leading to the upregulation of *FLT3*. The intra-TAD duplication examples given in the previous paragraph also hold true for deletions from the opposite perspective, as the loss of transcription factor binding sites in enhancer regions can lead to downregulation of the corresponding genes. In **Paper I**, we demonstrated a similar example involving a silencer element, where CRISPR-mediated deletion of this region led to the upregulation of its associated target, *NRAS*.

Insertions: These SVs represent another copy-number–altering mechanism, similar to duplications, yet the copied material does not originate from the same reference point but rather from an external source, such as mobile elements, short tandem repeats, viral or mitochondrial DNA[276]. A common metaphor used to distinguish these two mechanisms is “copy-and-paste” versus “cut-and-paste,” where the latter represents insertions. Inserted elements can cause regulatory rewiring through multiple

mechanisms, such as the addition of an insulator element that can alter the TAD landscape or block enhancer–promoter pairing[277], the introduction of new transcription factor binding sites that can modify modulate CRE activity[278], or even by inserting a CRE itself, such as a novel enhancer or silencer[279]. T-ALL again serves as a good model to exemplify such events, as insertion-induced neo-enhancers regulating oncogenes is an established phenomenon[280].

Translocations: Whether they are balanced or not, translocations are likely the most relevant class of SVs here, as they represent both well-established drivers of BCP ALL, and the most dominant source of enhancer hijacking events[268]. In the molecular classification section, I mentioned a brief history of translocation identifications in leukemia research, highlighting the Philadelphia chromosome, responsible for *BCR::ABL1* generation as the earliest cancer associated discovery. From a mechanistic point of view, this fusion gene also represents a textbook regulatory rewiring mechanism, as the active *BCR* promoter drives the expression of the fused *ABL1*[281]. The most common BCP ALL translocation, *ETV6::RUNX1* represents a similar scenario, where the *ETV6* promoter ends up controlling the *RUNX1* expression after fusion[282]. In T-ALL, t(5;14) leads to the activation of *TLX3* proto-oncogene via enhancer hijacking[283]. Numerous other examples for translocation driven regulatory rewiring can be cited, which is unsurprising, since a translocation event is not only relevant to the breakpoint of interest, but likely displaces an entire regulatory neighborhood from its native context. While the examples I've given so far illustrate local regulatory rewiring, translocations involving transcription factors often have global consequences as well, since the resulting fusion proteins can differ in their DNA-binding preferences and affinities for CREs. A noteworthy example is *KMT2A* rearrangements, where the diverse set of fusion proteins can cause oncogenic alterations on the transcriptional programs through a diverse set of mechanisms[284–287].

Inversions: This copy-number–neutral SVs represent genomic alterations in which a DNA segment is cut and reinserted in the reverse orientation. In the section on architectural proteins, I already discussed an elegant mechanistic example of inversion-induced regulatory rewiring, involving CTCF inversion experiments that demonstrated how such changes can alter 3D genome organization[152]. Since loop extrusion depends on convergently oriented CTCF binding sites, inversions affecting these sites can reorganize chromatin by enlarging existing TADs, fusing adjacent ones, or creating new boundaries when novel convergent motifs emerge[266]. A related example was also described in the transcription factors section with the interferon-β enhancer, highlighting how changes in the order of transcription factor binding sites can shift transcriptional output[144]. In regions where motif grammar is a key determinant of function, inversions can therefore cause pronounced changes in regulatory activity. In simpler cases, inversions may also lead to enhancer hijacking through CRE repositioning, for instance when a regulatory element at one end of the inverted

segment becomes rewired to a new target. A clear example of this is seen in AML cases with inv(3)(q21;q26), where the *GATA2* enhancer is repositioned to regulate *EVI1*, leading to upregulation of this proto-oncogene[288].

SNVs: A large portion of this work focuses on SNVs, particularly in **Paper I** and **Paper IV**. It is not possible to cover all potential mechanisms through which SNVs can affect regulatory wiring, but it would not be wrong to say that most of the functional outcomes described for the major SV subclasses above can also emerge from single-nucleotide alterations. For example, SNV-induced changes in CTCF binding affinity may influence TAD organization, leading to outcomes such as increased boundary strength, TAD fusion, or neo-TAD formation, depending on the genomic context. This possibility of functional convergence through various mechanisms highlights the need for detailed and context-aware evaluation of somatic events. Similarly, within E-P pairs, mutations altering transcription factor motifs can change binding patterns and ultimately transcriptional output. The most well-known CRE-associated example is perhaps the recurrent SNVs occurring in the *TERT* promoter region, which create additional ETS binding sites associated with gene upregulation[32,33]. In our own work (**Paper I**), we uncovered a similar mechanism, demonstrating how recurrent SNVs found in a hotspot region within the intron of *TSPAN2* altered CRE-driven reporter activity using luciferase assays. The functional relevance of SNVs occurring in introns is not limited to CRE-based alterations either, as evidence shows that intronic variation can also affect splicing[289,290]. Moreover, in **Paper IV**, we show how mutagenesis rates and patterns of SNVs are closely related to 3D genome organization, highlighting their potential roles in shaping and maintaining the overall regulatory architecture of the genome.

A summary of the well-established mechanisms discussed above is presented in **Figure 5** for ease of visualization. Of course, these schematics do not encompass the entire spectrum of somatic variation classes or the diverse ways in which they rewire regulatory architecture. More complex events—such as chromothripsis, chromoplexy, or break–fusion–bridge cycles—can also lead to profound structural rearrangements, exemplified in leukemia by the formation of iAMP21[248]. In contrast, simpler genetic changes affecting a single gene, such as *TET2* loss-of-function mutations in AML, can likewise exert broad regulatory consequences, in this specific case by driving global enhancer hypermethylation[291]. This is not surprising, as it is already known from perturbation experiments that genes central to 3D genome organization, such as the cohesin complex members *STAG2* and *RAD21*, can single-handedly alter global chromatin topology[153,292]. Altogether, this summary underscores the vast landscape of functional possibilities generated by somatic alterations across the non-coding genome—an area to which this thesis hopefully contributes a small but meaningful piece of understanding.

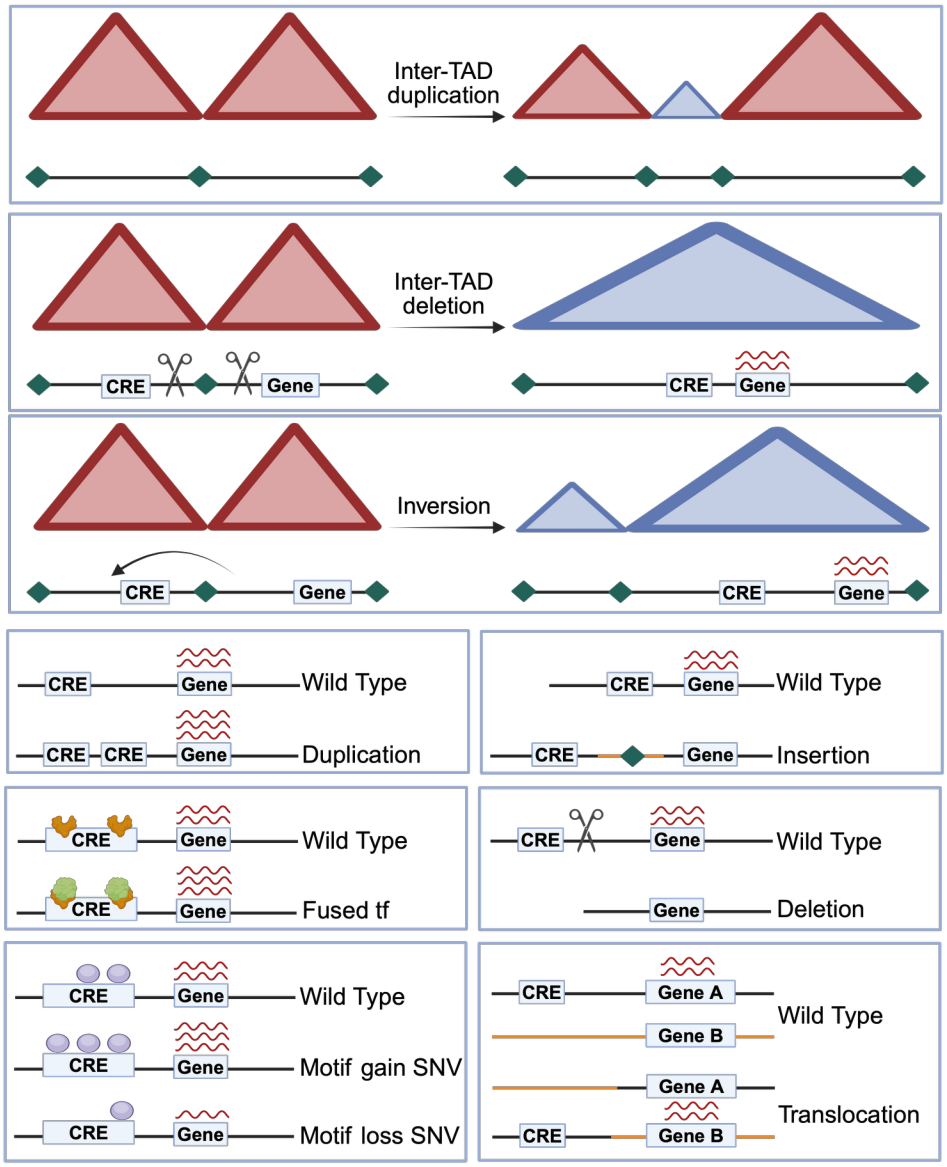


Figure 5 | Summary of potential regulatory rewiring mechanisms driven by somatic events. This schematic illustrates a range of possible rewiring mechanisms occurring in different genomic contexts. The first three panels, shown in larger rectangles, represent structural variations spanning topologically associating domain (TAD) boundaries, while the smaller panels depict intra-TAD events. Red triangles indicate wild-type TADs, whereas blue triangles denote neo-TADs. Green diamond symbols at boundaries represent CTCF binding sites. Differences in gene expression are visualized by the number of red mRNA transcripts shown above the corresponding genes. Abbreviations: CRE : cis-regulatory element, tf : transcription factor, TAD : topologically associating domain, SNV : single nucleotide variation

3. The present study

3.1. Methods

The bulk of the output in this thesis is the product of computational work, based on integrated data analysis. Naturally, such computational work requires the implementation of dozens of algorithms, which are not described in detail here, as they are adequately referenced in the corresponding papers. Similarly, software generated by me and my colleagues can be found in the associated repositories with full documentation: https://github.com/biomedswe/BASE (**Paper II**), https://github.com/efeaydin-bio/microc-all-db (**Paper III**).

Nevertheless, the big data processed in this study was derived from a variety of wet-lab methods. Therefore, before diving into the results and discussion, I believe that a conceptual introduction to the applied methods is of immense importance, as they are directly connected to how we approached the research questions throughout the work. In this section, I briefly introduce the main methods used across the four papers.

WGS

For all mutation-centric analyses presented in this thesis—including identification of SNVs, SVs, hotspots, and single base substitution (SBS) signatures—we relied on WGS as the main technology. WGS involves sequencing the entire genome of a sample and aligning it to the reference genome of the corresponding species. In broad terms, the process includes sample preparation, sequencing, quality control, and data analysis[293]. In this work, I analyzed WGS data derived from bone marrow samples of pediatric BCP ALL cases.

Three cohorts were used in this study: the Lund University cohort, the St. Jude cohort, and the Therapeutically Applicable Research to Generate Effective Treatments (TARGET) initiative. For the Lund and St. Jude cohorts, which were sequenced on Illumina platforms, we aligned the FASTQ reads to the reference genome using BWA, and identified SNVs with Mutect, Mutect2, and MUSE[294–296]. For SV detection, Manta and GRIDSS2 were used.[297,298]. For the TARGET cohort, which was generated on the Complete Genomics platform, we used the official TARGET WGS analysis pipeline (https://www.cancer.gov/ccg/research/genome-sequencing/target).

The annotated variants that passed post-filtering, as described in the corresponding publications, provided the foundation for the downstream analyses in **Paper I**, **Paper III**, and **Paper IV**, all of which focused on somatic events**. Paper II**, in contrast, included an ASE analysis, which required germline variant calls; for this, we used GATK HaplotypeCaller[299]. In **Paper IV**, the patterns of mutagenesis events for SNVs were of particular interest, and therefore we also analyzed the proportions of known COSMIC mutational signatures in our data using the SigProfilerAssignment tool[300].

SNP arrays

Although many of its applications—particularly in de novo variant discovery—have been overtaken by NGS, SNP arrays remain relevant in numerous fields, including BCP ALL research. Owing to their availability as a high-throughput technology for nearly three decades, a substantial body of data has accumulated from SNP arrays. In **Papers II and III**, we leveraged this abundance as a complementary layer of analysis.

Approximately 99% of the human genome is identical between individuals, but the remaining variation includes polymorphisms such as SNPs[301]. SNP arrays exploit this feature to assess genomic variation at scale. While arrays from different manufacturers differ in chemistry and computational pipelines, the underlying principle is the same. Briefly, hundreds of thousands of allele-specific hybridization probes target SNPs of interest. Hybridization of DNA to these probes generates signals corresponding to possible genotypes (AA, AB, BB). These raw intensity signals are normalized against background noise and processed to infer genotypes and, in many applications, CNVs.

In **Paper II**, we used SNP array data from The Cancer Cell Line Encyclopedia (CCLE), and a built in reference panel to derive corresponding copy number profiles, ensuring the robustness of our own analytical pipeline[302]. In **Paper III**, one of our objectives was to evaluate the potential of Micro-C to detect SVs using the EagleC algorithm[303]. To validate the robustness of this pipeline, we used previously published datasets, some of which (n=32) were generated by SNP arrays.

RNA-seq

Since its introduction in 2008, RNA-seq has dominated the field of gene expression studies and it is also one of the most frequent sources of data used in this doctoral work. Although numerous derivative methods have emerged from RNA-seq, the vast majority of available datasets—including those in this thesis—are short-read libraries sequenced on Illumina platforms. In broad steps, RNA-seq workflows include RNA extraction, reverse transcription to generate cDNA, and library preparation by fragmentation, adapter ligation, and polymerase chain reaction (PCR), followed by quality control and sequencing[304]. Raw data, which typically comprise ~20–30 million reads per sample in bulk RNA-seq, are analyzed with computational pipelines described in the papers included in the appendix. In brief, the analytical workflow involves aligning reads to a

reference transcriptome, normalizing counts to account for library size/composition effects, and measuring differences between the conditions of interest.

Although the newer derivative of bulk RNA-seq, which allows researchers to evaluate transcriptional outputs on a single-cell level (single-cell RNA-seq), provides valuable insights on the intra-sample heterogeneity of the disease that are largely masked on bulk data, the traditional bulk RNA-seq is still a major method to extract systematic patterns in between different conditions. In leukemia research, bulk RNA-seq can provide a molecular classification of the disease[28]. We used this feature of the method in **Paper I,** to ensure the correctness of our normalization pipeline, where we showed that despite the fact that we used samples from different cohorts, they cluster together based on their molecular subtypes and not their batch.

For the mutation-centric work that carries significant weight in this thesis, bulk RNA-seq provides a direct readout of the potentially affected genes. For example, **in Paper I**, after identifying a mutational hotspot interacting with the *NRAS* promoter, we compared RNA-seq profiles of mutant samples against wild-type. The expression pattern observed in RNA-seq was subsequently confirmed by other assays. In **Paper II**, to assess copy-number effects on gene expression, we again relied on RNA-seq paired with WGS. In **Paper III**, by pairing RNA-seq with Micro-C, we showed how closely regulatory transcriptomics and 3D genome architecture are connected: we observed effects spanning from 500-kb–scale compartmentalization down to 1-kb–scale looping, modulated by specific histone marks, and reflected in RNA-seq profiles. Moreover, we demonstrated a significant association between promoter interaction detection and gene-expression levels by jointly analyzing these modalities.

ChIP-seq

Throughout this thesis, I highlight how gene regulation at the DNA and protein levels is inter-connected and how information flows bidirectionally within this complex architecture occupying 3D space. By integrating multi-omics data, I aimed to account for multiple layers of regulatory architecture and model leukemia biology as faithfully as possible. ChIP-seq, as one of the methods I repeatedly relied on, directly serves this purpose by connecting the worlds of DNA and proteins in a singular experiment and allowing their interactome to be profiled genome-wide.

ChIP-seq maps the genomic regions bound by a specific protein of interest. The basic workflow includes; the crosslinking of protein of interest to the DNA using formaldehyde, fragmentation of the chromatin, immunoprecipitation of DNA-Protein pairs of interest using specific antibodies, reverse crosslinking, library preparation and sequencing[305]. Raw reads are analyzed with computational pipelines, some-of-which can be found within the papers in the appendix.

A significant portion of knowledge regarding the biochemical features of CREs come from ChIP-seq experiments, which makes ChIP-seq one of the most important methods in the field of epigenetics. Naturally, it also plays a big role in throughout this work. In **Paper I**, we investigated the presence of histone markers in mutational hotspots and showed their correlation with the differential expression of genes nearby these hotspots. In **Paper III**, we demonstrate the robustness of our Micro-C loop calling pipeline by showing the association between putative E-P pairs and canonical histone peaks derived from primary patient samples. Moreover, we provided novel insights on the 3D genome architecture of BCP ALL on multiple levels by investigating the architectural proteins found on TADs and presence of transcription factors found in 1kb loops, each of which were derived from ChIP-seq experiments. In **Paper IV**, where we discuss the possibilities of CTCF-independent TAD formation and its putative effects on mutational profiles, we leveraged ENCODE ChIP-seq to evaluate CTCF binding in regions of interest.

ATAC-seq

First introduced in 2013, ATAC-seq is an epigenomic mapping approach that enables genome-wide investigation of open chromatin in cells[306]. The method benefits from Tn5, a prokaryotic transposase, which uses a cut-and-paste mechanism, allowing integration of pre-loaded sequencing adapters to the targeted regions[306]. These cut fragments with pasted adapters are then sequenced, similar to other NGS methods described above. Closed chromatin provides a biochemical hindrance for this cut-and-paste mechanism to attach, thereby, the vast majority of sequenced fragments represent regions of open chromatin. This provides a powerful readout for researchers with broad potential discovery, as open chromatin is often a pre-requisite for transcription factor binding and regulatory interactions. Moreover, the nucleotide-level resolution enables footprinting-based inference of specific protein binding, which has drawn substantial attention in leukemia research, where lineage-specific transcription factors are major drivers of the continuous process of hematopoiesis.

Throughout this thesis, I repeatedly benefited from a previously generated ATAC-seq dataset comprising 156 primary bone marrow samples from diagnostic BCP ALL cases[307]. In **Paper I**, these data played a role analogous to histone marks, as we evaluated the presence of ATAC-seq peaks at mutational hotspots. In **Paper III**, we identified >55,000 chromatin loops at 1-kb resolution, of which >95% had an ATAC-seq peak at least at one anchor, underscoring the robustness of our experimental and computational pipeline. ATAC-seq was also essential for compartment annotation: after assigning compartments from chromatin interaction eigenvectors, we used ATAC-seq profiles from a previously generated REH cell line dataset and from our HeH samples (data not shown) to label A vs B compartments according to their accessibility patterns.

Hi-C/Micro-C

As emphasized throughout this thesis, my central focus has been to understand the roles of the 98% of DNA that has long been overlooked. The methods described earlier have provided valuable insights toward this goal by delineating epigenomic patterns linked to gene expression. Yet, these approaches remain, at their core, indirect inferences. As noted in the background section, no single epigenetic mark is sufficient to comprehensively annotate a regulatory locus.

It is important to remember that DNA—despite the many philosophical meanings attached to it—is still a polymer, shaped by the laws of physics acting on its biochemistry within three-dimensional space. In other words, the epigenetic marks we use to study DNA are inseparably tied to 3D organization, and regulatory architecture derives its meaning through its behavior in this spatial context. Thus, studying chromatin organization in its spatial framework is essential, and represents a perspective that forms the central theme of this work.

To achieve this, we primarily relied on the state-of-the art high-throughput chromosome conformation capture method Micro-C, and to a lesser extent, its predecessor Hi-C. The conceptual foundation of these methods dates back to 2002, when Dekker et al. introduced the original 3C assay in yeast[308]. This approach combined chromatin crosslinking, restriction digestion, ligation, reverse crosslinking, and quantitative PCR (qPCR) to quantify interactions between a specified pair of loci. This method was subsequently improved to one-to-all (4C), many-to-many (5C) and after the establishment of NGS technologies, all-to-all (Hi-C, Micro-C), enabling unbiased, genome-wide profiling of chromatin interactions[309]. The transformative impact of the 2009 Hi-C paper[165] has already been discussed in detail in the background section.

Methodologically, Hi-C involves crosslinking chromatin with formaldehyde, cleaving DNA with a restriction enzyme, filling in 5' overhangs with biotin, ligating DNA fragments, reversing crosslinks, sonication, and enrichment of ligation junctions for sequencing[165]. The original study used HindIII, whereas later studies employed MboI or DpnII to achieve more frequent cuts and thereby higher resolution[151,165,310]. However, one can see the obvious drawback here. Restriction enzymes by nature, are dependent on the sequence features of the DNA, which introduce biases and reduce uniformity. Micro-C, the newer derivative of Hi-C which was introduced in 2015, addressed this problem by modifying the protocol with the inclusion of an MNase, which cleaves linker DNA between nucleosomes, thereby producing a less biased /more uniform output[311].

In our own work, Micro-C allowed us to reach 1 kb resolution in **Paper III**, making it possible to investigate chromatin interactions at the level of CREs. In **Paper IV**, we further leveraged this resolution to demonstrate protective features of different classes

of chromatin loops. In **Paper I**, Micro-C was used in REH cells to generate an interaction matrix around our regions of interest, while publicly available promoter-capture Hi-C datasets of GM12878 were incorporated as supportive references[312]. For **Paper III**, we also integrated Hi-C data from peripheral blood mononuclear cells (PBMCs) and Hematopoietic stem and progenitor cells (HSPCs) samples, which provided additional context for interpreting the 3D genome organization of leukemia in relation to hematopoietic differentiation[313].

Dual luciferase assays

Since their introduction in the late 1980s, luciferase reporter assays have become a mainstay for studying gene regulation, initially using firefly luciferase as the readout[314]. In 1996, the dual-luciferase format was developed by Promega, in which a second reporter (typically Renilla luciferase) is co-transfected and measured sequentially to provide an internal normalization control. This simple but powerful modification corrects for technical variation and markedly improves reproducibility. Owing to their sensitivity and ease of use, dual-luciferase assays remain widely used today. Moreover, scalable derivatives have also emerged, allowing researchers to study regulatory elements from a high-throughput level[56].

In principle, dual-luciferase assays involve transfecting cells of interest with an experimental firefly-luciferase construct and a control Renilla-luciferase plasmid, then quantifying normalized activity (Firefly/Renilla) from cell lysates according to the experimental design. In **Paper I**, we were interested in a genomic region of 143 base-pairs within the intronic region of *TSPAN2*, where recurrent mutations occurred. We asked whether this region functions as a regulatory element and whether the mutations disrupt that function. We constructed three firefly reporters—wild-type insert, mutant insert, and empty vector—and co-transfected each with a CMV-driven Renilla control into REH cells. The wild-type construct produced a significantly higher normalized signal than the empty vector, indicating regulatory activity. Introducing the mutations significantly reduced reporter activity relative to the wild type, supporting a loss-of-function effect on this putative regulatory element.

It is important to highlight here that episomal reporter assays do not fully recapitulate native chromatin complexity. Therefore, one should exercise great caution in using these assays as a stand-alone method. Accordingly, we corroborated our reporter results with orthogonal data. Nevertheless, our readouts indicate that certain transcription factors present in the REH cell line actually bind to the wild-type enhancer region to increase reporter gene output. Moreover, introduced mutations likely disrupts these binding or leads to alternative formations. Altogether, reporter assays can still be relevant in CRE research, provided that the results are not overinterpreted and are supported by additional layers of information.

CRISPR-Cas9

Ever since the rediscovery of Mendel's work launched the field of genetics, the idea of manipulating the blueprint of life has been a central theme in both science and society. Few scientific pursuits have been so vividly represented outside academia, as evidenced by numerous cultural products, from Huxley's Brave New World to cinematic portrayals such as Jurassic Park and Gattaca. Yet until the 2010s, the ability to edit genes remained confined to a limited number of laboratories, relying on methods that were error-prone, costly, and time-consuming. This landscape changed dramatically with the introduction of CRISPR–Cas9, which transformed what was once reserved for a small group of scientists (or philosophical debates) into a practical molecular biology method, now routinely performed all around the world[315].

Originating from an adaptive immune system in bacteria, CRISPR-Cas9 systems are broadly composed of two main parts. The Cas9 protein, which is responsible for generating double strand breaks, and the tracrRNA:crRNA duplex, adding the site-specificity feature to the mechanism by interacting with target DNA based on Watson-Crick base-pairing[315]. For experimental convenience, this duplex can be synthesized as a single-guide RNA (sgRNA). Several transfection and transduction methods have been developed to deliver these components into cells, with varying degrees of efficiency depending on the experimental setting and cell type. In this thesis (**Paper I**), we employed a ribonucleoprotein (RNP)-based approach combined with electroporation. Specifically, we designed sgRNAs flanking the region to be deleted, assembled them with Cas9 protein in vitro, and introduced the resulting RNP complexes into REH cells by electroporation.

Despite the transformative impact of CRISPR–Cas9 on gene editing, the method is not flawless. Not all cells take up the construct during transfection, and even in successfully transfected cells, editing efficiency can vary. To enrich for edited cells and improve downstream analyses, we used a fluorescently tagged Cas9 to identify transfected populations, which were then sorted and assessed for cutting efficiency by qPCR.

Fluorescence activated cell sorting

In the background section, I highlighted the historical significance of flow cytometry and its transformative role in identifying cell surface markers, which fundamentally advanced our understanding of hematopoiesis. More than three decades have passed since the first prospective purification of mouse HSCs using fluorescence-activated cell sorting (FACS), yet the technique remains a cornerstone of hematology research and is routinely employed by researchers worldwide today.

In essence, FACS exploits cellular heterogeneity by distinguishing and isolating cells based on specific surface markers. Fluorescently conjugated antibodies are designed to

bind target proteins, and upon excitation, the emitted signals are detected and used to sort cells within a heterogeneous population according to defined marker combinations[316].

In **Paper I**, we applied FACS to enrich for transfected REH cells, where one sgRNA carried a GFP reporter and the other an RFP reporter. Because our strategy required simultaneous targeting of both flanking sites to generate a deletion, we specifically sorted for live cells expressing both GFP and RFP. In **Paper III**, we used FACS to isolate healthy B-cell progenitors from cord blood as controls, sorting based on expression of CD34, CD38, CD10, CD19, and CD20.

qPCR

One of the foundational methods developed in the 20th century within the field of molecular biology is, without doubt, PCR, which led the inventor Kary Mullis to receive a Nobel Prize in Chemistry in 1993[317]. Within a decade after the invention, an update on the method, real-time PCR, was developed, which allows users to amplify DNA material through thermal cycling steps involving denaturation, annealing, and elongation, while simultaneously quantifying (hence real-time) the product at every cycle based on the fluorescence signal emitted from the product[318]. In the early cycles, since there are limited copies, the signal is hard to detect. This exponentially grows until the product reaches a saturation point, eventually creating the characteristic sigmoid curve. The differences in the cycles where exponential growth is reached can be used as markers to quantitatively calculate the relative abundances between two conditions (relative quantitation), or the amount of material can be calculated with high sensitivity by comparing the experimental output against a known reference (absolute quantification), which is clinically relevant in certain diseases associated with agents such as hepatitis B or HIV[319,320].

For this work, we relied on qPCR only in **Paper I** for two occasions. First, we wanted to evaluate the cutting efficiency of our sgRNA–Cas9 complex. To do this, we designed forward and reverse primers targeting our region of interest, extracted the DNA of the transfected cells with the test and control complexes, and performed qPCR to compare the copy numbers against the wild type cells, which confirmed our pipeline's efficiency. After validating that our CRISPR design worked, we wanted to evaluate the putative effects of this deletion on our gene of interest, which we hypothesized to be targeted by this regulatory region. For this, we again relied on qPCR to compare the RNA abundances of *NRAS*, our gene of interest, normalized by the abundance of housekeeping gene *GUSB*. This demonstrated that our deletion caused upregulation of *NRAS*, consistent with the repressor function we had hypothesized based on the high-throughput data.

3.2. Results and discussion

Here, I summarize the outputs of the four papers included in this thesis by outlining the main results, explaining the rationale behind the study designs, and discussing the broader implications of the findings.

3.2.1. Paper I

Discovery of cis-regulatory mechanisms via non-coding mutations in acute lymphoblastic leukemia[321]

Main results:

- Subtype-specific hotspots are enriched near COSMIC Cancer Gene Census genes.
- Hotspots with differentially expressed genes nearby show greater enrichment for enhancer-associated marks than those without.
- A recurrently mutated cis-regulatory region ~368 kb from *NRAS* acts as a negative regulator.
- Beyond epigenomic annotations, recurrent non-coding variation itself can be integrated as a discovery signal for regulatory function.

In 2013, two back-to-back papers were published that provided important insights into the previously overlooked roles of non-coding regions in cancer[32,33]. These studies identified recurrent mutations in the promoter of *TERT*, which dysregulated transcriptional output in melanoma. Similar variations were soon confirmed in other tumors, firmly establishing the principle that non-coding alterations can act as bona-fide driver events in cancer[322,323].

This breakthrough spurred systematic efforts to analyze non-coding variants at a genome-wide scale. Although regulatory mutations continue to be discovered, large cohort studies have consistently reported a relative depletion of recurrent non-coding drivers compared to their protein-coding counterparts[324,325]. This observation is perhaps connected to our limited understanding of these regions with unique, heterogeneous properties and diverse elements with different mechanisms of action, thereby affecting the detection accuracy of available pipelines[289,326]. Furthermore, studies have shown that enhancer redundancy provides a protective layer against the phenotypic effects of loss-of-function mutations, ensuring robustness of transcriptional programs[327]. From another perspective, this feature indirectly indicates that variants occurring at different locations may converge on a similar functional output, which could be missed in the absence of relevant multi-omics data. Another reason for this depletion could be related

to the dynamic, fine-tuning features of CREs in regulating transcriptional programs, differing from the coding regions, where variations may have larger and more direct phenotypic impact[328,329]. On the other hand, functional studies and therapeutic interventions — for example, the use of BET inhibitors in acute myeloid leukemia (AML), which suppress oncogenic transcription by disrupting enhancer activity — clearly demonstrate that non-coding elements can nonetheless represent essential cancer dependencies[330,331].

In this paper, we set out to investigate this apparent dichotomy under the hypothesis that the observed depletion of non-coding drivers may in part reflect limitations of the traditional, linear definition of mutational hotspots, which relies solely on WGS data. Instead, we developed a systematic framework that treats mutational events as outputs of an integrated epigenomic context rather than as isolated DNA sequence changes. Specifically, we analyzed variant data from 345 WGS, representing primary BCP ALL cases across two cohorts, and integrated these with matched RNA-seq data as well as epigenetic profiles derived from primary samples through publicly available ChIP-seq and ATAC-seq datasets. We then further examined our findings from a functional perspective, using Micro-C, CRISPR and dual-luciferase assays.

BCP ALL is an especially compelling context for such an analysis. First, it is the most common pediatric cancer and typically has a relatively low mutational burden, which can make recurrence-based detection of drivers challenging. Second, the repertoire of established genetic drivers is limited, with many involving structural variation rather than point mutation, leaving gaps in our understanding of small non-coding contributions. Finally, treatment outcomes have plateaued, and further chemotherapy intensification is not feasible, underscoring the need for precision-medicine approaches. Together, these features make BCP ALL an ideal model in which to explore the contribution of non-coding variants, where every additional layer of insight may hold clinical and biological importance.

From a genome-only perspective, our results were consistent with previous systematic studies: mutational hotspots did not show enrichment at cancer-associated loci. However, when we focused on hotspots that were linked to differential expression of nearby genes, we observed clear enrichment of regulatory features, including H3K4me1, H3K27ac, and chromatin accessibility (ATAC-seq). Notably, many of these filtered regions were also annotated as strong enhancers in the GM12878 cell line according to the Enhancer Atlas. These findings underscore the close interconnection between epigenetic layers and DNA in gene regulation, and demonstrate how even a simple integration of additional regulatory data can reveal biological significance that remains hidden in genome-only analyses.

Another interesting finding emerged from the subtype-specific distribution of these non-coding variants. It is well established that BCP ALL subtypes follow distinctive

transcriptional programs that can be readily identified from RNA-seq profiles, which was also evident in our dataset (**Figure 4**). We found that several non-coding hotspots were enriched within specific subtypes, and these subtype-enriched hotspots were statistically more likely to be proximal to genes in the COSMIC Cancer Gene Census (Odds ratio = 1.81, $P = 0.038$, Fisher's exact test). These findings highlight subtype-specific heterogeneity in BCP ALL as an additional layer of complexity that can mask the impact of non-coding mutations from a WGS-only perspective.

As a proof-of-concept application, we prioritized a recurrently mutated hotspot in chromosome 1 that was associated with *NRAS* upregulation within the *BCR::ABL1-like* subgroup ($P = 0.021$, Two-tailed Mann–Whitney U-test, **Figure 6A**). First, to assess whether these mutations perturb cis-regulatory function, we performed dual-luciferase reporter assays, and found out that the wild-type sequence indeed exhibited regulatory activity, which was markedly disrupted by the patient-observed mutations (**Paper I, Figure 6B**). Next, we evaluated causality by CRISPR-mediated deletion of the locus; deletion led to significant *NRAS* upregulation ($P = 0.0286$, Two-tailed Mann-Whitney U-test, **Figure 6C**), consistent with a repressive function of the wild-type element and with GM12878 genome annotations of Roadmap Epigenomic Consortium's 15-state chromatin model[332]. Finally, Micro-C mapping revealed physical interactions between this hotspot and *NRAS* regulatory elements. Together, these results show that WGS variant data can be used to nominate and functionally validate regulatory elements of cancer-associated genes when integrated with multilayer epigenomic context.

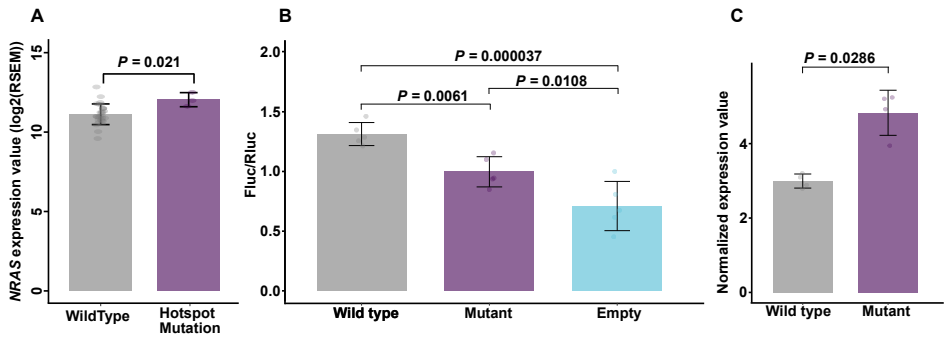


Figure 6 | Functional evaluation of the non-coding hotspot mutation. (A) RNA-seq analysis shows increased *NRAS* expression in cases harboring the putative regulatory mutation. (B) Dual-luciferase assays demonstrate regulatory activity in the region of interest compared to the empty vector, with activity significantly reduced upon introduction of the mutation. (C) CRISPR-Cas9–mediated deletion of the region leads to a significant increase in *NRAS* expression, indicating a repressive cis-regulatory role of the recurrently mutated region.

3.2.2. Paper II

Characterizing the allele-specific gene expression landscape in high hyperdiploid acute lymphoblastic leukemia with BASE[333]

Main results:

- BASE, an open-source computational tool that co-evaluates allele-specific copy number (asCN) with RNA-seq to identify ASE profiles in aneuploid tumors, was developed and applied to HeH ALL.
- The ASE landscape of HeH ALL is primarily driven by dosage effects from copy number gains.

In **Paper I**, we demonstrated how cis-regulatory regions, their gene targets, and functionality can be inferred from non-coding variants through integrated multi-omics analyses. Our results indicated the existence of potentially overlooked non-coding roles that remain hidden when relying on WGS data alone. Nevertheless, WGS still served as our starting point to identify mutational hotspots, and even by itself it represents a complex layer of information of biological relevance. One important aspect of this information lies in the distribution of variant allele frequency (VAF). In our dataset, the vast majority of variants passing filtering steps appeared to be heterozygous, with a median VAF of 0.3946 and a third quartile of 0.4750. While we acknowledge the possible influence of tumor purity and subclonality, this pattern is not unexpected. Indeed, genome-wide studies across tissues consistently show that somatic mutations are heterozygous events in the great majority of cases[334].

Naturally, this established phenomenon has implications from an allele-specific point of view. Heterozygous disruptions in CREs should have allele-specific consequences from a transcriptional point of view, which can be captured through RNA-seq. This is particularly relevant in cancer research, as pan-cancer studies have shown that roughly half of driver mutations are associated with allelic imbalance, with tumor cells often providing positive selection to oncogenic alleles through multiple routes[335]. One major route involves CNVs, which directly alter gene dosage. CNVs are of special importance in ALL, since the most frequent subtype—HeH ALL, accounting for 25-30% of cases—is defined by recurrent chromosomal gains[235]. In addition, many ALL drivers arise from structural variations, such as gene fusions, as discussed throughout this thesis. Thus, ASE and CNVs are inherently intertwined with the disease. However, the connection between ASE, CNVs, and ALL has not been fully clarified. How does the ASE landscape in ALL look like, and to what extent it is shaped by the well-characterized CNVs?

To answer these questions, one must first extract the ASE profile of the sample set. Several algorithms have been developed for this purpose[336–338]. Despite differences in implementation, such as modeling variance or accounting for technical biases, they all share a common foundation: ASE is inferred by comparing RNA-seq read counts between heterozygous alleles. Although this principle is logically sound and broadly applicable, it carries an obvious limitation in a tumor type like HeH ALL, where CNVs are pervasive across the genome. Regardless of how well-designed the analytical framework may be, ASE estimates that ignore the underlying genetic background risk systematic error. In this study, we addressed this challenge by developing a computational tool that integrates asCN states with ASE outputs, using paired WGS and RNA-seq data. We further demonstrated its utility in HeH ALL as a proof of concept.

Our tool, called BASE, independently analyzses paired WGS and RNA-seq data for CNV and unnormalized ASE counts respectively and then jointly analyzes these outputs to derive copy number adjusted ASE profiles (**Figure** 7). Beyond the built-in quality control steps of the algorithm, we evaluated the accuracy of each module against independent reference datasets. For asCN states, we analyzed 15 well-characterized cancer cell lines and compared our calls to those of CCLE, achieving recall rates of 98–100%. For ASE, we first tested phased versus unphased inputs, as well as expression-based normalization, and found that neither factor substantially altered the results[302]. As an external benchmark, we then compared ASE calls from BASE to those from two established pipelines, aScan[337] and MBASED[336], observing high concordance with Spearman's correlations of 0.93 and 0.99, respectively. Notably, in this comparison, phased inputs generated more accurate outputs. In summary, both asCN and ASE modules were independently validated against established methods before being jointly integrated to generate adjusted ASE profiles.

After validating the robustness of our computational framework, we applied it to a primary dataset of 12 matched WGS/RNA-seq cases of HeH ALL, aiming to characterize the allele-specific landscape of the most common pediatric ALL subtype in the context of CNVs. Our RNA-seq analysis identified 8,210 expressed genes, of which 44.5% also showed ASE. However, the majority of these ASE events (64.9%) were unique to a single sample. To determine how much CNVs influence this imbalance, we focused on autosomal chromosomes with recurrent gains (4, 6, 10, 14, 17, 18, 21). On these chromosomes, nearly a quarter of informative genes (22.8%) displayed recurrent ASE, compared to only 4.6% on the remaining chromosomes, indicating that copy-number gains are the predominant driver of ASE in HeH ALL. At the same time, we also identified CNV-independent ASE in some leukemia-associated genes, suggesting additional regulatory mechanisms. For example, ASE of *KMT2C* was observed in 3 cases (25%) with disomy 7. To further explore other mechanisms, we integrated somatic mutation profiles from paired remission WGS data. We highlight

two examples: (i) *FLT3*, where a promoter-proximal deletion caused cis-dysregulation and ASE favoring the mutant allele; and (ii) *WEE1*, where a nonsense mutation triggered degradation of the mutant transcript, leading to ASE of the wild-type allele.

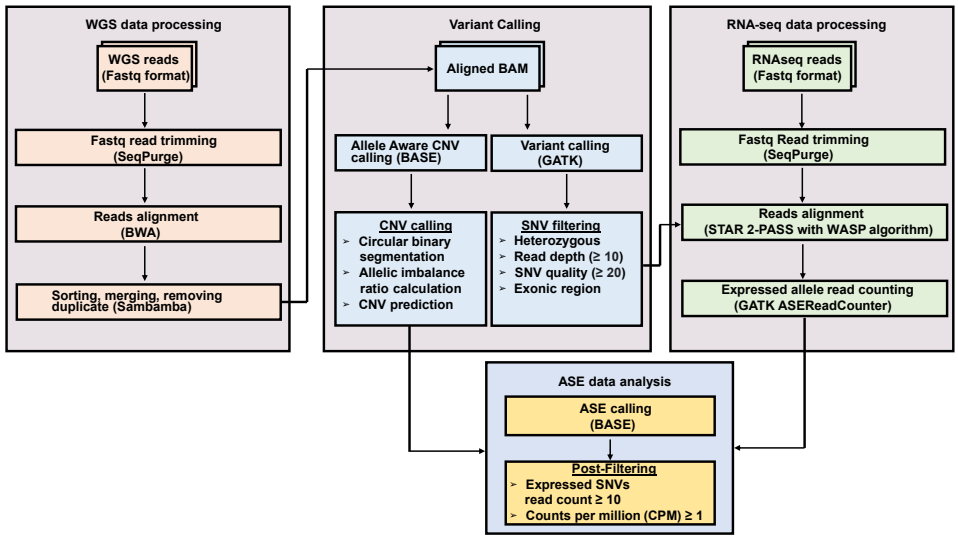


Figure 7 | BASE Workflow. A flowchart of the computational pipeline that integrates WGS and RNA-seq data to extract copy-number adjusted allele-specific expression profiles.

In summary, we developed a robust tool to delineate the diverse paths leading to allelic imbalance by integrating ASE with allele-specific copy number information. Using HeH ALL as a proof of concept, we mapped the allele-specific landscape of this leukemia subtype and showed to what extent ASE is driven by CNVs versus other mechanisms. Beyond ALL, our tool can help uncover oncogenic events that manifest as allelic imbalances in other diseases, with particular relevance for aneuploidy-driven cancers. The software is available at: https://github.com/biomedswe/BASE.

3.2.3. Paper III

The 3D genome of pediatric B-cell precursor acute lymphoblastic leukemia[339]

Main results:

- We generated the largest (to our knowledge) primary-sample 3D genome mapping dataset, using Micro-C on 35 primary BCP ALL samples with high resolution.
- Somatic events reshape the 3D genome of BCP ALL at all three layers; A/B compartments, TADs, and loops (**Figure 8**).
- TAD boundary strength can be used as a subtype classifier.
- We identified >25,000 chromatin loops anchored at regulatory elements, potentially influencing >10,000 protein-coding genes.
- Genes with alternative promoter interactions are enriched in the COSMIC Cancer Gene Census.
- Differential loop analyses reveal 3D rewiring of oncogene regulation even in the absence of somatic coding events.
- We developed a user-friendly web application to facilitate further research.

In **Section 2.1.3**, the invention of Hi-C in 2009 and its role in opening a new avenue for studying three-dimensional chromatin organization at a genome-wide scale has already been highlighted[165]. Through subsequent genome-wide studies, our current conceptual framework of A/B compartments, TADs, and chromatin loops has been gradually established, along with the underlying mechanistic principles such as loop extrusion and compartmental segregation. By today, chromosome conformation technologies have advanced substantially compared to the early 2010s, both in terms of resolution and accuracy. Notably, only weeks before I began writing this section, a *Nature* study reported an enhanced Micro-C method capable of achieving base-pair resolution (~10 bp) in *Escherichia coli*, revealing structural features that would remain hidden at coarser resolutions[340]. A somewhat similar observation on the mammalian side was published earlier, following the implementation of Region Capture Micro-C in mouse embryonic stem cells, which demonstrated the existence of nested microcompartments[341]. Such studies, among many others, highlight the field of 3D genomics as a rapidly advancing area of research, with many potential breakthroughs likely in the near-future.

The advent of genome-wide 3D chromatin technologies, together with the discovery that non-coding variants play critical roles in cancer, has transformed the field from basic studies in cell lines into a translational discipline with direct relevance to oncology. Leveraging these advances, numerous studies have now mapped the 3D genomic

landscape of cancers such as T-ALL, prostate cancer, colon cancer, and AML, often with cohorts of more than ten primary patient samples, underscoring the growing potential of 3D genomics to reveal disease mechanisms and therapeutic opportunities[313,342–344].

When it comes to BCP ALL, considering both the prevalence of the disease and the saturation of conventional chemotherapy approaches, 3D genome studies hold great promise, and this potential has only recently begun to be addressed. In one study, researchers performed integrated Hi-C, ATAC-seq, and RNA-seq on 12 primary samples to evaluate 3D genomic alterations associated with relapse and identified recurrent patterns of reorganization across multiple layers of chromatin organization[345]. In a more recent study, Micro-Capture-C was used to pinpoint *KMT2A::AFF1*-specific enhancers that regulate several leukemia-associated genes[346]. Another subtype focused study was performed by our lab before my arrival, where integrated proteogenomic and Hi-C were used to gain insights on the 3D organizational profile of HeH ALL[347]. Likewise, in **Paper I**, we used Micro-C to confirm the existence of a putative enhancer–promoter interaction and evaluated its causal role using CRISPR perturbation[321]. Nevertheless, as these examples illustrate, 3D genome studies in BCP ALL have largely been driven by targeted analyses, and a comprehensive 3D map of the BCP ALL genome with both large sample size and high resolution has been lacking. In this work, we aimed to fill this gap by providing, to our knowledge, the largest primary sample dataset of the BCP ALL 3D genome at the highest available resolution, together with an online web application to facilitate further research.

In this study, we performed WGS, RNA-seq, and Micro-C on 35 primary BCP ALL samples to evaluate the 3D genome landscape at multiple levels. On the compartment level, in addition to leukemia-specific analyses exploring alterations across samples and subgroups, we also compared the 3D genome of leukemia to that of HSPCs and more mature PBMCs and remission samples. We found that HSPCs, leukemia cells, and PBMC/remission samples formed three distinct clusters based on compartmentalization, consistent with the strong interconnection observed between chromatin accessibility profiles and hematopoietic cell differentiation stages[348]. As expected, leukemia cells were closer to HSPCs in their compartment profiles than to fully differentiated counterparts (**Figure 8A**). To ensure the robustness of our compartment calls, we compared shifted regions with RNA-seq outputs and confirmed that B-to-A shifts were associated with upregulation, and A-to-B shifts with downregulation. To further investigate the interplay between gene regulation and compartmentalization in leukemia, we examined B-to-A shifted regions in two well-established transcription factor fusion–driven subtypes, *ETV6::RUNX1* and *TCF3::PBX1*. Interestingly, when evaluating predicted binding sites of RUNX1 and PBX1, both transcription factors showed significant enrichment in the shifted A

compartments of their corresponding subtype, compared to shared A compartments (**Figure 8B-C**).

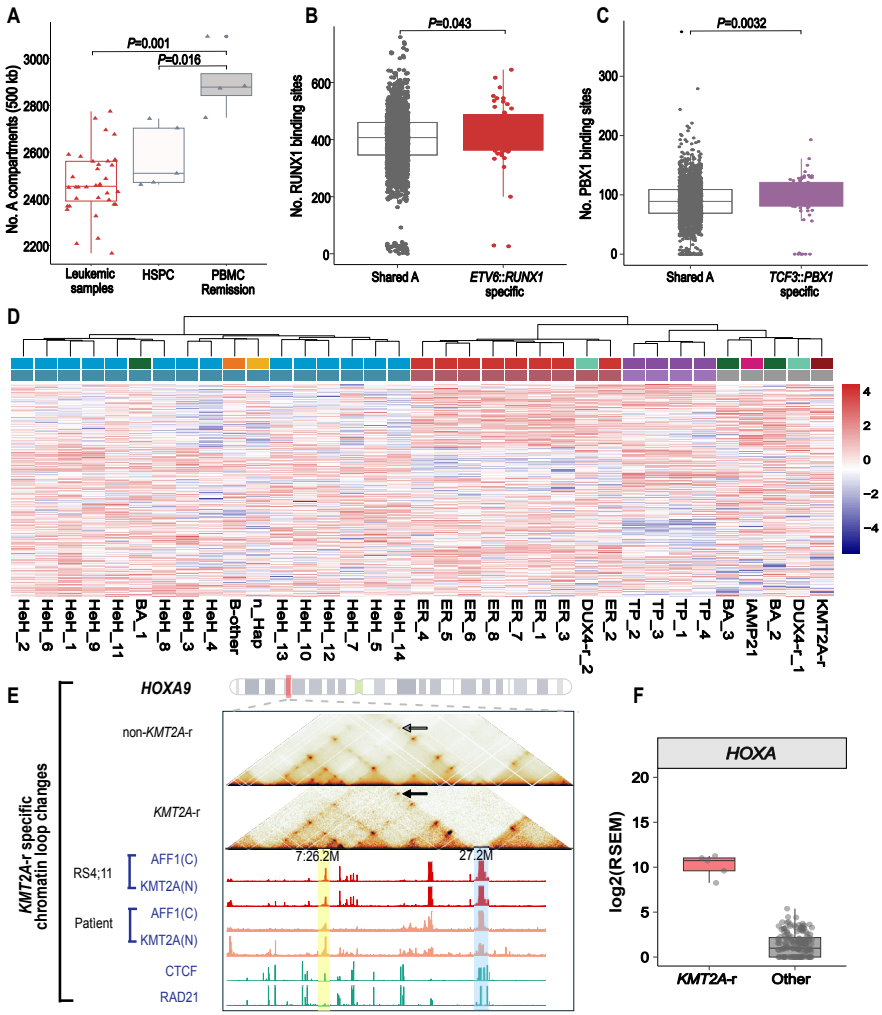


Figure 8 | Interplay between leukemia biology and 3D genome organization across multiple layers. (A) A-compartment frequency differs significantly between leukemia samples, hematopoietic stem/progenitor cells (HSPCs), and peripheral blood mononuclear cells (PBMC) or remission cases. (B–C) Somatic events are associated with subtype-specific compartment shifts: *ETV6::RUNX1*-specific A compartments are enriched for RUNX1 binding sites (B), while *TCF3::PBX1*-specific A compartments are enriched for PBX1 binding sites (C). (D) Unsupervised hierarchical clustering of the most variable topologically associating domain (TAD) boundary scores across 35 BCP-ALL samples reveals distinct clustering patterns associated with subtype. For clarity, a unique color code was assigned to each subtype, which can be seen above the heatmap and below the dendrogram. (E) Micro-C heatmaps demonstrate enhancer–promoter interactions on *HOXA* cluster that are specific to *KMT2A*-r cases, highlighted by the arrow. The bottom panel shows overlaid ChIP-seq tracks, where the enhancer and promoter of interest are marked by yellow and blue rectangles, respectively. (F) RNA-seq data show elevated expression of *HOXA* genes in *KMT2A*-r cases.

In the previous sections, the interesting dichotomy between the evolutionary conservation of TADs and the unexpected heterogeneity from single-cell studies have been mentioned. As the next level of analysis, we sought to examine how this dichotomy manifests in primary leukemia samples when viewed through the lens of inter-sample and inter-subtype variation. When evaluating the strengths of common TAD boundaries in our dataset, unsupervised clustering revealed strong subtype-specific patterns, particularly for the more frequent subtypes ($n > 3$), where every sample clustered with its corresponding subtype (**Figure 8D**). This highlights the important role of TAD boundaries in guiding subtype-specific gene regulatory programs. Notably, one *BCR::ABL1* sample clustered with the aneuploid cases. Upon further inspection, we found that this sample harbored 51 chromosomes, which lies at the lower end of HeH. Altogether, our analysis demonstrates that TAD boundary strength variance follows subtype-specific patterns and is highly dependent on ploidy.

One of the major strengths of combining Micro-C with a large cohort of primary samples was displayed at the loop level, where we achieved 1 kb resolution in the merged dataset, offering enormous potential for dissecting the cis-regulatory architecture of BCP ALL. At this resolution, we identified 11,360 promoter–promoter and 18,220 E-P contacts. To assess the robustness of these findings, we compared annotated enhancer and promoter anchors and observed significant enrichment for canonical histone marks. Moreover, 74% of transcripts with identified promoters within our contact matrix were expressed in our RNA-seq data. Anchors that were interacting with the promoters of unexpressed transcripts were significantly enriched for H3K27me3, an established silencing mark, further supporting our approach. When comparing our annotated enhancers to those reported for GM12878 in EnhancerAtlas, we found that the majority (53%) were novel[349]. We also noticed that promoter interaction detection was strongly associated with gene expression levels, with a correlation coefficient of 0.995 in a linear model, and 94% of transcripts with median log2(RSEM+1) values greater than 12 exhibited at least one promoter interaction in our contact matrix.

For the vast majority of genes (72.3%) identified in our 1 kb loops, a single promoter was observed. However, among the minority of genes with multiple promoters, we found a significant enrichment in the COSMIC cancer gene dataset (Odds Ratio = 2.14, $P = 8.55 \times 10^{-17}$; Fisher's exact test). These findings are consistent with pan-cancer studies demonstrating that alternative promoter usage is a common phenomenon with oncogenic relevance; however, to our knowledge, this represents the first investigation of this feature from a 3D genomic perspective and specifically in BCP ALL[350].

As with our analyses of compartmentalization and TADs, we also investigated subtype-specific patterns at the loop level. Due to sample size limitations, we performed this analysis at 10 kb resolution. Nevertheless, loop strength differences revealed strong

subtype-specific patterns, potentially influencing numerous established leukemia genes such as *HOXA9* (**Figure 8E**), *FLT3*, *TP53*, *CD44*, *ERG*, and *XBP1*. These findings were supported by the matched RNA-seq data, with one example shown in **Figure 8F**. Altogether, our findings suggest that loop architecture has considerable potential to explain regulatory mechanisms relevant to BCP ALL, both in general and in a subtype-specific manner. Given the more than 50,000 high-resolution loops identified in primary samples, it was beyond the scope of this study to investigate them individually. To enable further research, we developed a web-based application that allows users to query genes or genomic locations at 1 kb resolution for BCP ALL and 10 kb for subtype-specific findings through a user-friendly interface. The application is accessible online and, for more experienced users, can also be run locally from the terminal, as described in the repository: https://github.com/efeaydin-bio/microc-all-db.

Altogether, this study provides a comprehensive investigation of the 3D genome in BCP ALL and demonstrates its relevance to disease biology across multiple layers of regulation. We consider this work not only as a research study with defined questions and results, but also as a resource that can serve as a foundation for further investigations into the 3D genome architecture of BCP ALL and potentially other leukemias.

3.2.4. Paper IV

Interplay between somatic mutagenesis and the 3D genome in pediatric B-cell precursor acute lymphoblastic leukemia

Main results:

- By combining the 3D map generated in Paper III with WGS data from 801 primary specimens across three cohorts, we investigated the relationship between mutagenesis and the 3D genome in BCP ALL.
- A-compartments, TAD boundaries and chromatin loops (including the structural loops with unestablished functions) are depleted for mutations (**Figure 9**).
- The effect is not limited to mutation rates but also signatures, indicating the existence of biochemical constraints at play in addition to the evolutionary pressure.
- The protective effect from each of the 3D genome organization layers largely disappears after relapse.

In **Paper I**, we focused on the non-coding genome from a mutational perspective and highlighted its relevance in BCP ALL through spatial positioning within the 3D landscape. However, connecting these two aspects was challenging, as limited data on

the 3D genome of BCP ALL existed at the time and most of it were derived from low-input Hi-C studies on cell lines, constraining both resolution and generalizability. In **Paper III**, we sought to address this limitation by generating a high-resolution 3D genome dataset of BCP ALL using Micro-C on 35 primary samples. We demonstrated that each layer of genome organization contains extensive disease-relevant information with numerous potential applications. Finally, in **Paper IV**, we connect the story from **Paper I and III** together, and aim to address the interplay between the mutagenesis and the 3D genome in BCP ALL.

In 2012, researchers demonstrated that chromatin features explain more than half of the variation in mutation rates across the genome[35]. Follow-up work extended this link beyond mutation frequency to include mutational patterns[36]. Together, these studies established that the 3D genome is closely tied to mutagenesis and this connection was further explored under different disease contexts[351,352]. However, due to the limited availability of genome organization data from primary samples, this relationship has not been thoroughly examined in BCP ALL. In this study, we also targeted this gap by analyzing WGS data from 801 primary BCP ALL samples spanning three independent cohorts and integrating the results with the 3D chromatin map generated in **Paper III**.

Our results demonstrate that the 3D genome influences both mutagenesis rates and patterns across multiple layers of organization in BCP ALL. At the compartment level, we observed significantly higher mutation frequencies in B-compartments compared to A-compartments (**Figure 9A-B**), a pattern widely reported across tissues and generally attributed to the late-replicating nature of heterochromatin. We then explored this dichotomy under the lens of mutational signatures. SBS1 and SBS5, the two most frequent signatures in BCP ALL, demonstrated a specific enrichment within A-compartments[28,353]. By contrast, late-replication–associated signatures SBS8 and SBS18 were enriched in B-compartments, underscoring the robustness of our data integration[354,355]. Interestingly, SBS2, a signature typically localized to early-replicating regions in lung and breast cancers, was enriched in B-compartments in our cohort[356]. These findings indicate that both broad biological principles and disease-specific factors shape the mutational landscape, with compartmentalization emerging as a major determinant. Moreover, analysis of 64 matched relapse WGS samples revealed a significant increase in the proportion of A-compartment mutations compared to the diagnostic cohort. Considering the limited alterations reported in compartmentalization at relapse[345], these results suggest that evolutionary selection or changes in DNA repair mechanisms increasingly dominate the mutational landscape, rather than non–tissue-specific biochemical constraints imposed by compartmentalization.

Next, we investigated the impact of TAD boundaries on the mutational landscape (**Figure 9C**). As expected, these boundaries acted as a protective layer against mutations; however, this protective effect was more pronounced in B-compartments, consistent

with previous pan-cancer studies[36]. Interestingly, SBS5, the clock-like signature globally enriched in A-compartments, was significantly enriched in the TAD boundaries of B-compartments compared to the TAD-boundaries in A-compartments. When we combined our data with CTCF ChIP-seq outputs generated by ENCODE, we have seen that TAD boundaries within A-compartments had significantly more CTCF binding than those that are found on B-compartments. Notably, previous studies demonstrated SBS5 to be depleted on CTCF binding sites in hematologic malignancies[357], which may provide a rationale for this seemingly counterintuitive result.

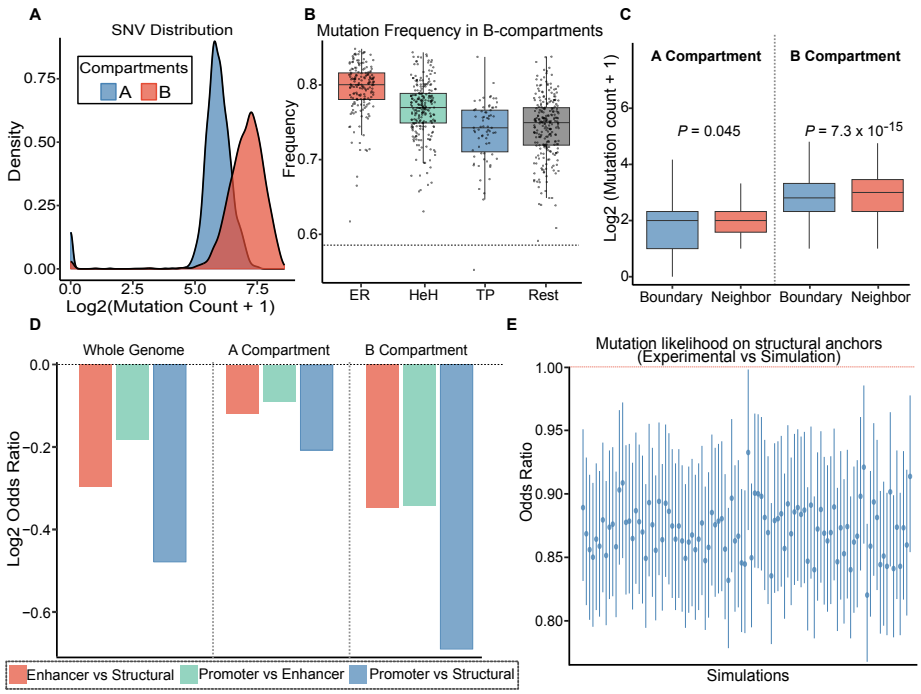


Figure 9 | Mutational landscape of BCP ALL from a 3D perspective. A) Density plot showing the mutational load difference between compartments. B) Boxplots showing the mutation frequencies of B compartments per sample across common genetic subtypes. Dashed line represents expected frequency based on the genomic coverage of B-compartments. C) Density of mutations in TAD boundaries and neighbours faceted by compartments (Fisher's exact test). D) Mutated/non-mutated comparison across each anchor classification faceted by compartments. For all comparisons, mutation likelihood follows the same order: promoter < enhancer < structural E) Mutated/non-mutated comparison between structural anchors against 100 simulated anchors to mimic the experimental dataset. Every comparison shows lower mutational likelihood in structural anchors, indicating functionality. Abbreviations: BCP ALL, B-cell precursor acute lymphoblastic leukemia; HeH, high hyperdiploidy; TAD, topologically associating domain.

Finally, we investigated the effect of chromatin looping on mutagenesis. We found that all classes of chromatin loops (promoter, enhancer, and structural) conferred a protective effect against mutations, albeit to varying degrees (**Figure 9D**). The putative

protective effect of structural anchors on mutation rates was evaluated by generating simulated anchor data that mimicked the experimental interaction matrix (**Figure 9E**). Promoter loops were the most strongly protected, followed by enhancer loops and then structural loops, which is consistent with evolutionary expectations. To highlight the disease relevance of this observation, we examined recurrently mutated promoter anchors and identified several well-established leukemia genes, including *BCL9*, *XPO1*, *PBX1*, and *MYB*[28,358–360]. Moreover, analysis of matched relapse samples revealed a significant increase in mutations at both promoter and enhancer anchors compared to diagnosis. Enhancer anchors that acquired at least two additional mutations at relapse were linked to 18 putative target genes, among which were important leukemia players such as *BCL6* and *PRMT1*[361,362].

Altogether, this study—to our knowledge—represents the most comprehensive investigation of the interplay between 3D genome organization and mutagenesis in BCP ALL. We demonstrate that organizational layers of the 3D genome influence mutagenesis at every level, yet the mutational output shaped by these biochemical constraints is markedly altered at relapse. Our findings provide a rationale for the repeatedly observed depletion of driver mutations in the non-coding genome, a topic discussed in detail both in this thesis and in **Paper I**. While we acknowledge that the fine-tuning functions of regulatory elements undoubtedly contribute to this depletion, our study underscores how misleading any systematic assessment of non-coding variation can be when spatial genome context is not considered.

4. Concluding remarks and future perspectives

In the introduction of this thesis, I noted that the age of polymaths is over, and due to the magnitude of information generated, scientific progress now depends on collaboration between specialists. Following this "new normal", I underscored the risk of a drawback in the same section, using Maslow's hammer-and-nail metaphor, to highlight how researchers tend to search for answers within the scope of their own expertise. This is a difficult problem to escape. Just as early 20th-century surgeons believed the answer to breast cancer lay in more extensive surgery, a bioinformatician working on the regulatory genome in the age of AI may also unconsciously follow familiar paths. Writing about the future of leukemia and the non-coding genome is therefore not straightforward, but it is necessary. After all, based what we have seen from the thousands of years old scientific pursuit so far, this road most likely does not have an end, and the inherent structure of any scientific endeavor requires a next step. In this closing section, I will humbly reflect on what those steps may look like, combining the work presented in this thesis with the current status of leukemia and non-coding genome biology.

As emphasized in axiom #1 from the introduction, the central problem in BCP ALL, as in many cancers, remains relapse and refractory disease. Single-cell studies show that leukemia is polyclonal at diagnosis[363], and our current understanding suggests that relapse is due to the survival of a chemoresistant clone[364]. Thereby, in theory, if biomarkers could be identified for every clone and matched with targeted therapies, relapse could be prevented. This parallels the logic of narrow-spectrum antibiotics, while also connecting to axiom #2, which underlined the toxicity of chemotherapy intensification and the urgent need for alternative combinations. Yet, neither identifying a comprehensive set of clonal biomarkers nor developing molecules to target them is an easy task. This challenge connects this thesis to axiom #3, which highlights the link between somatic events and cancer, as enhancing our understanding of leukemia biology through functional interpretation of observed alterations is one of the key steps in this difficult path.

Until very recently, the textbook definition of BCP ALL was the arrest occurring in immature pro-B cells due to somatic events, eventually causing the associated

symptoms. Thereby, the question of interest was to identify the hidden event(s) responsible for a shared outcome. This was already a challenging task, and yet single-cell studies demonstrated that the biological reality is even more complex. Current knowledge suggests that different subtypes of BCP ALL can originate from distinct hematopoietic maturation stages, with varying degrees of multipotency intact, giving a subset of cells the ability to switch to the myeloid lineage under selective pressure[206]. Heterogeneity cannot be comprehensively understood under a singular perspective either. From the 3D perspective, as already mentioned in the background section, organizational layers identified in genome-wide bulk chromosome conformation capture studies also display heterogeneity. Comparative studies incorporating single-cell data revealed that even these layers are represented in only a limited fraction of cells[177,178]. Nevertheless, the output of a Micro-C snapshot still carries substantial biological information. In our work (**Paper III**), we demonstrated the connection between the compartmentalization of the 3D genome and the cell differentiation stages of leukemia, highlighting subtype-level differences. Moreover, we demonstrated that TAD boundary strength is a strong classifier of leukemia subtypes, as samples of the same subtype tend to cluster together based on this feature. Together, the work of us and others demonstrates that biologically relevant patterns that complement each other can be derived from different approaches. On the other hand, we also now know that, in the absence of canonical mutations, epigenetic alterations may exert corresponding oncogenic effects[365], as well as modulate drug response[366]. Thereby, one must exercise caution when interpreting regulatory data from complementary resources, and yet identifying convergence patterns across different data layers may be a promising strategy for future research.

During the time of this doctoral work, neither the etiology of BCP ALL nor the "CRE code" was fully understood, and I doubt that we will reach such comprehensive knowledge in the near-future. Both concepts are marked by overwhelming complexity, with multiple layers of variables interacting in countless combinations. Due to this combinatorial nature, reconstruction of these systems becomes a complex problem, as investigating a single part in isolation can be insufficient, and perhaps even misleading. We now know that, by connecting with one another, the deciphered nodes of the regulatory system establish a vast range of possible meanings; yet neither the information nor the meanings are comprehensive, leaving us with sets of known unknowns and unknown unknowns. That is why we do not know the cause of BCP ALL, and why no single epigenetic mark is sufficient to single-handedly annotate a CRE. This challenge has been the origin of thought for this study. Thereby, at its core, I might say that what is presented here is a praise to multi-omics.

Throughout every paper, we integrated different omics data to minimize the error-proneness of a singular approach. We specifically relied on high-resolution 3D genome methodology and emphasized its basis in transcriptional programs. After all, although

we do not fully understand the rules or mechanisms guiding these self-forming units, we know that the architecture is associated with transcription, and whether it is a causal event or a byproduct, spatial context remains essential. We highlighted this in **Paper I** by criticizing the one-dimensional definition of a mutational hotspot and provided further support in **Paper IV** by showing significant associations between mutagenesis patterns and each layer of chromatin organization. A better understanding of the interplay between mutagenesis and epigenetics will likely provide important insights going forward.

I have repeatedly highlighted the limitations of coding-genome– or single-omics–focused studies throughout this book. Here, at the end, to complete the circle of thought, it is only fair to acknowledge the limitations of our studies as well. First of all, as mentioned earlier, we are limited by our incomplete understanding of both BCP ALL and the CRE code. For this reason, despite the vast number of primary samples included throughout the thesis, this work, I believe, has a stronger-than-expected fundamental biology perspective attached to it. In fact, I can even describe it as a basic science inquiry within the context of a specific disease, where the primary specimens' main benefit was to provide representativeness. It is possible that to some colleagues, such a definition may sound like downgrading the value of a study, but the gap between basic and clinical research is far smaller than in the past, and the intersection is far bigger. The clearest example is perhaps the transformation of the discrete hematopoiesis model into a continuous, heterogeneous one, as a result of single-cell studies. It is now shown that the projection of leukemia cells on this developmental map is tightly connected with prognosis[206]. A better understanding of this model has the potential to improve risk stratification and treatment decisions. Based on this, I also find it plausible that basic discoveries on the CRE code, perhaps even in model organisms, could have rapid clinical consequences in leukemia. In fact, when we look at **Paper III**, where we generated the largest primary-sample–derived 3D chromatin interaction map of the leukemia genome, we also noticed that none of the markers we included from the literature single-handedly explained any of the 3D layers. It is possible that, with new developments in fundamental biology, the vast amount of data we generated here could be reused for new discoveries. With the democratization of data access and sharing, many treasures can be found. This is further highlighted in **Paper IV**, where we showed that structural loops with no well-established functions provide a layer of protection against mutagenesis. Considering the rate of discoveries we have witnessed in the 3D genome within the last decade (such as the recent identification of tethering elements in fruit flies[367] and even more recently discovered extenders in mice[368]), I find it quite plausible that new classes of CREs or subclasses of existing ones will be identified, with potential clinical implications. For example, BET inhibitors, which cause downregulation of super-enhancer associated oncogenes are already under clinical trials, and yet the potential of such a CRE-associated drug is limited to diseases where the dominant drivers are super-enhancers[369]. If advances in fundamental biology indeed

uncover putative subclasses of enhancers, this could pave the way for the development of selective molecules that target CREs in a more disease- or clone-specific manner, while minimizing off-target risks. I strongly believe that when this future arrives, the data we generated throughout this doctoral work will be a part of it.

Another obvious limitation of our work lies in the methodological dependence on bulk data. We know that, at least to some extent, bulk and single-cell data agree, and in some contexts, the extraction of systematic output can be more useful than the output of the parts. A famous example is the heartbeat, which is regulated by stochastic and non-linear dynamics at the molecular level[370]. Nevertheless, we cannot deny that numerous heterogeneity-related features were potentially missed due to the limitations of these studies. For example, in **Paper II**, we highlighted the ASE landscape of BCP ALL, and yet we depended on external SNP data for phasing. If we tried to implement a similar approach on the Micro-C data we generated, this would mean the loss of >90% of the data due to dependence on SNPs within the limited genomic span of junctions. Numerous algorithms were developed to address this limitation[371–373], and yet, in the absence of reliable long reads, allele-specificity remains a problem. Other heterogeneity sources also remain unexplored. For example, in **Paper III** we showed enrichment of alternative promoter interactions among COSMIC genes, but since bulk chromatin data is a snapshot of a dynamic process, it cannot reveal whether this finding reflects allele-specificity, single-cell variability, or stochastic transcriptional bursting. As pan-cancer studies suggest strong prognostic associations for alternative promoter usage[350], unravelling the underlying biology here could also have a clinical relevance. On the other hand, since there is also a considerable amount of unknowns on the mechanical side of gene regulation, diving into single-cell resolution does not have the power to single-handedly decipher the CRE code either. Because of this, methods that incorporate both single-cell and multi-omics approaches are being developed, where multiple layers of regulatory and transcriptional information can be simultaneously detected at the single-cell level[374,375]. It is likely that these integrated single-cell multi-omics methods will illuminate many unknowns in gene regulation within the next decade. Of course, this does not mean that the single-cell perspective necessarily overthrows bulk data, it will still remain relevant. It should be remembered that, although a relatively older approach, bulk-scale omics is still a recent technology with significant room for improvement. For example, within the field of the 3D genome, an obvious obstacle of the Hi-C derived methods was the technical limitation of restricting the interaction landscape to pairs[376]. Anyone who has left cables unattended in a closed environment will recognize the complex self-assembly patterns that cannot be captured by pairwise approaches. The case with the chromatin is similar, and interactions are not necessarily limited to two genomic anchors in real space. To account for possible triplet, quartet, quintet, and higher-order formations, new methods such as genome architecture mapping (GAM)[377] and split-pool recognition of interactions by tag

extension (SPRITE)[378] are emerging. These developments may also provide additional information that could reshape our existing knowledge of CREs.

Finally, it will be anti-climactic to end a thesis written in 2025 without mentioning a thing or two about AI. The regulatory genome with all of its elements still depend on the laws of physics, which allows for accurate modelling of the chromatin behaviour through simulations under defined biochemical constraints. Members of the 4D nucleome project expect machine-learning to gain more weight in this field with the growing number of associated datasets[379]. A particularly interesting example is DeepMind's AlphaGenome, which aims to reproduce the success of AlphaFold in functional genomics. If such tools mature, they may fundamentally transform how we study chromatin architecture and regulation.

At present, numerous clinical trials are underway, testing different molecular combinations alongside extensive animal studies. Encouraging reports suggest that advances in combination therapy will continue to play a central role in the fight against leukemia. A widely held prediction is that precision medicine will become increasingly selective and patient-specific, an idea that directly connects to the three axioms introduced in this thesis. Improvements in risk stratification and prognostic prediction, along with developments on the pharmaceutical side, will hopefully help us achieve more with less toxicity. Undoubtedly, the deciphering of the CRE code and the broader non-coding genome will be integral to shaping this future.

Popular science summary

Cancer has been part of human life for as long as we have existed. Fossil evidence shows tumors not only in ancient *Homo sapiens* but also in Neanderthals and even earlier hominin ancestors. For most of history, humanity was on the losing side of this struggle. Only with modern science, through decades of concentrated effort and international collaboration, have we begun to fight back. The journey is far from over, as millions of patients and families still face cancer every year, and nearly everyone has known someone lost to the disease. Yet if we step back and look at what has been achieved in less than a century, there is reason for hope: many cancers that were uniformly fatal just two generations ago now have survival rates exceeding 90%. That being said, when it comes to cancer, even the best-case scenario can involve serious drawbacks, as many survivors still face an increased risk of long-term complications. If remaining life expectancy minus age at diagnosis is used as a metric to evaluate this impact, one patient group—children—stands out as carrying the greatest potential for cumulative harm.

The most common pediatric malignancy is B-cell precursor (BCP) acute lymphoblastic leukemia (ALL). Thanks to modern therapy regimens and risk stratification, BCP ALL, a disease once considered incurable, now has overall survival rates above 90% in developed countries. This progress stems largely from the development of combination chemotherapy in the 1970s and its steady refinement ever since. In simple terms, the goal of any cancer drug is to kill as many cancer cells as possible while causing as little harm as possible to the rest of the body. Unfortunately, no treatment has yet achieved perfect precision. Moreover, recent trials suggest that current therapy has reached its limit of tolerance, meaning that further intensification is unlikely to improve outcomes. Considering that thousands of children still succumb to the disease each year, and many more live with long-term complications from chemotherapy, two essential challenges now dominate pediatric ALL research: improving outcomes in high-risk patients and reducing toxicity in low-risk patients. To achieve either, we must generate new knowledge, both about ourselves and leukemia. This thesis represents a small attempt towards that goal. To do so, we turned to the blueprint of life: the genome and the hidden information it contains.

Anyone with a notion of interest in biology has likely encountered the grand metaphors attached to DNA over the years, such as “The Book” or “The Blueprint” of life. Since the discovery of the double helix, people have speculated about what might follow once

we could finally read what was written inside. This has not been a sole academic interest either, as evidenced from the global fame of Dolly the sheep or the cultural products such as Jurassic Park or GATTACA. In fact, 25 years ago, when Human Genome Project completed the first survey of the human genome, former US president Bill Clinton announced this achievement by referring to DNA as the language God used to create life. In other words, expectations on DNA were always sky-high. Yet such expectations risk a reductionist fallacy: in many complex systems, the whole is greater than the sum of its parts. This is particularly true for the life we know on earth. For example, there are hundreds of specialized cell types in a human body, yet all contain the same DNA (aside from the somatic alterations). How, then, can one blueprint give rise to such extraordinary diversity?

The answer lies in the complex regulatory architecture that evolved over time. These additional layers of control, collectively referred to as epigenetics (from the Greek epi, meaning "above"), vastly expand the functional possibilities of DNA. Thus, the metaphor of DNA as "The Book of Life" was not wrong, but incomplete. In reality, these regulatory layers interact in combinatorial ways, shaping when and where genes are expressed. This precise control is what allows a single fertilized egg to develop into a complex, highly organized mammal. Studying all of these layers together on a genome-wide scale is known as multi-omics, an approach that only became possible with the next-generation sequencing revolution of the 21st century.

One of the striking findings from the Human Genome Project was that only about 1.5–2% of the human genome coded for proteins. For decades, scientists mainly looked at this "protein-coding" part and the rest was often considered as "junk DNA". This view was challenged by the ENCODE project, which reported that roughly 80% of the genome shows signs of activity. To what extent the genome is functional is an ongoing debate, but one conclusion is clear: the non-coding genome plays a much greater functional role than previously imagined. In 2013, this shift in perspective produced a landmark discovery: a tiny change occurring within an on-switch of the *TERT* gene, was found to fuel carcinogenesis in melanoma[32,33]. Since then, alterations within the non-coding genome have been recognized in many cancers, though their full impact remains underexplored.

In this thesis, we investigate the relevance of these overlooked genomic regions in pediatric BCP ALL, with the hope of providing new insights into a disease where progress is urgently needed. Shortly before I began my thesis, an inspiring *Nature* paper was published that presented a pan-cancer analysis of non-coding mutations[325]. For non-coding genome enthusiasts like me, the results were somewhat sobering, as driver mutations were found to occur at significantly lower rates in these regions than in protein-coding ones. This pattern has since been reinforced by subsequent studies. Consistent with these findings, when we analyzed 345 samples for non-coding mutations in **Paper I**, we observed the same depletion.

The shortage of cancer-driving mutations in non-coding DNA likely comes from both biological and technical reasons. But in our view, it's also a matter of perspective. Most searches for important mutations have been done in a linear way, by comparing where mutations appear along the DNA sequence to what would be expected by chance. This makes sense for the parts of DNA that make proteins, because those regions follow clear rules about how the genetic code is read. However, this linear logic does not capture how the non-coding genome operates. Regulatory elements do not follow a single universal code but rather consist of multiple classes with diverse mechanisms of action. More importantly, their function is inherently three-dimensional. Because the genome folds and loops inside the nucleus, a control switch called an enhancer can activate a gene that lies millions of letters away along the linear sequence. The vast distances seen along the linear genome, when viewed through our 3D glasses, can often turn out to be mere illusions.

That's why, in brief words, this thesis is a praise to multi-omics and the 3D genome. Throughout our work, we did not only tackle the question of the functions of non-coding regions in BCP ALL but also tried to explore the fundamental biological aspects of our findings that are transferable to other disease contexts.

In **Paper I**, we highlighted the limitations of investigating somatic mutations as stand-alone events. When we combined mutation data with information about gene activity, epigenetic marks, and clinical features, hidden patterns began to emerge. We found that certain non-coding mutations, specific to particular subtypes of leukemia, tend to cluster near well-known cancer genes. This finding fits naturally with our later work in **Paper III**, which revealed that the 3D arrangement of the genome differs across subtypes on multiple levels. We also tested one of these mutated non-coding regions in the laboratory and confirmed that it helps control the activity of *NRAS*, a gene long recognized for its role in cancer.

Our aim was not only to point out the limits of current research on the non-coding genome but also to overcome some of those barriers ourselves. In **Paper II**, we developed a computational tool called BASE, designed to study how the two copies of each gene can behave differently inside cancer cells, a phenomenon known as allele-specific expression. This imbalance is common in leukemia and can be influenced by changes in non-coding DNA. Earlier tools could measure this expression imbalance, but they focused on the outcome itself rather than on how much it is driven by extra or missing chromosomes, which are frequently seen in ALL. We built BASE to address this gap and applied it to a common ALL subtype, showing that most of these imbalances were indeed driven by such alterations.

In **Paper III**, we mapped the 3D structure of the leukemia genome in 35 patient samples using a powerful technique called Micro-C. To our knowledge, this remains the largest 3D dataset based on primary blood-cancer samples. We demonstrated how

crucial 3D genome organization is in BCP ALL, and how somatic events can reshape every layer of chromatin architecture, that tightly regulates the behavior of leukemia cells. This vast dataset with thousands of newly identified regulatory loops highlights how limited our knowledge of the regulatory architecture of BCP ALL has been. And yet again, even our analysis was based on this limited previous knowledge combined with our generated data. With subsequent research and new insights in gene regulation, many important alterations that we missed in our own study can be rediscovered.

One potential aspect is how the 3D genome relates to mutagenesis, since both concepts are crucial in leukemia biology. In **Paper IV**, we addressed this question by integrating our 3D map with large-scale data from 801 BCP ALL samples. We show that the organizational layers of the 3D genome are tightly connected with mutagenesis, affecting both mutation rates and patterns. To facilitate future research, we developed a user-friendly web application where researchers can easily access and visualize the chromatin interaction map we annotated.

Altogether, this thesis explores the often-overlooked non-coding genome in BCP ALL. Throughout this work, I refused to see the well-established *TERT* promoter events as a coincidence. Instead, I thought of these findings as a small subset of signals coming from a larger, uncharted landscape, limited by the single dimensional interpretation of gene-regulation. Here, I aimed to offer a new perspective. These findings provide new insights into the functional roles of non-coding genome in this disease and strengthen the case for including these regions in leukemia research.

Acknowledgements

Despite beginning my PhD later than the average candidate, I can say with certainty that these four years have been the most transformative period of my life, both personally and intellectually. Yet without certain interactions along the way, this process would have remained only as a remote possibility within the vast event space. Therefore, I would like to acknowledge the people who have shaped and influenced this reality, knowingly or unknowingly, throughout these years.

First and foremost, I am grateful to **Kajsa** for giving me this opportunity. When I started my PhD, I was not sure what to expect, I was simply excited for the chance to pursue my doctoral studies under the supervision of a world-class expert in a topic I had long admired. The story did not disappoint. You introduced me to the fascinating world of the 3D genome and gave me the freedom to explore it from my own perspective, using state-of-the-art approaches that were so elegantly assembled thanks to your creativity and scientific intuition. You lead by helping people discover their own light, guiding them past obstacles, and somehow connecting those independent paths to form something greater than the sum of their parts. Looking back, I realize that you always treated me as a colleague rather than a student, even when I came to you with the most nonsensical ideas. You knew they were unlikely to work, yet you explained things in such a thoughtful way that not only helped me see the flaws but also opened new ways of thinking. The knowledge and questions that emerged from this journey now form the foundation of my scientific identity, and my interest in the non-coding genome has grown by orders of magnitude. Yet all of this remains secondary to the inspiration you have given me as a person and as a leader. Thank you for being such a humble and kind mentor, and for creating the best possible environment for us to work in.

Another person I cannot adequately highlight the importance to this thesis is **Minjun**, who co-supervised all the projects. You have been like a living, breathing and much less toxic version of Stack Overflow, not only to me or to our group, but to everyone around you. It is unbelievable how you manage to do so much and still find the time to help so many others. It has been an incredibly valuable experience to work under the supervision of a bioinformatician like you, and I'm deeply grateful for all the life-saving hacks you have shared throughout the years.

I would also like to thank all the lab members for their collaborative spirit and support throughout this work. A special thanks goes to **Larissa**, a fellow metalhead and chromosome whisperer, for optimizing the Micro-C and related protocols that became central to this thesis, and for the occasional reminders about data meetings that I was prone to forget. **Ellie**'s patience and guidance in teaching proper laboratory practice were essential, particularly for Paper I, which would not have been possible without her support. **Gladys** brought invaluable expertise in FACS, which played an important role in this work. I am also grateful to fellow PhD student **Charlotte** and former lab member **Jonas** for our collaborations. Also, big appreciation to **Henrik**, **Rebeqa**, **Bertil**, and **Thoas** for their valuable inputs as co-authors.

The importance of communities and institutions in the progression of science is well established. During my doctoral years at Lund University, I witnessed this firsthand at Clinical Genetics, where every moment — from lunch breaks to Friday seminars — offered opportunities for intellectual growth. This supportive and curious atmosphere continuously motivated me to improve, and I would like to acknowledge the PIs — **Anna Hagström**, **Marcus Järås**, **Karolin Hansén Nord**, **Fredrik Mertens**, **Bertil Johansson**, **David Gisselsson Nord**, **Thoas Fioretos**, and **Axel Hyrenius Wittsten** — for their inspiring science and leadership, which made this environment possible. I also extend special thanks to **Anna** for taking the time to be the opponent at my half-time seminar and for a discussion that challenged and broadened my perspectives. Additional thanks go to **Anette Welin** for expertly managing the department's administrative matters.

As a fortunate coincidence, this year marks the 50th anniversary of the department, and it is impossible to speak about Clinical Genetics without acknowledging the person who started it all. My deepest thanks go to **Felix Mitelman** for founding the department, which has inspired generations of scientists. His and **Nils Mandahl**'s ongoing hard work to push the boundaries of scientific knowledge stands as a testament to the spirit of this department. It has been my utmost honor to know these exceptional scientists personally.

Whether good or bad, every day in a PhD journey is a candidate for surprise (and more likely a bad one if you are a bioinformatician in the cell-lab). Yet, when I look back, I realize that one of the biggest surprises came from a social aspect. I never expected to meet so many inspirational people who would broaden my horizons on so many levels. Among them, I would like to acknowledge three prominent researchers who have become an integral part of my life.

I would like to begin by expressing my gratitude to **Ton** — the scientist, musician, photographer, local radio celebrity, licensed barbeque maker, and Barça enthusiast — for organizing 90% of the social events I took part in during my doctoral years, some of which changed my life in the best possible way. You are a full-stack nerd who can

discuss any topic for ten hours straight with the same motivation, because your enthusiasm, curiosity, and engagement come from within. It has been a privilege to have a friend like you.

Next is **Subhayan** — the prominent mathematician and sarcasm expert. I am grateful for our daily lunch trips in the Citadel, which often felt like a low-budget tech-bro podcast. I learned a great deal from our conversations about many things, and Kung Pao chicken without your company just sounds wrong. You are smart and creative, and one does not need machine learning to predict that good things are ahead of you. As Palpatine once said, *we will watch your career with great interest*.

Here comes the part about **Ping** — a stable node in a complex dynamic network. Many things can change in the world, but some things are certain: if one goes to BMC, one sees Ping doing serious science; if one hears Ping speak, one knows it is the truth and nothing else; and if one needs Ping, Ping will be there. I have enormous respect for this guy, both as a scientist and as a friend. Thank you for being who you are — and for making the greatest dumplings.

There are also key people who shaped my scientific journey before my arrival in Lund, and I would like to acknowledge them here as well. I will forever feel indebted to **Toni Gabaldón** for supervising my first independent research, giving me the opportunity to participate in his lncRNA project, and for his support and guidance. I also want to extend my gratitude to **Recep Demirgan**, for sparking my interest in the non-coding genome and teaching me about microRNAs. My college roommate **Canberk Sevim**, who has profoundly influenced my way of thinking, deserves special thanks for our philosophical discussions that continue to this day. I would also like to thank **Uğur Akyol** for always picking up the phone whenever I faced a coding emergency. All these people have, knowingly or not, contributed to this work, and I am grateful to them.

I would also like to extend my thanks to all C13 colleagues who made this journey so memorable and fun. Some specific names include Schrödinger's **Eva**, who can be both in Copenhagen and Lund until observed, for being an inspiring scientist and a fun friend who always brings good stuff to fika. My office-mate **Josephine**, for becoming a good friend despite the fact that we barely spoke at work — a paradoxical friendship born from silence. **Sibel**, for being such a strong character and for all the fun we had in numerous after-works, dinners, and trips. **Pablo**, for always coming up with the wisest comments at the lunch table and for the unexpectedly fun EHA trip. **Somadri**, for introducing me to everyone on my first day and for all the subsequent laughs. **Katrin**, for the legendary science joke that I hope to forget one day. **Ram**, for the good vibes and for sharing his cell biology expertise during our office discussions. **Maria**, for the engaging conversations at the Mediterranean lunch table and the occasional after-works. **Jakob**, for the inspiring CAR-T cell research and making sure all is smooth in the office during the weekends. To the party pixies, **Hanna**, **Louise**, **Noelia**, **Valeria**,

Karim, **Natalie**, and **Mattias**, for taking the time to organize memorable events that brought everyone together. **Kaname**, **Michele**, **Bahar**, **Ludvig**, **Valeriia**, and **Maggie**, for leaving your mark on the history of the middle office. To fellow genomics people, **Saskia**, **Vendela**, and **Valeria**, for walking the same path together. And to **Andrea**, for always being kind and helpful whenever I needed something in the lab.

My doctoral work took four years to complete, yet it felt like a moment because of the inexplicable joy of researching something you are truly curious about. I am deeply aware of what a privilege it is to devote such concentrated effort of mind and body to something over the years, and to find joy in it, without even knowing what to expect. Reaching this state-of-mind in life took me decades longer than PhD, and without my family providing confidence through their unconditional love and support, I would not have had the freedom, courage, or opportunity to follow this calling. That is why I dedicated this book to my mom and dad — for the simple reason that none of this would have happened without them. An honorable mention also goes to my late grandmother, who was inexplicably fascinated by science and the workings of nature despite her age and limited education. It would have been nice to share my findings with her.

Now, how to distil a final reflection of it all…

If you ask any scientist about their doctoral years, most will describe them as a time of sharp oscillations, where hope, excitement, frustration, and doubt alternate on chaotic patterns—and breakdowns are simply part of the process. In my case, to be honest, I didn't feel these oscillations as much. Yes, I spent countless hours trying to solve problems I barely understood, received cold rejections from journals, failed experiments, and questioned my intelligence like any proper PhD student would—and yet, it was always okay. None of those feelings ever lingered for long, and that was not because of my spiritual maturity (not sure if I have any), but because of you, **Hannah**. From the moment our paths crossed, your presence in my life has been so fulfilling that it left no room for worries about the future, or regrets about the past. You are not only my best friend, but also my favorite scientist, pianist, painter, and fighter of all time. As this journey ends, and uncharted territories filled with thousands of variables begin to unfold before our path, I am free of all worry—and for that, I thank you.

Notes on illustrations and text

The front cover of this thesis features a Penrose triangle, originally created by the Swedish artist Oscar Reutersvärd in 1934, and introduced to the academic community after Lionel and Roger Penrose independently published it in the *British Journal of Psychology* in 1958 (https://doi.org/10.1111/j.2044-8295.1958.tb00634.x).

My interest in such optical illusions originated from Douglas Hofstadter's "*Gödel, Escher, Bach: an Eternal Golden Braid*", which famously explores the properties of self-referential systems. I found the concept of tangled hierarchies discussed in that book strikingly relevant to the 3D genome, where many relationships defy strict hierarchical organization.

This goes together with how we visualize the layers of these structures; compartments and TADs, as checkerboard patterns and triangles on heatmaps. The cover integrates these visual motifs by portraying a Penrose triangle with a checkerboard pattern interwoven by chromatin loops, symbolizing the recursive and interconnected nature of genome organization in three-dimensional space. This depiction also resonates from a leukemia perspective, as the refutation of strict hierarchical models has also shaped our current understanding of hematopoiesis. It is a pleasing coincidence that this impossible object was first drawn by a Swedish artist, given that this work was carried out at Lund University.

The Penrose triangle icon and the black background composition were generated using ChatGPT 5 (OpenAI) through iterative visual prompting. The resulting image was refined in Affinity Designer 2 for final adjustments.

The back cover was created using BioRender, featuring schematic icons representing chromatin, compartments, TADs, and loops. The loop is depicted as an orange ribbon, referencing leukemia awareness. The accompanying Hofstadter quote feels particularly relevant, as we continue to explore the dichotomy between evolutionary conservation and heterogeneity within these structures.

Figure 5 was also prepared using BioRender.

Throughout the main text, I used ChatGPT 5 for grammar checks and stylistic refinements. All suggestions were critically reviewed to ensure that no scientific meaning or clarity was altered.

References

1. Lam, C. G., Howard, S. C., Bouffet, E. & Pritchard-Jones, K. Science and health for all children with cancer. *Science* **363**, 1182–1186 (2019).
2. Hanahan, D. Hallmarks of Cancer: New Dimensions. *Cancer Discov* **12**, 31–46 (2022).
3. Hanahan, D. & Weinberg, R. A. Hallmarks of Cancer: The Next Generation. *Cell* **144**, 646–674 (2011).
4. Hanahan, D. & Weinberg, R. A. The Hallmarks of Cancer. *Cell* **100**, 57–70 (2000).
5. DeVita, V. T. & Chu, E. A History of Cancer Chemotherapy. *Cancer Res* **68**, 8643–8653 (2008).
6. Pui, C.-H. & Evans, W. E. A 50-Year Journey to Cure Childhood Acute Lymphoblastic Leukemia. *Semin Hematol* **50**, 185–196 (2013).
7. Laetsch, T. W., DuBois, S. G., Bender, J. G., Macy, M. E. & Moreno, L. Opportunities and Challenges in Drug Development for Pediatric Cancers. *Cancer Discov* **11**, 545–559 (2021).
8. Robison, L. L. & Hudson, M. M. Survivors of childhood and adolescent cancer: life-long risks and responsibilities. *Nat Rev Cancer* **14**, 61–70 (2014).
9. Milo, R., Jorgensen, P., Moran, U., Weber, G. & Springer, M. BioNumbers—the database of key numbers in molecular and cell biology. *Nucleic Acids Res* **38**, D750–D753 (2010).
10. Weinberg, R. A. *The Biology of Cancer*. (W.W. Norton & Company, 2013).
11. Matthews, H. K., Bertoli, C. & de Bruin, R. A. M. Cell cycle control in cancer. *Nat Rev Mol Cell Biol* **23**, 74–88 (2022).
12. Alexandrov, L. B. & Stratton, M. R. Mutational signatures: the patterns of somatic mutations hidden in cancer genomes. *Curr Opin Genet Dev* **24**, 52–60 (2014).
13. Sweet-Cordero, E. A. & Biegel, J. A. The genomic landscape of pediatric cancers: Implications for diagnosis and treatment. *Science* **363**, 1170–1175 (2019).
14. Skinner, A. M. & Turker, M. S. Oxidative Mutagenesis, Mismatch Repair, and Aging. *Sci Aging Knowl Environ* **2005**, re3 (2005).

15. Charlesworth, B. Fisher, Medawar, Hamilton and the Evolution of Aging. *Genetics* **156**, 927–931 (2000).
16. Crick, F. Central Dogma of Molecular Biology. *Nature* **227**, 561–563 (1970).
17. Gröbner, S. N. *et al.* The landscape of genomic alterations across childhood cancers. *Nature* **555**, 321–327 (2018).
18. Kandoth, C. *et al.* Mutational landscape and significance across 12 major cancer types. *Nature* **502**, 333–339 (2013).
19. Kaatsch, P. Epidemiology of childhood cancer. *Cancer Treat Rev* **36**, 277–285 (2010).
20. Schwab, C. & Harrison, C. J. Advances in B-cell precursor acute lymphoblastic leukemia genomics. *Hemasphere* **2**, (2018).
21. Nowell, P. C. & Hungerford, D. A. Chromosome studies on normal and leukemic human leukocytes. *J Natl Cancer Inst* **25**, 85–109 (1960).
22. Lapidot, T. *et al.* A cell initiating human acute myeloid leukaemia after transplantation into SCID mice. *Nature* **367**, 645–648 (1994).
23. Maude, S. L. *et al.* Chimeric antigen receptor T cells for sustained remissions in leukemia. *N Engl J Med* **371**, 1507–1517 (2014).
24. van Dongen, J. J. *et al.* Prognostic value of minimal residual disease in acute lymphoblastic leukaemia in childhood. *Lancet* **352**, 1731–1738 (1998).
25. Rheingold, S. R. *et al.* Determinants of survival after first relapse of acute lymphoblastic leukemia: a Children's Oncology Group study. *Leukemia* **38**, 2382–2394 (2024).
26. Mulrooney, D. A. *et al.* The changing burden of long-term health outcomes in survivors of childhood acute lymphoblastic leukaemia: a retrospective analysis of the St Jude Lifetime Cohort Study. *Lancet Haematol* 6, e306–e316 (2019).
27. Inaba, H. & Mullighan, C. G. Pediatric acute lymphoblastic leukemia. *Haematologica* **105**, 2524–2539 (2020).
28. Brady, S. W. *et al.* The genomic landscape of pediatric acute lymphoblastic leukemia. *Nat Genet* **54**, 1376–1389 (2022).
29. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
30. Wittkopp, P. J. & Kalay, G. Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nat Rev Genet* **13**, 59–69 (2012).
31. Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
32. Huang, F. W. *et al.* Highly recurrent TERT promoter mutations in human melanoma. *Science* **339**, 957–959 (2013).

33. Horn, S. *et al. TERT* Promoter Mutations in Familial and Sporadic Melanoma. *Science* **339**, 959–961 (2013).
34. Maslow, A. H. *The Psychology of Science: A Reconnaissance*. (Harper & Row, 1966), pp. 15–16.
35. Schuster-Böckler, B. & Lehner, B. Chromatin organization is a major influence on regional mutation rates in human cancer cells. *Nature* **488**, 504–507 (2012).
36. Akdemir, K. C. *et al.* Somatic mutation distributions in cancer genomes vary with three-dimensional chromatin structure. *Nat Genet* **52**, 1178–1188 (2020).
37. Kuhn, T. *The Structure of Scientific Revolutions*. (1962).
38. Gerstein, M. B. *et al.* What is a gene, post-ENCODE? History and updated definition. *Genome Res* **17**, 669–681 (2007).
39. Morgan, T. H. *The Mechanism of Mendelian Heredity*. (Holt, New York, 1915).
40. McClintock, B. A cytological and genetical study of triploid maize. *Genetics* **14**, 180–222 (1929).
41. Watson, J. D. & Crick, F. H. C. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature* **171**, 737–738 (1953).
42. Watson, J. *The Double Helix: A Personal Account of the Discovery of the Structure of DNA*. (1968).
43. Nirenberg, M. & Leder, P. RNA Codewords and Protein Synthesis. *Science* **145**, 1399–1407 (1964).
44. Söll, D. *et al.* Studies on polynucleotides, XLIX. Stimulation of the binding of aminoacyl-sRNA's to ribosomes by ribotrinucleotides and a survey of codon assignments for 20 amino acids. *Proc Natl Acad Sci USA* **54**, 1378–1385 (1965).
45. Holley, R. W. *et al.* Structure of a Ribonucleic Acid. *Science* **147**, 1462–1465 (1965).
46. Jou, W. M., Haegeman, G., Ysebaert, M. & Fiers, W. Nucleotide Sequence of the Gene Coding for the Bacteriophage MS2 Coat Protein. *Nature* **237**, 82–88 (1972).
47. Sanger, F. *et al.* Nucleotide sequence of bacteriophage φX174 DNA. *Nature* **265**, 687–695 (1977).
48. Venter, J. C. *et al.* The Sequence of the Human Genome. *Science* **291**, 1304–1351 (2001).
49. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002).
50. Jacob, F. & Monod, J. Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol* **3**, 318–356 (1961).

51. Butler, J. E. F. & Kadonaga, J. T. The RNA polymerase II core promoter: a key component in the regulation of gene expression. *Genes Dev* **16**, 2583–2592 (2002).

52. Shiraki, T. *et al.* Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc Natl Acad Sci USA* **100**, 15776–15781 (2003).

53. Robertson, G. *et al.* Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat Methods* **4**, 651–657 (2007).

54. Patwardhan, R. P. *et al.* High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. *Nat Biotechnol* **27**, 1173–1175 (2009).

55. Kwasnieski, J. C., Mogno, I., Myers, C. A., Corbo, J. C. & Cohen, B. A. Complex effects of nucleotide variants in a mammalian *cis*-regulatory element. *Proc Natl Acad Sci USA* **109**, 19498–19503 (2012).

56. Melnikov, A. *et al.* Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat Biotechnol* **30**, 271–277 (2012).

57. Haberle, V. *et al.* Two independent transcription initiation codes overlap on vertebrate core promoters. *Nature* **507**, 381–385 (2014).

58. Carninci, P. *et al.* Genome-wide analysis of mammalian promoter architecture and evolution. *Nat Genet* **38**, 626–635 (2006).

59. Haberle, V. & Stark, A. Eukaryotic core promoters and the functional basis of transcription initiation. *Nat Rev Mol Cell Biol* **19**, 621–637 (2018).

60. Roy, A. L. & Singer, D. S. Core promoters in transcription: Old problem, new insights. *Trends Biochem Sci* **40**, 165–171 (2015).

61. Saxonov, S., Berg, P. & Brutlag, D. L. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc Natl Acad Sci USA* **103**, 1412–1417 (2006).

62. Lee, M. P. *et al.* ATG deserts define a novel core promoter subclass. *Genome Res* **15**, 1189–1197 (2005).

63. Yuan, G.-C. *et al.* Genome-Scale Identification of Nucleosome Positions in *S. cerevisiae*. *Science* **309**, 626–630 (2005).

64. Mavrich, T. N. *et al.* Nucleosome organization in the Drosophila genome. *Nature* **453**, 358–362 (2008).

65. Schones, D. E. *et al.* Dynamic Regulation of Nucleosome Positioning in the Human Genome. *Cell* **132**, 887–898 (2008).

66. Jin, C. *et al.* H3.3/H2A.Z double variant–containing nucleosomes mark 'nucleosome-free regions' of active promoters and other regulatory regions. *Nat Genet* **41**, 941–945 (2009).

67. Voigt, P., Tee, W. W. & Reinberg, D. A double take on bivalent promoters. *Genes Dev* **27**, 1318–1338 (2013).

68. Schoenfelder, S. & Fraser, P. Long-range enhancer–promoter contacts in gene expression control. *Nat Rev Genet* **20**, 437–455 (2019).

69. Lenhard, B., Sandelin, A. & Carninci, P. Metazoan promoters: Emerging characteristics and insights into transcriptional regulation. *Nat Rev Genet* **13**, 233–245 (2012).

70. Gotea, V. *et al.* Homotypic clusters of transcription factor binding sites are a key component of human promoters and enhancers. *Genome Res* **20**, 565–577 (2010).

71. Weingarten-Gabbay, S. *et al.* Systematic interrogation of human promoters. *Genome Res* **29**, 171–183 (2019).

72. Lambert, S. A. *et al.* The Human Transcription Factors. *Cell* **172**, 650–665 (2018).

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

74. Banerji, J., Olson, L. & Schaffner, W. A lymphocyte-specific cellular enhancer is located downstream of the joining region in immunoglobulin heavy chain genes. *Cell* **33**, 729–740 (1983).

75. Heintzman, N. D. *et al.* Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459**, 108–112 (2009).

76. Corces, M. R. *et al.* Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nat Genet* **48**, 1193–1203 (2016).

77. Heintzman, N. D. *et al.* Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet* **39**, 311–318 (2007).

78. Creyghton, M. P. *et al.* Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci USA* **107**, 21931–21936 (2010).

79. Arnold, M. & Stengel, K. R. Emerging insights into enhancer biology and function. *Transcription* **14**, 68–87 (2023).

80. Wang, Z. *et al.* Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat Genet* **40**, 897–903 (2008).

81. Hu, G. *et al.* H2A.Z Facilitates Access of Active and Repressive Complexes to Chromatin in Embryonic Stem Cell Self-Renewal and Differentiation. *Cell Stem Cell* **12**, 180–192 (2013).

82. Wang, Q., Carroll, J. S. & Brown, M. Spatial and Temporal Recruitment of Androgen Receptor and Its Coactivators Involves Chromosomal Looping and Polymerase Tracking. *Mol Cell* **19**, 631–642 (2005).

83. Narita, T. *et al.* Enhancers are activated by p300/CBP activity-dependent PIC assembly, RNAPII recruitment, and pause release. *Mol Cell* **81**, 2166-2182.e6 (2021).

84. Neumayr, C. *et al.* Differential cofactor dependencies define distinct types of human enhancers. *Nature* **606**, 406–413 (2022).

85. Villar, D. *et al.* Enhancer Evolution across 20 Mammalian Species. *Cell* **160**, 554–566 (2015).

86. Wong, E. S. *et al.* Deep conservation of the enhancer regulatory code in animals. *Science* **370**, (2020).

87. Payne, J. L. & Wagner, A. The Robustness and Evolvability of Transcription Factor Binding Sites. *Science* **343**, 875–877 (2014).

88. Mehta, S. *et al.* Temporal resolution of gene derepression and proteome changes upon PROTAC-mediated degradation of BCL11A protein in erythroid cells. *Cell Chem Biol* **29**, 1273-1287.e8 (2022).

89. Blackwood, E. M. & Kadonaga, J. T. Going the Distance: A Current View of Enhancer Action. *Science* **281**, 60–63 (1998).

90. Sanyal, A., Lajoie, B. R., Jain, G. & Dekker, J. The long-range interaction landscape of gene promoters. *Nature* **489**, 109–113 (2012).

91. Lettice, L. A. A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum Mol Genet* **12**, 1725–1735 (2003).

92. Deng, W. *et al.* Reactivation of Developmentally Silenced Globin Genes by Forced Chromatin Looping. *Cell* **158**, 849–860 (2014).

93. Hou, C., Zhao, H., Tanimoto, K. & Dean, A. CTCF-dependent enhancer-blocking by alternative chromatin loop formation. *Proc Natl Acad Sci USA* **105**, 20398–20403 (2008).

94. Yang, J. H. & Hansen, A. S. Enhancer selectivity in space and time: from enhancer–promoter interactions to promoter activation. *Nat Rev Mol Cell Biol* **25**, 574–591 (2024).

95. Crump, N. T. *et al.* BET inhibition disrupts transcription but retains enhancer-promoter contact. *Nat Commun* **12**, 223 (2021).

96. Calo, E. & Wysocka, J. Modification of Enhancer Chromatin: What, How, and Why? *Mol Cell* **49**, 825–837 (2013).

97. Bulger, M. & Groudine, M. Functional and Mechanistic Diversity of Distal Transcription Enhancers. *Cell* **144**, 327–339 (2011).

98. Panigrahi, A. & O'Malley, B. W. Mechanisms of enhancer action: the known and the unknown. *Genome Biol* **22**, 108 (2021).

99. Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012).

100. Li, X. & Noll, M. Compatibility between enhancers and promoters determines the transcriptional specificity of gooseberry and gooseberry neuro in the Drosophila embryo. *EMBO J* **13**, 400–406 (1994).

101. van Arensbergen, J., van Steensel, B. & Bussemaker, H. J. In search of the determinants of enhancer–promoter interaction specificity. *Trends Cell Biol* **24**, 695–702 (2014).

102. Segert, J. A., Gisselbrecht, S. S. & Bulyk, M. L. Transcriptional Silencers: Driving Gene Expression with the Brakes On. *Trends Genet* **37**, 514–527 (2021).

103. Brand, A. H., Breeden, L., Abraham, J., Sternglanz, R. & Nasmyth, K. Characterization of a "silencer" in yeast: A DNA sequence with properties opposite to those of a transcriptional enhancer. *Cell* **41**, 41–48 (1985).

104. Brand, A. H., Micklem, G. & Nasmyth, K. A yeast silencer contains sequences that can promote autonomous plasmid replication and transcriptional activation. *Cell* **51**, 709–719 (1987).

105. Gisselbrecht, S. S. *et al.* Transcriptional Silencers in Drosophila Serve a Dual Role as Transcriptional Enhancers in Alternate Cellular Contexts. *Mol Cell* 77, 324-337.e8 (2020).

106. Bandara, T. A. M. K. *et al.* A dual enhancer-silencer element, DES-K16, in mouse spermatocyte-derived GC-2spd(ts) cells. *Biochem Biophys Res Commun* **534**, 1007–1012 (2021).

107. Cao, R. *et al.* Role of Histone H3 Lysine 27 Methylation in Polycomb-Group Silencing. *Science* **298**, 1039–1043 (2002).

108. Ngan, C. Y. *et al.* Chromatin interaction analyses elucidate the roles of PRC2-bound silencers in mouse development. *Nat Genet* **52**, 264–272 (2020).

109. Huang, D., Petrykowska, H. M., Miller, B. F., Elnitski, L. & Ovcharenko, I. Identification of human silencers by correlating cross-tissue epigenetic profiles and gene expression. *Genome Res* **29**, 657–667 (2019).

110. Geyer, P. K. & Corces, V. G. DNA position-specific repression of transcription by a Drosophila zinc finger protein. *Genes Dev* **6**, 1865–1873 (1992).

111. Kellum, R. & Schedl, P. A Group of scs Elements Function as Domain Boundaries in an Enhancer-Blocking Assay. *Mol Cell Biol* **12**, 2424–2431 (1992).

112. Sun, F.-L. & Elgin, S. C. R. Putting Boundaries on Silence. *Cell* **99**, 459–462 (1999).

113. Bell, A. C., West, A. G. & Felsenfeld, G. The Protein CTCF Is Required for the Enhancer Blocking Activity of Vertebrate Insulators. *Cell* **98**, 387–396 (1999).

114. Kentepozidou, E. *et al.* Clustered CTCF binding is an evolutionary mechanism to maintain topologically associating domains. *Genome Biol* **21**, 5 (2020).

115. Phillips-Cremins, J. E. & Corces, V. G. Chromatin Insulators: Linking Genome Organization to Cellular Function. *Mol Cell* **50**, 461–474 (2013).

116. Parelho, V. *et al.* Cohesins Functionally Associate with CTCF on Mammalian Chromosome Arms. *Cell* **132**, 422–433 (2008).

117. Ortabozkoyun, H. *et al.* CRISPR and biochemical screens identify MAZ as a cofactor in CTCF-mediated insulation at Hox clusters. *Nat Genet* **54**, 202–212 (2022).

118. Kornberg, R. D. Chromatin Structure: A Repeating Unit of Histones and DNA. *Science* **184**, 868–871 (1974).

119. Olins, A. L. & Olins, D. E. Spheroid Chromatin Units (ν Bodies). *Science* **183**, 330–332 (1974).

120. Woodcock, C. L. F., Safer, J. P. & Stanchfield, J. E. Structural repeating units in chromatin. *Exp Cell Res* **97**, 101–110 (1976).

121. Finch, J. T. & Klug, A. Solenoidal model for superstructure in chromatin. *Proc Natl Acad Sci USA* **73**, 1897–1901 (1976).

122. Allfrey, V. G., Faulkner, R. & Mirsky, A. E. Acetylation and methylation of histones and their possible role in the regulation of RNA synthesis. *Proc Natl Acad Sci USA* **51**, 786–794 (1964).

123. Vidali, G., Gershey, E. L. & Allfrey, V. G. Chemical studies of histone acetylation. The distribution of epsilon-N-acetyllysine in calf thymus histones. *J Biol Chem* **243**, 6361–6 (1968).

124. Brownell, J. E. *et al.* Tetrahymena Histone Acetyltransferase A: A Homolog to Yeast Gcn5p Linking Histone Acetylation to Gene Activation. *Cell* **84**, 843–851 (1996).

125. Taunton, J., Hassig, C. A. & Schreiber, S. L. A Mammalian Histone Deacetylase Related to the Yeast Transcriptional Regulator Rpd3p. *Science* **272**, 408–411 (1996).

126. Millán-Zambrano, G., Burton, A., Bannister, A. J. & Schneider, R. Histone post-translational modifications — cause and consequence of genome function. *Nat Rev Genet* **23**, 563–580 (2022).

127. Bannister, A. J. & Kouzarides, T. Regulation of chromatin by histone modifications. *Cell Res* **21**, 381–395 (2011).

128. Strahl, B. D. & Allis, C. D. The language of covalent histone modifications. *Nature* **403**, 41–45 (2000).

129. Neumann, H. *et al.* A Method for Genetically Installing Site-Specific Acetylation in Recombinant Histones Defines the Effects of H3 K56 Acetylation. *Mol Cell* **36**, 153–163 (2009).

130. Takahashi, K. & Yamanaka, S. Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors. *Cell* **126**, 663–676 (2006).

131. Wasserman, W. W. & Sandelin, A. Applied bioinformatics for the identification of regulatory elements. *Nat Rev Genet* **5**, 276–287 (2004).

132. Vokes, S. A., Ji, H., Wong, W. H. & McMahon, A. P. A genome-scale analysis of the *cis* -regulatory circuitry underlying sonic hedgehog-mediated patterning of the mammalian limb. *Genes Dev* **22**, 2651–2663 (2008).

133. Ucar, D., Beyer, A., Parthasarathy, S. & Workman, C. T. Predicting functionality of protein–DNA interactions by integrating diverse evidence. *Bioinformatics* **25**, i137–i144 (2009).

134. Mahendrawada, L., Warfield, L., Donczew, R. & Hahn, S. Low overlap of transcription factor DNA binding and regulatory targets. *Nature* **642**, 796–804 (2025).

135. Zentner, G. E., Kasinathan, S., Xin, B., Rohs, R. & Henikoff, S. ChEC-seq kinetics discriminates transcription factor binding sites by DNA sequence and shape in vivo. *Nat Commun* **6**, 8733 (2015).

136. Giorgetti, L. *et al.* Noncooperative Interactions between Transcription Factors and Clustered DNA Binding Sites Enable Graded Transcriptional Responses to Environmental Inputs. *Mol Cell* **37**, 418–428 (2010).

137. Spitz, F. & Furlong, E. E. M. Transcription factors: from enhancer binding to developmental control. *Nat Rev Genet* **13**, 613–626 (2012).

138. Gebhardt, J. C. M. *et al.* Single-molecule imaging of transcription factor binding to DNA in live mammalian cells. *Nat Methods* **10**, 421–426 (2013).

139. Junion, G. *et al.* A Transcription Factor Collective Defines Cardiac Cell Fate and Reflects Lineage History. *Cell* **148**, 473–486 (2012).

140. Slattery, M. *et al.* Cofactor Binding Evokes Latent Differences in DNA Binding Specificity between Hox Proteins. *Cell* **147**, 1270–1282 (2011).

141. Cirillo, L. A. *et al.* Opening of Compacted Chromatin by Early Developmental Transcription Factors HNF3 (FoxA) and GATA-4. *Mol Cell* **9**, 279–289 (2002).

142. Rippe, K. *et al.* DNA sequence- and conformation-directed positioning of nucleosomes by chromatin-remodeling complexes. *Proc Natl Acad Sci USA* **104**, 15635–15640 (2007).

143. Leddin, M. *et al.* Two distinct auto-regulatory loops operate at the PU.1 locus in B cells and myeloid cells. *Blood* **117**, 2827–2838 (2011).

144. Thanos, D. & Maniatis, T. Virus induction of human IFNβ gene expression requires the assembly of an enhanceosome. *Cell* **83**, 1091–1100 (1995).

145. Cubeñas-Potts, C. & Corces, V. G. Architectural proteins, transcription, and the three-dimensional organization of the genome. *FEBS Lett* **589**, 2923–2930 (2015).

146. Gosalia, N., Neems, D., Kerschner, J. L., Kosak, S. T. & Harris, A. Architectural proteins CTCF and cohesin have distinct roles in modulating the higher order structure and expression of the CFTR locus. *Nucleic Acids Res* **42**, 9612–9622 (2014).

147. Kim, T. H. *et al.* Analysis of the Vertebrate Insulator Protein CTCF-Binding Sites in the Human Genome. *Cell* **128**, 1231–1245 (2007).

148. Li, M. *et al.* Architectural proteins for the formation and maintenance of the 3D genome. *Sci China Life Sci* **63**, 795–810 (2020).

149. Ciosk, R. *et al.* Cohesin's Binding to Chromosomes Depends on a Separate Complex Consisting of Scc2 and Scc4 Proteins. *Mol Cell* **5**, 243–254 (2000).

150. Sanborn, A. L. *et al.* Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proc Natl Acad Sci USA* **112**, (2015).

151. Rao, S. S. P. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).

152. Guo, Y. *et al.* CRISPR Inversion of CTCF Sites Alters Genome Topology and Enhancer/Promoter Function. *Cell* **162**, 900–910 (2015).

153. Rao, S. S. P. *et al.* Cohesin Loss Eliminates All Loop Domains. *Cell* **171**, 305-320.e24 (2017).

154. Nora, E. P. *et al.* Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains from Genomic Compartmentalization. *Cell* **169**, 930-944.e22 (2017).

155. Nuebler, J., Fudenberg, G., Imakaev, M., Abdennur, N. & Mirny, L. A. Chromatin organization by an interplay of loop extrusion and compartmental segregation. *Proc Natl Acad Sci USA* **115**, (2018).

156. Huang, H. *et al.* CTCF mediates dosage- and sequence-context-dependent transcriptional insulation by forming local chromatin domains. *Nat Genet* **53**, 1064–1074 (2021).

157. Zhou, Q. *et al.* ZNF143 mediates CTCF-bound promoter–enhancer loops required for murine hematopoietic stem and progenitor cell function. *Nat Commun* **12**, 43 (2021).

158. Weintraub, A. S. *et al.* YY1 Is a Structural Regulator of Enhancer-Promoter Loops. *Cell* **171**, 1573-1588.e28 (2017).

159. Cubeñas-Potts, C. *et al.* Different enhancer classes in Drosophila bind distinct architectural proteins and mediate unique chromatin interactions and 3D architecture. *Nucleic Acids Res* **45**, 1714–1730 (2017).

160. Zuin, J. *et al.* Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells. *Proc Natl Acad Sci USA* **111**, 996–1001 (2014).

161. Cremer, T. & Cremer, M. Chromosome Territories. *Cold Spring Harb Perspect Biol* **2**, a003889–a003889 (2010).

162. Allshire, R. C. & Madhani, H. D. Ten principles of heterochromatin formation and function. *Nat Rev Mol Cell Biol* **19**, 229–244 (2018).

163. Fawcett, D. W. On the occurrence of a fibrous lamina on the inner aspect of the nuclear envelope in certain cells of vertebrates. *Am J Anat* **119**, 129–145 (1966).

164. Cremer, T. *et al.* Rabl's model of the interphase chromosome arrangement tested in Chinese hamster cells by premature chromosome condensation and laser-UV-microbeam experiments. *Hum Genet* **60**, 46–56 (1982).

165. Lieberman-Aiden, E. *et al.* Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science* **326**, 289–293 (2009).

166. Finn, E. H. & Misteli, T. Molecular basis and biological function of variability in spatial genome organization. *Science* **365**, (2019).

167. Stamatoyannopoulos, J. A. *et al.* Human mutation rate associated with DNA replication timing. *Nat Genet* **41**, 393–395 (2009).

168. Nora, E. P. *et al.* Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* **485**, 381–385 (2012).

169. Rowley, M. J. & Corces, V. G. Organizational principles of 3D genome architecture. *Nat Rev Genet* **19**, 789–800 (2018).

170. Mirny, L. A., Imakaev, M. & Abdennur, N. Two major mechanisms of chromosome organization. *Curr Opin Cell Biol* **58**, 142–152 (2019).

171. Schwarzer, W. *et al.* Two independent modes of chromatin organization revealed by cohesin removal. *Nature* **551**, 51–56 (2017).

172. Lupiáñez, D. G. *et al.* Disruptions of Topological Chromatin Domains Cause Pathogenic Rewiring of Gene-Enhancer Interactions. *Cell* **161**, 1012–1025 (2015).

173. Espinola, S. M. *et al.* Cis-regulatory chromatin loops arise before TADs and gene activation, and are independent of cell fate during early Drosophila development. *Nat Genet* **53**, 477–486 (2021).

174. Misteli, T. The Self-Organizing Genome: Principles of Genome Architecture and Function. *Cell* **183**, 28–45 (2020).

175. Jost, D., Carrivain, P., Cavalli, G. & Vaillant, C. Modeling epigenome folding: formation and dynamics of topologically associated chromatin domains. *Nucleic Acids Res* **42**, 9553–9561 (2014).

176. Falk, M. *et al.* Heterochromatin drives compartmentalization of inverted and conventional nuclei. *Nature* **570**, 395–399 (2019).

177. Finn, E. H. *et al.* Extensive Heterogeneity and Intrinsic Variation in Spatial Genome Organization. *Cell* **176**, 1502-1515.e10 (2019).

178. Cattoni, D. I. *et al.* Single-cell absolute contact probability detection reveals chromosomes are organized by multiple low-frequency yet specific interactions. *Nat Commun* **8**, 1753 (2017).

179. Stevens, T. J. *et al.* 3D structures of individual mammalian genomes studied by single-cell Hi-C. *Nature* **544**, 59–64 (2017).

180. Chen, J. *et al.* Single-Molecule Dynamics of Enhanceosome Assembly in Embryonic Stem Cells. *Cell* **156**, 1274–1285 (2014).

181. Altschuler, S. J. & Wu, L. F. Cellular Heterogeneity: Do Differences Make a Difference? *Cell* **141**, 559–563 (2010).

182. Li, J., Lin, Y., Tang, Q. & Li, M. Understanding three-dimensional chromatin organization in diploid genomes. *Comput Struct Biotechnol J* **19**, 3589–3598 (2021).

183. Roix, J. J., McQueen, P. G., Munson, P. J., Parada, L. A. & Misteli, T. Spatial proximity of translocation-prone gene loci in human lymphomas. *Nat Genet* **34**, 287–291 (2003).

184. Calvi, L. M. & Link, D. C. Cellular Complexity of the Bone Marrow Hematopoietic Stem Cell Niche. *Calcif Tissue Int* **94**, 112–124 (2014).

185. Olson, O. C., Kang, Y.-A. & Passegué, E. Normal Hematopoiesis Is a Balancing Act of Self-Renewal and Regeneration. *Cold Spring Harb Perspect Med* **10**, a035519 (2020).

186. Lorenz, E., Uphoff, D., Reid, T. R. & Shelton, E. Modification of irradiation injury in mice and guinea pigs by bone marrow injections. *J Natl Cancer Inst* **12**, 197–201 (1951).

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

188. Till, J. E. & McCulloch, E. A. A direct measurement of the radiation sensitivity of normal mouse bone marrow cells. *Radiat Res* **14**, 213–22 (1961).

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

190. Abramson, S., Miller, R. & Phillips, R. The identification in adult bone marrow of pluripotent and restricted stem cells of the myeloid and lymphoid systems. *J Exp Med* **145**, 1567–1579 (1977).

191. Lemischka, I. R., Raulet, D. H. & Mulligan, R. C. Developmental potential and dynamic behavior of hematopoietic stem cells. *Cell* **45**, 917–927 (1986).

192. Spangrude, G. J., Heimfeld, S. & Weissman, I. L. Purification and Characterization of Mouse Hematopoietic Stem Cells. *Science* **241**, 58–62 (1988).

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

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

195. Liggett, L. A. & Sankaran, V. G. Unraveling Hematopoiesis through the Lens of Genomics. *Cell* **182**, 1384–1400 (2020).

196. Zhang, Y., Gao, S., Xia, J. & Liu, F. Hematopoietic Hierarchy – An Updated Roadmap. *Trends Cell Biol* **28**, 976–986 (2018).

197. Cabezas-Wallscheid, N. *et al.* Identification of Regulatory Networks in HSCs and Their Immediate Progeny via Integrated Proteome, Transcriptome, and DNA Methylome Analysis. *Cell Stem Cell* **15**, 507–522 (2014).

198. Buenrostro, J. D. *et al.* Integrated Single-Cell Analysis Maps the Continuous Regulatory Landscape of Human Hematopoietic Differentiation. *Cell* **173**, 1535-1548.e16 (2018).

199. Wilson, N. K. *et al.* Combined Single-Cell Functional and Gene Expression Analysis Resolves Heterogeneity within Stem Cell Populations. *Cell Stem Cell* **16**, 712–24 (2015).

200. Giladi, A. *et al.* Single-cell characterization of haematopoietic progenitors and their trajectories in homeostasis and perturbed haematopoiesis. *Nat Cell Biol* **20**, 836–846 (2018).

201. Dharampuriya, P. R. *et al.* Tracking the origin, development, and differentiation of hematopoietic stem cells. *Curr Opin Cell Biol* **49**, 108–115 (2017).

202. Zhang, Y. & Liu, F. The evolving views of hematopoiesis: from embryo to adulthood and from in vivo to in vitro. *J Genet Genomics* **51**, 3–15 (2024).

203. Morrison, S. J. & Scadden, D. T. The bone marrow niche for haematopoietic stem cells. *Nature* **505**, 327–334 (2014).

204. Raaijmakers, M. H. G. P. *et al.* Bone progenitor dysfunction induces myelodysplasia and secondary leukaemia. *Nature* **464**, 852–857 (2010).

205. Kode, A. *et al.* Leukaemogenesis induced by an activating β-catenin mutation in osteoblasts. *Nature* **506**, 240–244 (2014).

206. Iacobucci, I. *et al.* Multipotent lineage potential in B cell acute lymphoblastic leukemia is associated with distinct cellular origins and clinical features. *Nat Cancer* **6**, 1242–1262 (2025).

207. Garcia, C., Miller-Awe, M. D. & Witkowski, M. T. Concepts in B cell acute lymphoblastic leukemia pathogenesis. *J Leukoc Biol* **116**, 18–32 (2024).

208. Klug, C. A. *et al.* Hematopoietic stem cells and lymphoid progenitors express different Ikaros isoforms, and Ikaros is localized to heterochromatin in immature lymphocytes. *Proc Natl Acad Sci USA* **95**, 657–662 (1998).

209. Heizmann, B., Kastner, P. & Chan, S. Ikaros is absolutely required for pre-B cell differentiation by attenuating IL-7 signals. *J Exp Med* **210**, 2823–2832 (2013).

210. Gu, Z. *et al.* PAX5-driven subtypes of B-progenitor acute lymphoblastic leukemia. *Nat Genet* **51**, 296–307 (2019).

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

212. Sun, C., Chang, L. & Zhu, X. Pathogenesis of *ETV6/RUNX1* -positive childhood acute lymphoblastic leukemia and mechanisms underlying its relapse. *Oncotarget* **8**, 35445–35459 (2017).

213. Wiemels, J. *et al.* Prenatal origin of acute lymphoblastic leukaemia in children. *Lancet* **354**, 1499–1503 (1999).

214. Hjalgrim, L. L. *et al.* Presence of clone-specific markers at birth in children with acute lymphoblastic leukaemia. *Br J Cancer* **87**, 994–999 (2002).

215. Mori, H. *et al.* Chromosome translocations and covert leukemic clones are generated during normal fetal development. *Proc Natl Acad Sci USA* **99**, 8242–8247 (2002).

216. Guest, E. M. *et al.* Outstanding outcomes in infants with KMT2A-germline acute lymphoblastic leukemia treated with chemotherapy alone: results of the Children's Oncology Group AALL0631 trial. *Haematologica* **107**, 1205–1208 (2022).

217. Wei, J. *et al.* Microenvironment Determines Lineage Fate in a Human Model of MLL-AF9 Leukemia. *Cancer Cell* **13**, 483–495 (2008).

218. Friend, C. Cell-free transmission in adult Swiss mice of a disease having the character of a leukemia. *J Exp Med* **105**, 307–318 (1957).

219. Gross, L. 'Spontaneous' Leukemia Developing in G3H Mice Following Inoculation, In Infancy, with AK-Emkemic. *Exp Biol Med* **76**, 27–32 (1951).

220. Ellermann, V. & Bang, O. Experimentelle Leukämie bei Hühnern. II. *Zeitschrift für Hygiene und Infektionskrankheiten* **63**, 231–272 (1909).

221. Greaves, M. A causal mechanism for childhood acute lymphoblastic leukaemia. *Nat Rev Cancer* **18**, 471–484 (2018).

222. Greaves, M. F. Speculations on the cause of childhood acute lymphoblastic leukemia. *Leukemia* **2**, 120–5 (1988).

223. Kinlen, L. Evidence for an infective cause of childhood leukaemia: Comparison of a Scottish new town with nuclear reprocessing sites in Britain. *Lancet* **332**, 1323–1327 (1988).

224. Ford, A. M. *et al.* The TEL-AML1 leukemia fusion gene dysregulates the TGF-β pathway in early B lineage progenitor cells. *J Clin Invest* **119**, 826–836 (2009).

225. Rodríguez-Hernández, G. *et al.* Infection Exposure Promotes *ETV6-RUNX1* Precursor B-cell Leukemia via Impaired H3K4 Demethylases. *Cancer Res* 77, 4365–4377 (2017).

226. Baruchel, A. *et al.* Down syndrome and leukemia: from basic mechanisms to clinical advances. *Haematologica* **108**, 2570–2581 (2023).

227. Soulier, J. *et al.* Amplification of band q22 of chromosome 21, including AML1, in older children with acute lymphoblastic leukemia: an emerging molecular cytogenetic subgroup. *Leukemia* **17**, 1679–1682 (2003).

228. Harewood, L. *et al.* Amplification of AML1 on a duplicated chromosome 21 in acute lymphoblastic leukemia: a study of 20 cases. *Leukemia* **17**, 547–553 (2003).

229. Papaemmanuil, E. *et al.* Loci on 7p12.2, 10q21.2 and 14q11.2 are associated with risk of childhood acute lymphoblastic leukemia. *Nat Genet* **41**, 1006–1010 (2009).

230. Treviño, L. R. *et al.* Germline genomic variants associated with childhood acute lymphoblastic leukemia. *Nat Genet* **41**, 1001–1005 (2009).

231. Perez-Andreu, V. *et al.* Inherited GATA3 variants are associated with Ph-like childhood acute lymphoblastic leukemia and risk of relapse. *Nat Genet* **45**, 1494–1498 (2013).

232. Kharazmi, E. *et al.* Familial risks for childhood acute lymphocytic leukaemia in Sweden and Finland: far exceeding the effects of known germline variants. *Br J Haematol* **159**, 585–588 (2012).

233. Bennett, J. M. *et al.* Proposals for the classification of the acute leukaemias. French-American-British (FAB) co-operative group. *Br J Haematol* **33**, 451–458 (1976).

234. Duffield, A. S., Mullighan, C. G. & Borowitz, M. J. International Consensus Classification of acute lymphoblastic leukemia/lymphoma. *Virchows Archiv* **482**, 11–26 (2023).

235. Paulsson, K. & Johansson, B. High hyperdiploid childhood acute lymphoblastic leukemia. *Genes Chromosomes Cancer* **48**, 637–660 (2009).

236. Paulsson, K. *et al.* The genomic landscape of high hyperdiploid childhood acute lymphoblastic leukemia. *Nat Genet* **47**, 672–676 (2015).

237. Carroll, A. J. *et al.* Masked hypodiploidy: Hypodiploid acute lymphoblastic leukemia (ALL) mimicking hyperdiploid ALL in children: A report from the Children's Oncology Group. *Cancer Genet* **238**, 62–68 (2019).

238. Rowley, J. D. A New Consistent Chromosomal Abnormality in Chronic Myelogenous Leukaemia identified by Quinacrine Fluorescence and Giemsa Staining. *Nature* **243**, 290–293 (1973).

239. Schultz, K. R. *et al.* Long-term follow-up of imatinib in pediatric Philadelphia chromosome-positive acute lymphoblastic leukemia: Children's Oncology Group Study AALL0031. *Leukemia* **28**, 1467–1471 (2014).

240. Cario, G. *et al.* BCR-ABL1-like acute lymphoblastic leukemia in childhood and targeted therapy. *Haematologica* **105**, 2200–2204 (2020).

241. Roberts, K. G. & Mullighan, C. G. The Biology of B-Progenitor Acute Lymphoblastic Leukemia. *Cold Spring Harb Perspect Med* **10**, a034835 (2020).

242. Inaba, T. *et al.* Fusion of the Leucine Zipper Gene *HLF* to the *E2A* Gene in Human Acute B-Lineage Leukemia. *Science* **257**, 531–534 (1992).

243. Lilljebjörn, H. *et al.* Identification of ETV6-RUNX1-like and DUX4-rearranged subtypes in paediatric B-cell precursor acute lymphoblastic leukaemia. *Nat Commun* 7, 11790 (2016).

244. Yasuda, T. *et al.* Recurrent DUX4 fusions in B cell acute lymphoblastic leukemia of adolescents and young adults. *Nat Genet* **48**, 569–574 (2016).

245. Gu, Z. *et al.* Genomic analyses identify recurrent MEF2D fusions in acute lymphoblastic leukaemia. *Nat Commun* 7, 13331 (2016).

246. Hormann, F. M. *et al. NUTM1* is a recurrent fusion gene partner in B-cell precursor acute lymphoblastic leukemia associated with increased expression of genes on chromosome band 10p12.31-12.2. *Haematologica* **104**, e455–e459 (2019).

247. Hirabayashi, S. *et al. ZNF384* -related fusion genes define a subgroup of childhood B-cell precursor acute lymphoblastic leukemia with a characteristic immunotype. *Haematologica* **102**, 118–129 (2017).

248. Li, Y. *et al.* Constitutional and somatic rearrangement of chromosome 21 in acute lymphoblastic leukaemia. *Nature* **508**, 98–102 (2014).

249. Moorman, A. V. *et al.* Risk-Directed Treatment Intensification Significantly Reduces the Risk of Relapse Among Children and Adolescents With Acute Lymphoblastic Leukemia and Intrachromosomal Amplification of Chromosome 21: A Comparison of the MRC ALL97/99 and UKALL2003 Trials. *J Clin Oncol* **31**, 3389–3396 (2013).

250. Mullighan, C. G. *et al.* Deletion of *IKZF1* and Prognosis in Acute Lymphoblastic Leukemia. *N Engl J Med* **360**, 470–480 (2009).

251. Stanulla, M. *et al. IKZF1* ^{plus} Defines a New Minimal Residual Disease–Dependent Very-Poor Prognostic Profile in Pediatric B-Cell Precursor Acute Lymphoblastic Leukemia. *J Clin Oncol* **36**, 1240–1249 (2018).

252. Hayashi, H., Makimoto, A. & Yuza, Y. Treatment of Pediatric Acute Lymphoblastic Leukemia: A Historical Perspective. *Cancers (Basel)* **16**, 723 (2024).

253. Hunger, S. P. & Mullighan, C. G. Acute Lymphoblastic Leukemia in Children. *N Engl J Med* **373**, 1541–1552 (2015).

254. Pinkel, D. Five-year follow-up of 'total therapy' of childhood lymphocytic leukemia. *JAMA* **216**, 648–52 (1971).

255. Riehm, H., Gadner, H. & Welte, K. [The west-berlin therapy study of acute lymphoblastic leukemia in childhood--report after 6 years (author's transl)]. *Klin Padiatr* **189**, 89–102 (1977).

256. Smith, M. *et al.* Uniform approach to risk classification and treatment assignment for children with acute lymphoblastic leukemia. *J Clin Oncol* **14**, 18–24 (1996).

257. Meyer, C. *et al.* The MLL recombinome of acute leukemias in 2017. *Leukemia* **32**, 273–284 (2018).

258. Hutchinson, R. J. *et al.* Intensification of Therapy for Children With Lower-Risk Acute Lymphoblastic Leukemia: Long-Term Follow-Up of Patients Treated on Children's Cancer Group Trial 1881. *J Clin Oncol* **21**, 1790–1797 (2003).

259. Bhatia, S. *et al.* 6MP adherence in a multiracial cohort of children with acute lymphoblastic leukemia: a Children's Oncology Group study. *Blood* **124**, 2345–2353 (2014).

260. von Stackelberg, A. *et al.* Phase I/Phase II Study of Blinatumomab in Pediatric Patients With Relapsed/Refractory Acute Lymphoblastic Leukemia. *J Clin Oncol* **34**, 4381–4389 (2016).

261. Kantarjian, H. M. *et al.* Inotuzumab Ozogamicin versus Standard Therapy for Acute Lymphoblastic Leukemia. *N Engl J Med* **375**, 740–753 (2016).

262. Maude, S. L. *et al.* Tisagenlecleucel in Children and Young Adults with B-Cell Lymphoblastic Leukemia. *N Engl J Med* **378**, 439–448 (2018).

263. Vora, A. *et al.* Treatment reduction for children and young adults with low-risk acute lymphoblastic leukaemia defined by minimal residual disease (UKALL 2003): a randomised controlled trial. *Lancet Oncol* **14**, 199–209 (2013).

264. Pui, C.-H. *et al.* Treating Childhood Acute Lymphoblastic Leukemia without Cranial Irradiation. *N Engl J Med* **360**, 2730–2741 (2009).

265. Angiolillo, A. L. *et al.* Excellent Outcomes With Reduced Frequency of Vincristine and Dexamethasone Pulses in Standard-Risk B-Lymphoblastic Leukemia: Results From Children's Oncology Group AALL0932. *J Clin Oncol* **39**, 1437–1447 (2021).

266. Spielmann, M., Lupiáñez, D. G. & Mundlos, S. Structural variation in the 3D genome. *Nat Rev Genet* **19**, 453–467 (2018).

267. Franke, M. *et al.* Formation of new chromatin domains determines pathogenicity of genomic duplications. *Nature* **538**, 265–269 (2016).

268. Mulet-Lazaro, R. & Delwel, R. Oncogenic Enhancers in Leukemia. *Blood Cancer Discov* **5**, 303–317 (2024).

269. Herranz, D. *et al.* A NOTCH1-driven MYC enhancer promotes T cell development, transformation and acute lymphoblastic leukemia. *Nat Med* **20**, 1130–1137 (2014).

270. Kühn, M. W. M. *et al.* High-resolution genomic profiling of adult and pediatric core-binding factor acute myeloid leukemia reveals new recurrent genomic alterations. *Blood* **119**, e67–e75 (2012).

271. Radtke, I. *et al.* Genomic analysis reveals few genetic alterations in pediatric acute myeloid leukemia. *Proc Natl Acad Sci USA* **106**, 12944–12949 (2009).

272. Lancho, O. & Herranz, D. The MYC Enhancer-ome: Long-Range Transcriptional Regulation of MYC in Cancer. *Trends Cancer* **4**, 810–822 (2018).

273. Bahr, C. *et al.* A Myc enhancer cluster regulates normal and leukaemic haematopoietic stem cell hierarchies. *Nature* **553**, 515–520 (2018).

274. Hnisz, D. *et al.* Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science* **351**, 1454–1458 (2016).

275. Yang, M. *et al.* 13q12.2 deletions in acute lymphoblastic leukemia lead to upregulation of FLT3 through enhancer hijacking. *Blood* **136**, 946–956 (2020).

276. Collins, R. L. & Talkowski, M. E. Diversity and consequences of structural variation in the human genome. *Nat Rev Genet* **26**, 443–462 (2025).

277. Wang, J. *et al.* MIR retrotransposon sequences provide insulators to the human genome. *Proc Natl Acad Sci USA* **112**, E4428-37 (2015).

278. Kunarso, G. *et al.* Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nat Genet* **42**, 631–634 (2010).

279. Chuong, E. B., Elde, N. C. & Feschotte, C. Regulatory evolution of innate immunity through co-option of endogenous retroviruses. *Science* **351**, 1083–1087 (2016).

280. Mansour, M. R. *et al.* An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element. *Science* **346**, 1373–1377 (2014).

281. Sharma, N. *et al.* BCR/ABL1 and BCR are under the transcriptional control of the MYC oncogene. *Mol Cancer* **14**, 132 (2015).

282. Romana, S. *et al.* The t(12;21) of acute lymphoblastic leukemia results in a tel-AML1 gene fusion. *Blood* **85**, 3662–3670 (1995).

283. Botten, G. *et al.* Structural variation cooperates with permissive chromatin to control enhancer hijacking-mediated oncogenic transcription. *Blood* **142**, 336–351 (2023).

284. Wang, Q.-F. *et al.* MLL fusion proteins preferentially regulate a subset of wild-type MLL target genes in the leukemic genome. *Blood* **117**, 6895–905 (2011).

285. Xu, J. *et al.* MLL1 and MLL1 fusion proteins have distinct functions in regulating leukemic transcription program. *Cell Discov* **2**, 16008 (2016).

286. Crump, N. T. *et al.* MLL-AF4 cooperates with PAF1 and FACT to drive high-density enhancer interactions in leukemia. *Nat Commun* **14**, 5208 (2023).

287. Takahashi, S. & Yokoyama, A. The molecular functions of common and atypical MLL fusion protein complexes. *Biochim Biophys Acta Gene Regul Mech* **1863**, 194548 (2020).

288. Yamazaki, H. *et al.* A Remote GATA2 Hematopoietic Enhancer Drives Leukemogenesis in inv(3)(q21;q26) by Activating EVI1 Expression. *Cancer Cell* **25**, 415–427 (2014).

289. Cuykendall, T. N., Rubin, M. A. & Khurana, E. Non-coding genetic variation in cancer. *Curr Opin Syst Biol* **1**, 9–15 (2017).

290. Anna, A. & Monika, G. Splicing mutations in human genetic disorders: examples, detection, and confirmation. *J Appl Genet* **59**, 253–268 (2018).

291. Rasmussen, K. D. *et al.* Loss of TET2 in hematopoietic cells leads to DNA hypermethylation of active enhancers and induction of leukemogenesis. *Genes Dev* **29**, 910–22 (2015).

292. Smith, J. S. *et al.* Chronic loss of STAG2 leads to altered chromatin structure contributing to de-regulated transcription in AML. *J Transl Med* **18**, 339 (2020).

293. Lu, Y. *et al.* Advances in Whole Genome Sequencing: Methods, Tools, and Applications in Population Genomics. *Int J Mol Sci* **26**, 372 (2025).

294. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).

295. Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol* **31**, 213–219 (2013).

296. Fan, Y. *et al.* MuSE: accounting for tumor heterogeneity using a sample-specific error model improves sensitivity and specificity in mutation calling from sequencing data. *Genome Biol* **17**, 178 (2016).

297. Chen, X. *et al.* Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* **32**, 1220–1222 (2016).

298. Cameron, D. L. *et al.* GRIDSS2: comprehensive characterisation of somatic structural variation using single breakend variants and structural variant phasing. *Genome Biol* **22**, 202 (2021).

299. Van der Auwera, G. A. *et al.* From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. *Curr Protoc Bioinformatics* **43**, (2013).

300. Díaz-Gay, M. *et al.* Assigning mutational signatures to individual samples and individual somatic mutations with SigProfilerAssignment. *Bioinformatics* **39**, (2023).

301. LaFramboise, T. Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. *Nucleic Acids Res* **37**, 4181–4193 (2009).

302. Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603–607 (2012).

303. Wang, X., Luan, Y. & Yue, F. EagleC: A deep-learning framework for detecting a full range of structural variations from bulk and single-cell contact maps. *Sci Adv* **8**, (2022).

304. Stark, R., Grzelak, M. & Hadfield, J. RNA sequencing: the teenage years. *Nat Rev Genet* **20**, 631–656 (2019).

305. Park, P. J. ChIP–seq: advantages and challenges of a maturing technology. *Nat Rev Genet* **10**, 669–680 (2009).

306. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* **10**, 1213–1218 (2013).

307. Barnett, K. R. *et al.* Epigenomic mapping reveals distinct B cell acute lymphoblastic leukemia chromatin architectures and regulators. *Cell Genomics* **3**, (2023).

308. Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing Chromosome Conformation. *Science* **295**, 1306–1311 (2002).

309. Denker, A. & de Laat, W. The second decade of 3C technologies: detailed insights into nuclear organization. *Genes Dev* **30**, 1357–1382 (2016).

310. Belaghzal, H., Dekker, J. & Gibcus, J. H. Hi-C 2.0: An optimized Hi-C procedure for high-resolution genome-wide mapping of chromosome conformation. *Methods* **123**, 56–65 (2017).

311. Hsieh, T.-H. S. *et al.* Mapping Nucleosome Resolution Chromosome Folding in Yeast by Micro-C. *Cell* **162**, 108–119 (2015).

312. Mifsud, B. *et al.* Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nat Genet* **47**, 598–606 (2015).

313. Xu, J. *et al.* Subtype-specific 3D genome alteration in acute myeloid leukaemia. *Nature* **611**, 387–398 (2022).

314. de Wet, J. R., Wood, K. V., DeLuca, M., Helinski, D. R. & Subramani, S. Firefly Luciferase Gene: Structure and Expression in Mammalian Cells. *Mol Cell Biol* **7**, 725–737 (1987).

315. Doudna, J. A. & Charpentier, E. The new frontier of genome engineering with CRISPR-Cas9. *Science* **346**, (2014).

316. McKinnon, K. M. Flow Cytometry: An Overview. *Curr Protoc Immunol* **120**, 5.1.1-5.1.11 (2018).

317. Mullis, K. *et al.* Specific Enzymatic Amplification of DNA In Vitro: The Polymerase Chain Reaction. *Cold Spring Harb Symp Quant Biol* **51**, 263–273 (1986).

318. Heid, C. A., Stevens, J., Livak, K. J. & Williams, P. M. Real time quantitative PCR. *Genome Res* **6**, 986–994 (1996).

319. Terrault, N. A. *et al.* Update on prevention, diagnosis, and treatment of chronic hepatitis B. *Hepatology* **67**, 1560–1599 (2018).

320. Mellors, J. W. *et al.* Prognosis in HIV-1 Infection Predicted by the Quantity of Virus in Plasma. *Science* **272**, 1167–1170 (1996).

321. Aydın, E. *et al.* Discovery of Cis-Regulatory Mechanisms via Non-Coding Mutations in Acute Lymphoblastic Leukemia. *Genes Chromosomes Cancer* **64**, (2025).

322. Rachakonda, P. S. *et al. TERT* promoter mutations in bladder cancer affect patient survival and disease recurrence through modification by a common polymorphism. *Proc Natl Acad Sci USA* **110**, 17426–17431 (2013).

323. Nault, J. C. *et al.* High frequency of telomerase reverse-transcriptase promoter somatic mutations in hepatocellular carcinoma and preneoplastic lesions. *Nat Commun* **4**, 2218 (2013).

324. Fredriksson, N. J., Ny, L., Nilsson, J. A. & Larsson, E. Systematic analysis of noncoding somatic mutations and gene expression alterations across 14 tumor types. *Nat Genet* **46**, 1258–1263 (2014).

325. Rheinbay, E. *et al.* Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. *Nature* **578**, 102–111 (2020).

326. Dietlein, F. *et al.* Genome-wide analysis of somatic noncoding mutation patterns in cancer. *Science* **376**, (2022).

327. Osterwalder, M. *et al.* Enhancer redundancy provides phenotypic robustness in mammalian development. *Nature* **554**, 239–243 (2018).

328. Wu, K. *et al.* Exploring noncoding variants in genetic diseases: from detection to functional insights. *J Genet Genomics* **51**, 111–132 (2024).

329. Castro, C. P., Diehl, A. G. & Boyle, A. P. Challenges in screening for de novo noncoding variants contributing to genetically complex phenotypes. *Human Genetics and Genomics Advances* **4**, 100210 (2023).

330. Roe, J.-S., Mercan, F., Rivera, K., Pappin, D. J. & Vakoc, C. R. BET Bromodomain Inhibition Suppresses the Function of Hematopoietic Transcription Factors in Acute Myeloid Leukemia. *Mol Cell* **58**, 1028–1039 (2015).

331. Zuber, J. *et al.* RNAi screen identifies Brd4 as a therapeutic target in acute myeloid leukaemia. *Nature* **478**, 524–528 (2011).

332. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–329 (2015).

333. Andersson, J. *et al.* Characterizing the allele-specific gene expression landscape in high hyperdiploid acute lymphoblastic leukemia with BASE. *Sci Rep* **14**, 23181 (2024).

334. Chatsirisupachai, K. & de Magalhães, J. P. Somatic mutations in human ageing: New insights from DNA sequencing and inherited mutations. *Ageing Res Rev* **96**, 102268 (2024).

335. Bielski, C. M. *et al.* Widespread Selection for Oncogenic Mutant Allele Imbalance in Cancer. *Cancer Cell* **34**, 852-862.e4 (2018).

336. Mayba, O. *et al.* MBASED: allele-specific expression detection in cancer tissues and cell lines. *Genome Biol* **15**, 405 (2014).

337. Zambelli, F. *et al.* aScan: A Novel Method for the Study of Allele Specific Expression in Single Individuals. *J Mol Biol* **433**, 166829 (2021).

338. Harvey, C. T. *et al.* QuASAR: quantitative allele-specific analysis of reads. *Bioinformatics* **31**, 1235–42 (2015).

339. Moura-Castro, L. H. *et al.* The 3D genome of pediatric B-cell precursor acute lymphoblastic leukemia. Preprint at https://doi.org/10.1101/2025.07.22.666073 (2025).

340. Gavrilov, A. A. *et al.* Elementary 3D organization of active and silenced E. coli genome. *Nature* **645**, 1060–1070 (2025).

341. Goel, V. Y., Huseyin, M. K. & Hansen, A. S. Region Capture Micro-C reveals coalescence of enhancers and promoters into nested microcompartments. *Nat Genet* **55**, 1048–1056 (2023).

342. Yang, L. *et al.* 3D genome alterations associated with dysregulated HOXA13 expression in high-risk T-lineage acute lymphoblastic leukemia. *Nat Commun* **12**, 3708 (2021).

343. Hawley, J. R. *et al.* Reorganization of the 3D Genome Pinpoints Noncoding Drivers of Primary Prostate Tumors. *Cancer Res* **81**, 5833–5848 (2021).

344. Johnstone, S. E. *et al.* Large-Scale Topological Changes Restrain Malignant Progression in Colorectal Cancer. *Cell* **182**, 1474-1489.e23 (2020).

345. Narang, S. *et al.* Clonal evolution of the 3D chromatin landscape in patients with relapsed pediatric B-cell acute lymphoblastic leukemia. *Nat Commun* **15**, 7425 (2024).

346. Smith, A. *et al.* Enhancer heterogeneity in acute lymphoblastic leukemia drives differential gene expression in patients. *Blood* (2025). https://doi.org/10.1182/blood.2024028019

347. Yang, M. *et al.* Proteogenomics and Hi-C reveal transcriptional dysregulation in high hyperdiploid childhood acute lymphoblastic leukemia. *Nat Commun* **10**, 1519 (2019).

348. Martin, E. W. *et al.* Dynamics of Chromatin Accessibility During Hematopoietic Stem Cell Differentiation Into Progressively Lineage-Committed Progeny. *Stem Cells* **41**, 520–539 (2023).

349. Gao, T. & Qian, J. EnhancerAtlas 2.0: An updated resource with enhancer annotation in 586 tissue/cell types across nine species. *Nucleic Acids Res* **48**, D58–D64 (2020).

350. Demircioğlu, D. *et al.* A Pan-cancer Transcriptome Analysis Reveals Pervasive Regulation through Alternative Promoters. *Cell* **178**, 1465-1477.e17 (2019).

351. Yost, K. E. *et al.* Three-dimensional genome landscape of primary human cancers. *Nat Genet* **57**, 1189–1200 (2025).

352. Zhao, S. G. *et al.* Integrated analyses highlight interactions between the three-dimensional genome and DNA, RNA and epigenomic alterations in metastatic prostate cancer. *Nat Genet* **56**, 1689–1700 (2024).

353. Studd, J. B. *et al.* Cancer drivers and clonal dynamics in acute lymphoblastic leukaemia subtypes. *Blood Cancer J* **11**, 177 (2021).

354. Singh, V. K., Rastogi, A., Hu, X., Wang, Y. & De, S. Mutational signature SBS8 predominantly arises due to late replication errors in cancer. *Commun Biol* **3**, 421 (2020).

355. Yaacov, A., Rosenberg, S. & Simon, I. Mutational signatures association with replication timing in normal cells reveals similarities and differences with matched cancer tissues. *Sci Rep* **13**, 7833 (2023).

356. Kazanov, M. D. *et al.* APOBEC-Induced Cancer Mutations Are Uniquely Enriched in Early-Replicating, Gene-Dense, and Active Chromatin Regions. *Cell Rep* **13**, 1103–1109 (2015).

357. Otlu, B. *et al.* Topography of mutational signatures in human cancer. *Cell Rep* **42**, 112930 (2023).

358. Verbeke, D. *et al.* The XPO1 Inhibitor KPT-8602 Synergizes with Dexamethasone in Acute Lymphoblastic Leukemia. *Clin Cancer Res* **26**, 5747–5758 (2020).

359. Shao, X. *et al.* Transcriptional regulatory program controlled by MYB in T-cell acute lymphoblastic leukemia. *Leukemia* **38**, 2573–2584 (2024).

360. Suzuki, K. *et al. MEF2D - BCL9* Fusion Gene Is Associated With High-Risk Acute B-Cell Precursor Lymphoblastic Leukemia in Adolescents. *J Clin Oncol* **34**, 3451–3459 (2016).

361. Duy, C. *et al.* BCL6 enables Ph+ acute lymphoblastic leukaemia cells to survive BCR-ABL1 kinase inhibition. *Nature* **473**, 384–8 (2011).

362. Zhu, Y. *et al.* Targeting PRMT1-mediated FLT3 methylation disrupts maintenance of MLL-rearranged acute lymphoblastic leukemia. *Blood* **134**, 1257–1268 (2019).

363. Toma, M. M. & Skorski, T. Star wars against leukemia: attacking the clones. *Leukemia* **38**, 2293–2302 (2024).

364. Mengxuan, S., Fen, Z. & Runming, J. Novel Treatments for Pediatric Relapsed or Refractory Acute B-Cell Lineage Lymphoblastic Leukemia: Precision Medicine Era. *Front Pediatr* **10**, (2022).

365. Jost, E. *et al.* Epimutations mimic genomic mutations of DNMT3A in acute myeloid leukemia. *Leukemia* **28**, 1227–1234 (2014).

366. Brown, R., Curry, E., Magnani, L., Wilhelm-Benartzi, C. S. & Borley, J. Poised epigenetic states and acquired drug resistance in cancer. *Nat Rev Cancer* **14**, 747–753 (2014).

367. Batut, P. J. *et al.* Genome organization controls transcriptional dynamics during development. *Science* **375**, 566–570 (2022).

368. Bower, G. *et al.* Range extender mediates long-distance enhancer activity. *Nature* **643**, 830–838 (2025).

369. Crump, N. T. & Milne, T. A. Is Enhancer Function Driven by Protein–Protein Interactions? From Bacteria to Leukemia. *BioEssays* **47**, (2025).

370. Qu, Z., Hu, G., Garfinkel, A. & Weiss, J. N. Nonlinear and stochastic dynamics in the heart. *Phys Rep* **543**, 61–162 (2014).

371. Edge, P., Bafna, V. & Bansal, V. HapCUT2: Robust and accurate haplotype assembly for diverse sequencing technologies. *Genome Res* **27**, 801–812 (2017).

372. Ye, T. & Ma, W. ASHIC: hierarchical Bayesian modeling of diploid chromatin contacts and structures. *Nucleic Acids Res* **48**, e123–e123 (2020).

373. Tan, L., Xing, D., Chang, C.-H., Li, H. & Xie, X. S. Three-dimensional genome structures of single diploid human cells. *Science* **361**, 924–928 (2018).

374. Chai, H. *et al.* Tri-omic single-cell mapping of the 3D epigenome and transcriptome in whole mouse brains throughout the lifespan. *Nat Methods* **22**, 994–1007 (2025).

375. Wen, X. *et al.* Single-cell multiplex chromatin and RNA interactions in ageing human brain. *Nature* **628**, 648–656 (2024).

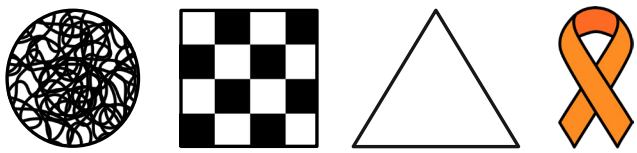
376. Chai, H. & Ruan, Y. 3D genome sequencing technology in its mid-teens: past, present, and future. *Trends Genet* (2025). https://doi.org/10.1016/j.tig.2025.07.010

377. Beagrie, R. A. *et al.* Complex multi-enhancer contacts captured by genome architecture mapping. *Nature* **543**, 519–524 (2017).

378. Quinodoz, S. A. *et al.* Higher-Order Inter-chromosomal Hubs Shape 3D Genome Organization in the Nucleus. *Cell* **174**, 744-757.e24 (2018).

379. Dekker, J. *et al.* Spatial and temporal organization of the genome: Current state and future aims of the 4D nucleome project. *Mol Cell* **83**, 2624–2640 (2023).

Printed by Media-Tryck, Lund 2025 NORDIC SWAN ECOLABEL 3041 0903



"It turns out that an eerie type of chaos can lurk just behind a facade of order—and yet, deep inside the chaos lurks an even eerier kind of order"

-Douglas Hofstadter-

LUND UNIVERSITY

FACULTY OF MEDICINE

Division of Clinical Genetics
Department of Laboratory Medicine

Lund University, Faculty of Medicine
Doctoral Dissertation Series 2025:130
ISBN 978-91-8021-783-5
ISSN 1652-8220