Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine 2179

Deciphering the molecular landscape of childhood acute lymphoblastic leukemia through multiomics data integration

OLGA KRALI

ACTA UNIVERSITATIS
UPSALIENSIS
2025

ISSN 1651-6206
ISBN 978-91-513-2588-0
urn:nbn:se:uu:diva-563468

UPPSALA
UNIVERSITET

Dissertation presented at Uppsala University to be publicly examined in Triple room Navet, Uppsala Biomedicinska Centrum (BMC), Husargatan 3, entrance C11, Uppsala, Friday, 24 October 2025 at 09:15 for the degree of Doctor of Philosophy (Faculty of Medicine). The examination will be conducted in English. Faculty examiner: Professor Jonathan Bond (University College Dublin, School of Medicine, Systems Biology, Ireland).

Abstract
Krali, O. 2025. Deciphering the molecular landscape of childhood acute lymphoblastic leukemia through multiomics data integration. *Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine* 2179. 61 pp. Uppsala: Acta Universitatis Upsaliensis. ISBN 978-91-513-2588-0.

Pediatric acute lymphoblastic leukemia (ALL), the most common childhood cancer, is characterized by aberrant lymphopoiesis that disrupts normal hematopoiesis in the bone marrow. Diverse genetic drivers, and epigenetic and transcriptomic alterations contribute to disease heterogeneity. By applying standard clinical diagnostic procedures and next generation sequencing (NGS) methods, 22 B-cell precursor ALL (BPC-ALL) and 15 T-cell ALL (T-ALL) subtypes have been identified. The aim of this thesis is to integrate molecular, clinical, and *ex vivo* drug response data from pediatric ALL patients diagnosed and treated in the Nordic countries to identify cross-modal interactions underlying therapy responses and clinical outcomes. Ultimately, this work aims to provide insights that support treatment optimization, with the dual goal of preventing relapses and reducing toxicity. The datasets included DNA methylation (DNAm), gene expression (GEX), somatic single nucleotide variants (SNVs), fusion genes, copy number alterations (CNA), *ex vivo* drug response (EVDR) data, and clinical variables from over 1,000 patients. Bioinformatic approaches included supervised classification, survival analysis and modelling, differential DNAm and GEX analyses, and data integration.

In paper I, we built ALLIUM, a multimodal machine learning classifier that predicts 17 ALL subtypes using either DNAm or GEX data. Applied to 1,131 ALL patients, ALLIUM successfully subtyped previously unclassified cases, usually aligning with existent molecular findings. **In paper II**, we designed two DNAm-based risk prediction models, the relapse risk predictor (RRP) and the mortality risk predictor (MRP), which outperformed the traditional risk stratification models and generalized well to independent cohorts. **In paper III**, we integrated EVDR data with DNAm and GEX data and found pre-treatment epigenetic and transcriptomic differences between *ex vivo* resistant and sensitive BCP-ALL patients. The resulting signatures were enriched in pathways commonly altered in cancer. **In paper IV**, we utilized a framework to integrate DNAm, GEX, SNV, and EVDR data from 1,231 BCP-ALL patients. We identified cross-modal signatures and associations, reducing feature dimensionality and enabling pathway, correlation, and network analyses. These signatures improved risk modelling beyond standard clinical stratification.

Together, these studies enhance our understanding of the molecular underpinnings of ALL and may form a basis for supporting future clinical decision-making, particularly in ambiguous cases.

Keywords: pediatric acute lymphoblastic leukemia, DNA methylation, gene expression, *ex vivo* drug responses, bioinformatics, machine learning, data integration

Olga Krali, Science for Life Laboratory, SciLifeLab, Box 256, Uppsala University, SE-75105 Uppsala, Sweden. Department of Medical Sciences, Molecular Precision Medicine, Box 1432, BMC, Uppsala University, SE-751 44 Uppsala, Sweden.

ISSN 1651-6206
ISBN 978-91-513-2588-0
URN urn:nbn:se:uu:diva-563468 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-563468)

To all the little cancer warriors and their families.
Life has not been easy for you, yet you teach us how to fight.
Great lessons are given even when a school class is not involved.

To my parents who have been my safe place.
To my husband who has been my light in the dark.
To my new family who has loved me unconditionally.
To my research group who became a place to call home.

List of Papers

This thesis is based on the following papers, which are referred to in the text by their Roman numerals.

I. **Krali, O.**, Marincevic-Zuniga, Y., Arvidsson, G., Enblad, AP., Lundmark, A., Sayyab, S., Zachariadis, V., Heinäniemi, M., Suhonen, J., Oksa, L., Vepsäläinen, K., Öfverholm, I., Barbany, G., Nordgren, A., Lilljebjörn, H., Fioretos, T., Madsen, HO., Marquart, HV., Flaegstad, T., Forestier, E., Jónsson, ÓG., Kanerva, J., Lohi, O., Norén-Nyström, U., Schmiegelow, K., Harila, A., Heyman, M., Lönnerholm, G., Syvänen, AC. & Nordlund, J[+]. (2023) Multimodal classification of molecular subtypes in pediatric acute lymphoblastic leukemia. *npj Precis. Onc.*, 7, 131.

II. Mosquera Orgueira, A.[*+], **Krali, O.**[*], Pérez Míguez, C., Peleteiro Raíndo, A., Díaz Arias, JÁ., González Pérez, MS., Pérez Encinas, MM., Fernández Sanmartín, M., Sinnet, D., Heyman, M., Lönnerholm, G., Norén-Nyström, U., Schmiegelow, K. & Nordlund, J. (2024) Refining risk prediction in pediatric acute lymphoblastic leukemia through DNA methylation profiling. *Clin Epigenet*, 16, 49.

III. Enblad, AP., **Krali, O**, Gezelius, H., Lundmark, A., Blom, K., Andersson, C., Palle, J., Frost, BM., Ryhänen, S., Flægstad, T., Jónsson, ÓG., Schmiegelow, K., Heyman, M., Harila, A., Nygren, P., Larsson, R., Lönnerholm, G. & Nordlund, J[+]. (2025) Ex Vivo Drug Responses and Molecular Profiles of 597 Pediatric Acute Lymphoblastic Leukemia Patients. *HemaSphere*, 9, e70176.

IV. **Krali, O.**, Enblad, AP., Sulyaeva, J., Gogishvili, D., Lundmark, A., Harila, A., Andersson, C., Erkers, T., Heinäniemi, M., Lönnerholm, G. & Nordlund, J[+]. An integrative molecular map of pediatric B-cell precursor acute lymphoblastic leukemia. *Manuscript under revision*.

*Equal contribution to this project
+Corresponding author

Other papers by the author

I. Gabbutt, C. [*+], Duran-Ferrer, M.[*], Grant, H., Mallo, D., Nadeu, F., Househam, J., Villamor, N., Müller, M., Heath, S., Raineri, E., **Krali, O.**, Nordlund, J., Zenz, T., Gut, IG., Campo, E., Lopez-Guillermo, A., Fitzgibbon, J., Barnes, CP., Shibata, D., Martin-Subero, JI.[#+] & Graham, A.[#+]. (2025) Fluctuating DNA methylation tracks cancer evolution at clinical scale. *Nature, In Press*.

II. Marchi, F., Shastri, VM., Marrero, RJ., Nguyen, NHK., Öttl, A., Schade, AK., Landwehr, M., **Krali, O.**, Nordlund, J., Ghavami, M., Sckaff, F., Mansinghka, VK., Cao, X., Slayton, W., Starostik, P., Cogle, CR., Ribeiro, RC., Rubnitz, JE., Klco, J., Elsayed, A., Gamis, AS., Triche, TJ., Ries, R., Kolb, EA., Aplenc, R., Alonzo, T., Pounds, S., Meshinchi, S. & Lamba, JK [+]. (2025) Epigenomic diagnosis and prognosis of Acute Myeloid Leukemia. *Nature Communications*, 16, 6961.

III. Oksa, L.[*+], Moisio, S.[*], Maqbool, K., Kramer, R., Nikkilä, A., Jayasingha, B., Mäkinen, A., Foroughi-Asl, H., Rounioja, S., Suhonen, J., **Krali, O.**, Voutilainen, M., Lahnalampi, M., Vepsäläinen, K., Huang, S., Duque-Afonso, J., Hauer, J., Nordlund, J., Wirta, V., Lohi, O.[#+] & Heinäniemi, M[#+]. (2025) Genomic determinants of therapy response in *ETV6*::*RUNX1* leukemia. *Leukemia*, 39, 2125–2139.

IV. Lysenkova Wiklander, M., Zachariah, D., **Krali, O.** & Nordlund, J[+]. (2025) Error reduction in leukemia machine learning classification with conformal prediction. *JCO Clin. Cancer Inform.* e2400324.

V. Ghantous, A.[*+], Gonseth Nusslé, S.[*], Nassar, FJ.[*], Spitz, N., Novoloaca, A., **Krali, O**, Nickels, E., Cahais, V., Cuenin, C., Roy, R., Li, S., Caron, M., Lam, D., Fransquet, PD., Casement, J., Strathdee, G., Pearce, MS., Hansen, HM., Lee, HH., Lee, YS., de Smith, AJ., Sinnett, D., Håberg, SE., McKay, JA., Nordlund, J., Magnus, P., Dwyer, T., Saffery, R., Wiemels, JL.[#], Munthe-Kaas, MC.[#] & Herceg, Z[#+]. (2024) Epigenome-wide analysis across the development span of pediatric acute lymphoblastic leukemia: backtracking to birth. *Mol Cancer*, 23, 238.

VI. **Krali, O.*** , Palle, J.*+, Bäcklin, CL., Abrahamsson, J., Norén-Nyström, U., Hasle, H., Jahnukainen, K., Jónsson, ÓG., Hovland, R., Lausen, B., Larsson, R., Palmqvist, L., Staffas, A., Zeller, B. & Nordlund, J+. (2021) DNA Methylation Signatures Predict Cytogenetic Subtype and Outcome in Pediatric Acute Myeloid Leukemia (AML). *Genes*, 12(6), 895.

*Equal contribution to this project
# These authors jointly supervised this work
+Corresponding author

Supervisors and Examination Committee

Main Supervisor	***Associate Professor Jessica Nordlund*** Dept. of Medical Sciences, Molecular Precision Medicine, Uppsala University
Co-supervisors	***Professor Arja Harila*** Dept. of Women's and Children's Health, Pediatric Oncological and Neurological Research, Uppsala University
	Professor Merja Heinäniemi Institute of Biomedicine, School of Medicine, Faculty of Health Sciences, University of Eastern Finland
Examiner	***Professor Tove Fall*** Dept. of Medical Sciences, Molecular Epidemiology, Uppsala University
Opponent	***Professor Jonathan Bond*** School of Medicine, University College Dublin
Evaluation Committee	***Associate Professor Jordi Carreras Puigvert*** Dept. of Pharmaceutical Biosciences, Pharmaceutical Bioinformatics, Uppsala University
	Doctor Anna Robelius Dept. of Medical Sciences, Hematology, Uppsala University
	Associate Professor Nick Tobin Dept. of Oncology and Pathology, Karolinska Institutet and University Hospital & Breast Center, Karolinska Comprehensive Cancer Center, Karolinska University Hospital

Contents

Abbreviations

AI	Artificial Intelligence
ALL	Acute lymphoblastic leukemia
ALLIUM	Acute lymphoblastic leukemia subtype identification using machine learning
AML	Acute myeloid leukemia
BCP-ALL	B-cell precursor acute lymphoblastic leukemia
BM	Bone marrow
CGI	CpG island
c-index	Concordance index
CIMP	CpG island methylator phenotype
CME	Cross-modal element
CNA	Copy number alteration
CpG	Cytosine-phosphate-guanine
DEA	Differential expression analysis
DEG	Differentially expressed gene
DL	Deep learning
DMA	Differential methylation analysis
DMC	Differentially methylated CpG site
DNAm	DNA methylation
DNMT	DNA methyltransferase
ETP ALL	Early T cell precursor acute lymphoblastic leukemia
EVDR	*Ex vivo* drug response
FMCA	Fluorometric microculture cytotoxicity assay
GEX	Gene expression
HeH	High hyperdiploidy
HSC	Hematopoietic stem cell
HSCT	Hematopoietic stem cell transplantation
iAMP21	Intrachromosomal amplification of chromosome 21
ICC	International Consensus Classification
IC_{25}	Inhibitory concentration 25
LC_{50}	Lethal concentration 50
log_2FC	log2 fold change
ML	Machine Learning

MOFA	Multi-Omics Factor Analysis
MSC	Mesenchymal stem cell
MTT	3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyl tetrazolium bromide
MRD	Minimal residual disease
mRNA	messenger RNA
MRP	Mortality risk predictor
ncRNA	non-coding RNA
NGS	Next-generation sequencing
NOPHO	Nordic Society of Pediatric Oncology and Hematology
NSC	Nearest Shrunken Centroid
OS	Overall survival
PB	Peripheral blood
PCA	Principal component analysis
RFS	Relapse free survival
RNA-seq	RNA sequencing
RRP	Relapse risk predictor
scRNA-seq	Single-cell RNA sequencing
SI	Survival index
SNV	Single nucleotide variant
T-ALL	T-cell acute lymphoblastic leukemia
TK	Tyrosine kinase
TKI	Tyrosine kinase inhibitor
UMAP	Uniform Manifold Approximation and Projection for Dimension Reduction
WBC	White blood cell
WGBS	Whole genome bisulfite sequencing
WES	Whole exome sequencing
WGS	Whole genome sequencing

Introduction

Childhood leukemia

Acute lymphoblastic leukemia (ALL) is the most common malignancy in children worldwide[1], accounting for approximately 80% of all childhood leukemia cases[2]. It is characterized by the uncontrollable growth of immature lymphoid cells, known as lymphoblasts, in the bone marrow. This expansion disrupts normal lymphopoiesis (**Figure 1**), and impairs overall hematopoiesis, reducing the production of other essential blood cells and leading to a range of pathological conditions, such as severe infections, anemia, and bleeding. As a fast-progressing disease, ALL leads to dismal outcomes if not treated in time. Based on the lineage of the affected cells, ALL is classified into B-cell precursor ALL (BCP-ALL) and T-cell ALL (T-ALL). ALL is diagnosed using immunophenotyping (e.g., flow cytometry) on bone marrow (BM) aspirates or peripheral blood (PB) samples. In contrast to healthy individuals, who have <5% of blasts in the BM and no blasts in the PB, ALL patients have >20% of lymphoblasts in the BM.

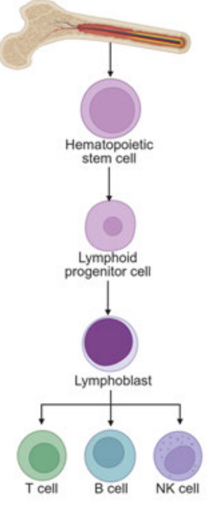


Figure 1. Lymphopoiesis in the bone marrow (created with Biorender.com).

In the Nordic countries, approximately 200 children are diagnosed with ALL every year[3,4]. BCP-ALL is the most common immunophenotype accounting for ~85-90% of all ALL cases. Despite the high survival rates (~90%), the chances of cure decrease significantly after relapse to approximately 50%[4–6]. In addition, T-ALL, which is more aggressive and until recently less characterized[7–9], requires refined risk stratification to achieve similar survival rates as BCP-ALL[10,11]. Except for relapse, patients with ALL are at substantial risk of experiencing treatment-related toxicities and adverse effects, which can be immediately life-threatening or lead to serious health issues later in life.

The Nordic countries were among the first to introduce treatment protocols including risk-based stratification into treatment groups[3,4,12,13]. Specifically, ALL patients treated according to the Nordic Society of Pediatric Hematology and Oncology (NOPHO) protocols are stratified into risk groups, based on white blood cell (WBC) count, age at diagnosis, and cytogenetic subtype. In the NOPHO ALL-2008 protocol[4], minimal residual disease (MRD) was incorporated as a clinical variable used for treatment stratification. Multiple studies highlight MRD at the end of induction as a prognostic marker for ALL[14,15]. The treatment is risk adapted and involves the administration of several types of chemotherapeutic agents in combination with glucocorticoids and in some selected cases targeted agents. For some high-risk patients, hematopoietic stem cell transplantation (HSCT) is an important option for achieving lasting remission. Immunotherapy, such as chimeric antigen receptor T cell or blinatumomab (bispecific T-cell engager), may also be administered to children with refractory or relapsed disease[16]. A recent phase 3 trial demonstrated that blinatumomab combined with induction chemotherapy improved disease-free survival in children compared to those who received chemotherapy alone[17]. The aim of the treatment is to quickly eliminate leukemic cells from the body and prevent recurrent disease. Although these intensive treatment protocols have resulted in significant improvements in survival rates, this progress is accompanied by serious toxicities and other effects that impact the overall health and quality of life of pediatric leukemia survivors[18]. To address these challenges, new initiatives aim to reduce overtreatment in patients at particularly low risk of relapse, and thereby minimizing toxicity, while simultaneously enhancing treatment strategies for improving outcomes for high-risk patients, who currently face poor outcomes. Achieving these goals requires large-scale international collaborations, such as the ongoing European ALL protocol, ALLTogether[19], which enhances statistical power for studying rare subtypes. Additionally, analyzing retrospective cohorts remains a valuable tool for identifying ways to optimize treatment.

Childhood ALL subtypes

Most of the pediatric leukemia subtypes are characterized by large-scale chromosomal alterations, including missing or extra chromosomal copies, intrachromosomal amplifications, or translocations, which may lead to expressed fusion genes (**Figure 2**). Such recurrent aberrations are subtype defining in ALL and have been widely used to aid treatment stratification and monitoring of therapy response[20,21]. Cytogenetic analyses were previously performed with the help of Giemsa banding karyotyping, fluorescence in situ hybridization, and reverse transcription polymerase chain reaction[22,23]. However, many patients carry non-recurrent aberrations or complex karyotypes, which pose challenges for interpreting molecular genetic subtypes with the traditional clinical diagnostic methods[24].

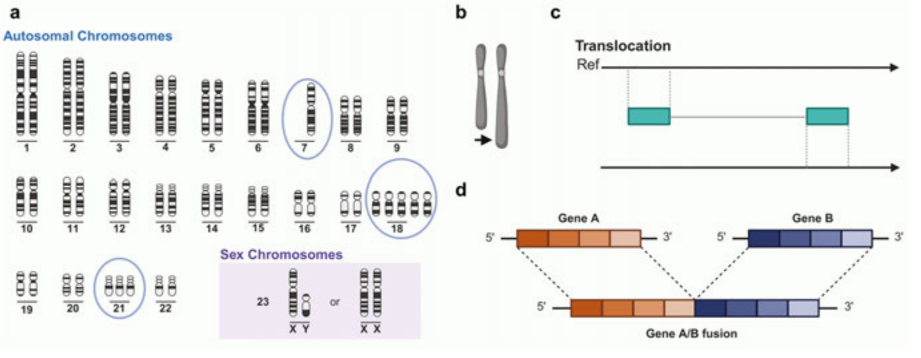


Figure 2. Examples of chromosomal abnormalities found in ALL cells. These include numerical aberrations as a) abnormal number of chromosomes (extra or missing copies), b) intrachromosomal amplifications (e.g., iAMP21), or structural aberrations c) such as translocations, and d) expression of fusion genes (created with Biorender.com).

The introduction of next-generation sequencing (NGS) techniques has led to the discovery of recurrent genetic aberrations, allowing the classification of patients into new subtype groups[7,8,21,25–27]. This has resulted in a more comprehensive understanding of the oncogenic subtypes in BCP-ALL and T-ALL[7,20], resolving a large proportion of previously unclassified cases (**Figure 3**).

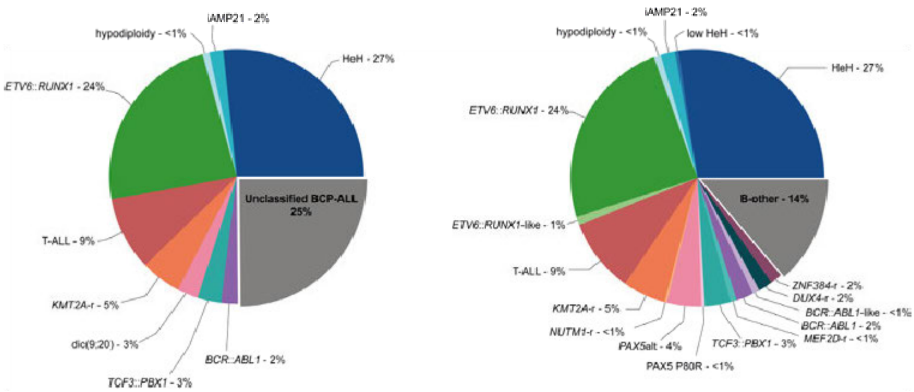


Figure 3. Distribution of ALL subtypes in 1,131 pediatric patients diagnosed in the Nordic countries between 1996 and 2013, as determined by traditional methods at diagnosis (left) and after molecular screening (right, **Paper I**). On the right panel, dic(9;20) has been re-labelled as *PAX*alt .

Subtype-defining alterations in BCP-ALL

According to the latest International Consensus Classification (ICC), there are 22 molecular genetic subtypes and five provisional entities of BCP-ALL[20]. The canonical subtypes are defined by large-scale recurrent chromosomal alterations, such as aberrant chromosomal numbers, e.g., high hyperdiploidy (HeH, > 50 chromosomes), or hypodiploidy subtypes (< 40 chromosomes). Furthermore, translocations with expressed fusion genes appear in t(12;21)(p13.2;q22.1)/*ETV6*::*RUNX1* (*ETV6*::*RUNX1*), t(9;22)(q34.1;q11.2) /*BCR*::*ABL1* (*BCR*::*ABL1*), t(1;19)(q23.3;p13.3)/*TCF3*::*PBX1* (*TCF3*::*PBX1*), and t(v;11q23.3)/*KMT2A*-rearranged (*KMT2A*-r) cases. Additionally, intrachromosomal amplifications may occur, such as the amplification of chromosome 21 (iAMP21).

During the last decade, NGS methods such as RNA sequencing (RNA-seq) enabled the identification of non-canonical subtypes with expressed fusion genes. These involve the genes *DUX4*[28,29], *MEF2D*[30,31], *ZNF384*[32–34], and *NUTM1*[35,36] genes, or subtypes similar to their canonical counterparts, i.e., *BCR*::*ABL1*-like/"Ph-like"[37,38] *ETV6*::*RUNX1*-like/"ER-like"[28,39]. Finally, *PAX5*-driven subtypes, i.e., *PAX5* alteration (*PAX5*alt) and point mutations in PAX5 P80R were also identified[40–42].

Subtype-defining alterations in T-ALL

T-ALL includes a well-defined subtype known as early T cell precursor ALL (ETP ALL), with approximately a third of the cases harboring *BCL11B* rearrangements. A phenotypically similar group, referred to as near-ETP ALL, shares clinical features with ETP ALL but is genetically distinct. The remaining T-ALL cases are characterized by the aberrant activation of various transcription factors[20]. These subtypes are not commonly utilized for risk stratification in current clinical treatment protocols, largely due to technical challenges in detecting the subtype-defining aberrations at diagnosis, and are therefore referred by ICC as provisional entities. However, Pölönen et al.[7,8] employed whole genome sequencing (WGS), whole exome sequencing (WES), and RNA-seq and identified 15 distinct molecular genetic T-ALL subtypes. Unlike the current ICC stratification for T-ALL, the Pölönen-based stratification system suggests a more biologically relevant and prognostic classification, emphasizing data integration to reveal both coding and noncoding alterations.

Collectively, these findings underscore the importance of integrating diverse molecular datasets to unravel the genetic complexity of pediatric ALL.

Molecular and functional profiling

DNA methylation

A cytosine-phosphate-guanine (CpG) site is a DNA sequence region, where a cytosine nucleotide is followed by a guanine in the 5' to 3' direction, separated by a phosphate group. There are approximately 28 million CpGs in the human genome[43], some of which cluster into CpG-rich regions known as CpG islands (CGI). There are ~30,000 CpG islands in the human genome[44], spanning at least 200 base pairs and containing more than 50% C+G content[45]. CGIs can be found in promoter regions, within gene bodies (intragenic) or between genes (intergenic)[46]. Methylation of CpG dinucleotides (DNAm) is an epigenetic modification that does not alter the DNA sequence, but can influence gene expression (GEX) by affecting the DNA structure and accessibility. In mammals, the most common type of DNAm involves the addition of a methyl group on the 5th carbon of the cytosine in CpG dinucleotides by DNA methyltransferases (DNMTs), resulting in the formation of 5-methylcytosine (**Figure 4a**). Two *de novo* DNMTs, DNMT3A and DNMT3B, create new methylation patterns, while DNMT1, known as maintenance DNMT, is responsible for copying the DNAm patterns to the replicated DNA[47] (**Figure 4b**). DNAm plays an essential role during cellular development and differentiation. It contributes to gene silencing, genomic imprinting, X-chromosome inactivation, tissue- and cell-type specific regulation, and maintenance of genomic stability during mitosis[46–49]. In most genomic contexts, CpGs are methylated, including CGIs in intragenic and intergenic regions, as well as CGI promoters of repressed genes, such as imprinted genes and genes on the inactive X chromosome[46,47,50,51]. However, CGIs in and around transcription start sites often remain unmethylated and are associated with the expression of essential genes, including housekeeping genes and those important for development and differentiation, as well as cell-type or tissue-specific genes[46–49]. To date, the progression between DNAm and gene silencing remains a subject of debate[46]. Some studies suggest that DNAm occurs after gene silencing, maintaining the silent state[52–54]. In contrast, Challen et al. provided evidence that DNAm is necessary for initiating gene silencing rather than preserving it. Their study showed that the loss of DNMT3A compromised hematopoietic stem cell differentiation (HSC) and led to the accumulation of HSCs in the BM of mice[55].

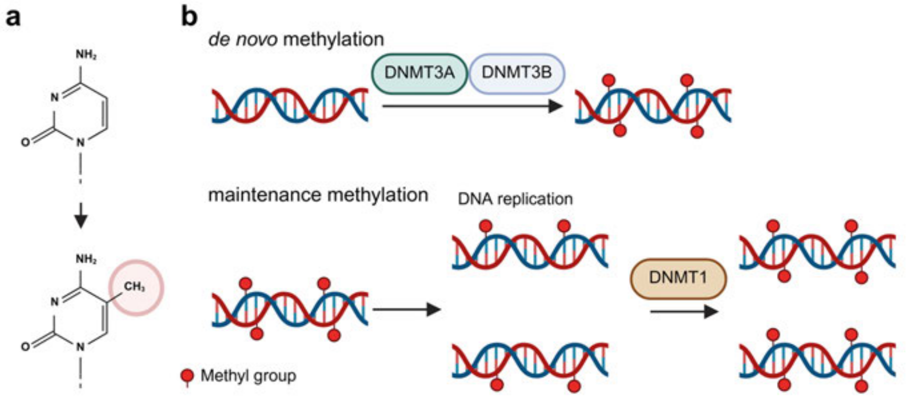


Figure 4. a) The addition of a methyl group at the 5[th] carbon of cytosine is catalyzed by DNA methyltransferases (DNMTs). b) DNMT3A and DNMT3B mediate *de novo* methylation (top), while DNMT1 maintains methylation during DNA replication (bottom, created with Biorender.com).

Methylation patterns are altered in various diseases, including cancer, where hypermethylation of CGIs that facilitates gene silencing often co-occurs with a global loss of methylation[56–58]. In contrast to other cancer types, ALL methylomes have been shown to retain high levels of DNAm globally[59], but differences are observed among subtypes[60,61]. Strong associations between DNAm and recurrent chromosomal aberrations (subtype), as well as outcome, have been observed in pediatric ALL[59,60,62,63]. Additionally, a distinct hypermethylation pattern that occurs on multiple CGIs in cancer is referred to as the CpG island methylator phenotype (CIMPs). CIMPs have been associated with variable outcomes across different cancer types[64], including pediatric T-ALL[11,65,66] and BCP-ALL[67]. In T-ALL, it has been proposed that the combination of CIMP and MRD status has potential utility in guiding risk stratification[11]. These findings highlight the value of incorporating DNAm as a biomarker, especially in diseases where genetic alterations do not fully explain heterogeneity[11]. In addition, DNAm has been used as a proxy for assessing the mitotic history and evolutionary dynamics of various hematological malignancies, underscoring its pivotal role in tracking disease progression from an epigenetic perspective[68,69].

The association between leukemic subtypes and DNAm modifications is well established[70–72]. DNA is stable and less prone to degradation compared to RNA, making DNAm profiling a promising analyte for investigating biobanked, degraded, or low-quality samples[73]. Such retrospective studies are important for analyzing rare subtypes and for investigating subtype-specific outcomes with the help of long follow-up patient data.

Several methods exist for DNAm profiling, including targeted methylation sequencing, whole genome bisulfite sequencing (WGBS), and microarrays. These approaches rely on bisulfite conversion, although bisulfite-free alternatives such as enzymatic methyl-seq and long read sequencing have been developed[74,75]. Among the bisulfite-based methods, the Infinium BeadChip assays (450k and EPIC arrays from Illumina, Inc.) have been widely used in epigenetic studies[76–79] and served as the platform for generating DNAm data analyzed in **Papers I-IV**. A key step to both array-based and sequencing-based methylation methods is bisulfite conversion, in which unmethylated cytosines in CpG sites are chemically converted to uracils, while the methylated cytosines remain unchanged. Despite arrays only cover ~5% of CpG sites, they continue to provide cost and computationally effective, high-throughput profiling of multiple samples in a single run, covering most promoter, enhancer regions, and CpG islands across the genome, and are supported by many computational packages for data normalization and analysis[80–85].

Array-based methylation platforms, such as the Illumina 450k and EPIC arrays, enable the interrogation of the methylation status of ~450K and ~800K CpG sites, respectively. Each CpG site is assayed by a 50-base probe designed to hybridize to bisulfite-converted DNA. The arrays utilize two types of probes, Type I and Type II, which differ in detection chemistry and genomic coverage. Type I uses two separate bead types, one for methylated and one for the unmethylated state, and rely on a single fluorescent channel for detection. In contrast, Type II probes use a single bead type and determine the methylation status through single base extension, requiring two channels for the quantification of the methylated (green channel) and unmethylated (red channel) signals. This design difference can introduce technical variability, particularly because Type II probes have compressed dynamic range and constitute the majority of the probes on the array. Therefore, appropriate normalization is crucial to ensure comparability across probes and reduce bias.

The intensities of the methylated and unmethylated signals are quantified and further used to calculate either beta or M values, which are further utilized in downstream analyses. Beta values represent the ratio of methylated signal to the total signal, ranging from 0 (fully unmethylated) to 1 (fully methylated). A constant value a (commonly set to 100) is usually added to the denominator to prevent division errors by small intensity values (1). M values represent the $\log_2$ ratio of the methylated to the unmethylated intensities (2). An M value of 0 indicates that a CpG site is equally methylated and unmethylated, while positive and negative values indicate hypermethylation and hypomethylation, respectively. M values are often used for statistical analysis, as they offer improved interpretability and more suitable statistical properties for modelling differences between groups compared to beta values.

$$beta\ value = \frac{Methylated}{Methylated + Unmethylated + a} \quad (1)$$

$$M\ value = log2\left(\frac{Methylated}{Unmethylated}\right) \quad (2)$$

Finally, the signal intensities can additionally be used to infer copy number alterations (CNAs)[86,87]. This gives the arrays an additional modality read-out by providing data on chromosomal aberrations. These include large-scale aneuploidies as well as focused gene-level deletions or amplifications, given the gene of interest is covered by enough probes, and potential fusion gene breakpoints.

Gene expression

The human genome contains approximately 20,000 protein-coding and 40,000 non-coding genes[88]. GEX refers to the transcription of DNA (genes) into RNA products, which include messenger RNAs (mRNA) that are translated into proteins, and functional non-coding RNAs (ncRNA). Together, these RNA molecules comprise the transcriptome. GEX patterns can be cell type-specific or ubiquitous, and their disruption may signal disease onset.

Among protein-coding genes, tumor suppressors and oncogenes play a critical role in maintaining normal cell growth and division. Genetic alterations affecting those genes, such as loss-of-function mutations in tumor suppressors or gain-of-function mutations in oncogenes, can drive tumorigenesis[89]. ncRNAs include long non-coding, ribosomal, transfer, micro, and small RNAs such as small nucleolar, small nuclear, small interfering, and piwi-interacting RNAs. These molecules constitute important regulators of GEX and are often implicated in cancer[90,91].

Approximately 60% of genes have CGI promoters, which are typically unmethylated, and their expression levels are regulated with the help of transcription factors[46]. Some promoters, however, are methylated, contributing to transcriptional silencing[56,58]. While these findings suggest the significant role of DNAm on gene regulation, the DNAm levels of the majority of CpG sites do not appear to directly influence GEX levels[46,92].

Genome-wide GEX profiling can aid in assessing the functional impact of genetic, epigenetic, and other regulatory factors on GEX. This can be achieved using, for example, microarrays (e.g., Affymetrix) or RNA-seq. In both methods, RNA molecules are reverse-transcribed into cDNA, typically using oligo(dT) primers to capture poly-A RNA or random primers to provide broader transcript coverage. In Affymetrix arrays, GEX is quantified by measuring the fluorescent signal produced after the hybridization of labeled cDNA fragments to the oligonucleotide probes on the microarray[93]. In RNA-seq, GEX is profiled by high-throughput sequencing of cDNA fragments after poly-A selection, ribosomal RNA depletion, or size selection[94]. Compared to microarrays, RNA-seq offers several advantages, including the ability to detect novel transcripts, quantify rare or lowly expressed genes, identify fusion genes, and expressed mutations in coding regions[94–96]. In a growing number of centers, RNA-seq is being used to aid diagnosis of patients with ALL and was applied to generate the GEX data for **Papers I, III, and V**.

Subtype-specific GEX dysregulation is widely observed in ALL[7,97]. Apart from expression-level changes, other aberrations are associated with ALL subtypes, including fusion genes, are detectable by RNA-seq. For example, when chromosomal rearrangements occur, fusion genes can emerge[98]. These fusion events are typically mutually exclusive across BCP-ALL subtypes, suggesting their role in disease etiology[97,99]. The data produced by RNA-seq are particularly valuable for patient subtype identification, including cases for which standard cytogenetic analyses fail[7,25,100–105]. Importantly, some fusions also encode oncogenic proteins that are clinically relevant, underscoring the importance of their detection. For instance, in *BCR*::*ABL1* BCP-ALL leukemia, the *BCR*::*ABL1* fusion gene encodes an active tyrosine kinase (TK), which contributes to uncontrollable cell growth and proliferation. Such cases can be treated using targeted agents, such as tyrosine kinase inhibitors (TKI)[106]. In contrast, in *BCR*::*ABL1*-like BCP-ALL leukemia, which shares gene expression profiles with *BCR*::*ABL1* cases but lacks the *BCR*::*ABL1* fusion gene, is characterized by diverse kinase-activating fusions. The wide variety of underlying genetic lesions poses a challenge in evaluating individual therapy responses. Patients who harbor expressed ABL-class fusions, or JAK-STAT alterations can benefit from TKIs[107]. As *BCR*::*ABL1*-like is classified as high-risk, it is critical to identify this subtype at diagnosis, although not all of the patients carry such actionable alterations. Systematic identification of these lesions may support the development of more effective treatment strategies[107].

Ex vivo drug screening

Ex vivo drug response profiling is a technique that evaluates how patient-derived cells respond to pharmacological agents outside the body (*ex vivo*). Numerous studies have shown that *ex vivo* responses are associated with *in vivo* responses and clinical outcomes, making *ex vivo* screening a promising tool for assessing the risk of treatment resistance and failure[108–113]. In pediatric ALL, a recent pharmacotyping study demonstrated that molecular subtypes often exhibit distinct *ex vivo* responses to different drug classes, while treatment response, measured by MRD levels, was significantly associated with *ex vivo* drug sensitivity for certain drugs. Altogether, these findings highlight the additive value of integrating *ex vivo* profiling, molecular characterization, and *in vivo* response data to guide precision therapeutic strategies[108].

Ex vivo drug response profiling is commonly performed using cell viability assays[114], including the 4-day in vitro 3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyl tetrazolium bromide (MTT) drug resistance assay[108,115,116], or the mesenchymal stem cell (MSC) co-culture system assay with flowcytometry[108,117]. These methods enable the determination of the lethal concentration 50 (LC_{50}), the drug dose that kills 50% of the leukemic cells. Other cell viability-based methods include the fluorometric microculture cytotoxicity assay (FMCA)[118,119] and ATP-based assays[109], which measure esterase activity and cellular metabolic activity, respectively, to quantify the percentage of surviving cells after drug exposure. Similar to LC_{50}, the inhibitory concentration 25 (IC_{25}) reflects the concentration that reduces the metabolic activity of a cell by 25% and has been used to assess drug response in leukemia cells[110]. Among these methods, FMCA offers several advantages, including short incubation period (2-4 days), high sensitivity due to fluorescence-based detection, and a simplified staining procedure[119]. FMCA, as a cell viability assay, measures the proportion of surviving cells after 72h of drug exposure. Cells with intact plasma membranes produce fluorescent signal due through the conversion of fluorescein diacetate to fluorescein by esterases. The survival index (SI) is calculated as the ratio of the mean fluorescent signal from drug-treated leukemic cells to that of untreated cells (controls), with the background fluorescence from the blank wells (culture medium) subtracted [119]. These data were used in **Papers III-IV** as a proxy to assess *ex vivo* cellular drug responses.

Computational applications in biomedical research

Machine learning and survival modelling

Machine learning (ML), a subset of artificial intelligence (AI), involves training models capable of performing intelligent tasks (**Figure 5a**). Within ML, Deep Learning (DL) uses multi-layered neural networks to learn hierarchical patterns from data. These deep architectures typically demand substantial computational power but excel in handling large volumes of unstructured data, making them particularly valuable for complex pattern recognition tasks. ML approaches are versatile, with applications ranging from email spam detection[120], stock market prediction[121], and weather forecasting[122], to the development of intelligent agents powered by language models[123,124]. In biomedical research, ML is particularly well poised for analyzing high-dimensional NGS data, array-based measurements, and clinical records. The application of ML has demonstrated promising results for disease outcome prediction[125,126], cell type annotation[127,128], and biomarker discovery[129]. Specifically in leukemia, ML-based methods have shown promising results for subtype classification and clinical decision support[70,71,100–103,105]. Building on this knowledge, such computational approaches hold the potential to become fast, accurate, and cost-effective diagnostic and prognostic tools, supporting both fundamental research and clinical translation.

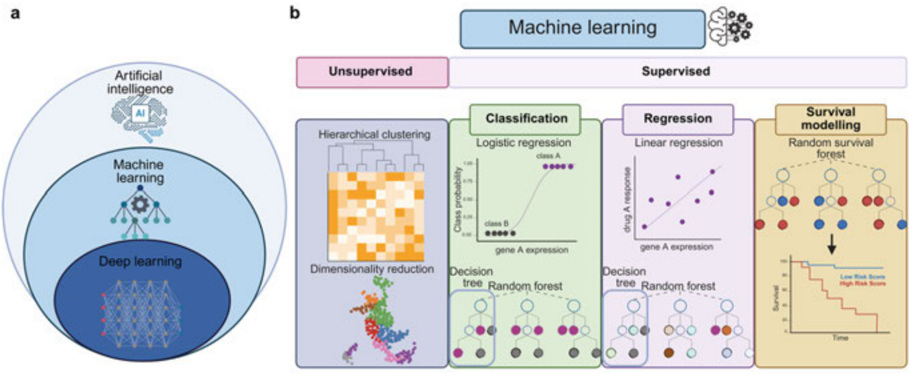


Figure 5. a) The relationship between artificial intelligence (AI), machine learning (ML), and Deep learning (DL). b) Overview of the two ML types (unsupervised vs. supervised learning) that are explored in this thesis with examples. The supervised methods are extended to classification, regression, and survival models (Created with Biorender.com).

ML techniques are categorized into four types–unsupervised, supervised, semi-supervised, and reinforcement learning[130]. Unsupervised learning does not rely on labeled data to identify underlying patterns and structures, whereas supervised learning leverages labeled data to build models. Semi-supervised learning combines unlabeled and labeled data to improve performance when the annotated data are scarce. Reinforcement learning employs an internal reward-and-penalty system to improve its learning capacity from feedback and is commonly applied in tasks such as autonomous vehicle navigation[131]. This thesis focuses on unsupervised and supervised learning, given their relevance to the scientific questions explored herein (**Figure 5b**).

Unsupervised learning includes methods such as clustering algorithms and dimensionality reduction techniques, which are fundamental for exploratory analysis and data preprocessing. Hierarchical clustering is often employed for generating heatmap visualizations or identifying sample groupings. Dimensionality reduction methods, such as principal component analysis (PCA)[132] and Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP)[133,134], enable the projection of complex, high-dimensional data into a two dimensional space, revealing potential hidden structures within the data. When the resulting clusters are well-defined, these methods can aid the identification of previously unrecognized subgroups[7,8].

Supervised learning uses labeled data to develop predictive models, with performance evaluation based on the capacity of the model to predict the correct labels on unseen data. Classification tasks involve discrete categorical outcomes (e.g., ALL subtypes), while regression tasks predict continuous numerical variables (e.g., drug response).

Depending on the research question, linear models can be employed to capture simple, direct relationships between the input features (independent variables) and the target (dependent variables), while non-linear models are suited to detect complex interactions, which are commonly encountered in real-world datasets. The *Nearest Shrunken Centroid* (NSC) classifier exemplifies a supervised learning approach that combines a linear model with integrated feature selection capabilities[135]. Initially developed for GEX microarray data, NSC has demonstrated robust performance when applied to RNA-seq[136] and DNAm for ALL subtype prediction[71], and was successfully used in **Paper I** of this thesis for both DNAm (450k array) and GEX (RNA-seq) data. NSC works by applying a shrinkage threshold to the standardized differences between the class-specific feature means (centroids) and the overall centroid across all classes, shrinking small differences toward zero. As a result, the shrunken centroids are pulled closer to the overall centroid. Features with low centroid variability across all classes are considered non-informative and excluded, retaining only those with substantial differences across classes for model training (**Figure 6**). This shrinkage-based approach allows the model to focus on putative clinically relevant signals, while reducing the noise. Alternative feature-selecting classifiers include decision trees, random forests, k-nearest neighbors, and support vector machines–each of which offering linear or non-linear configurations[137].

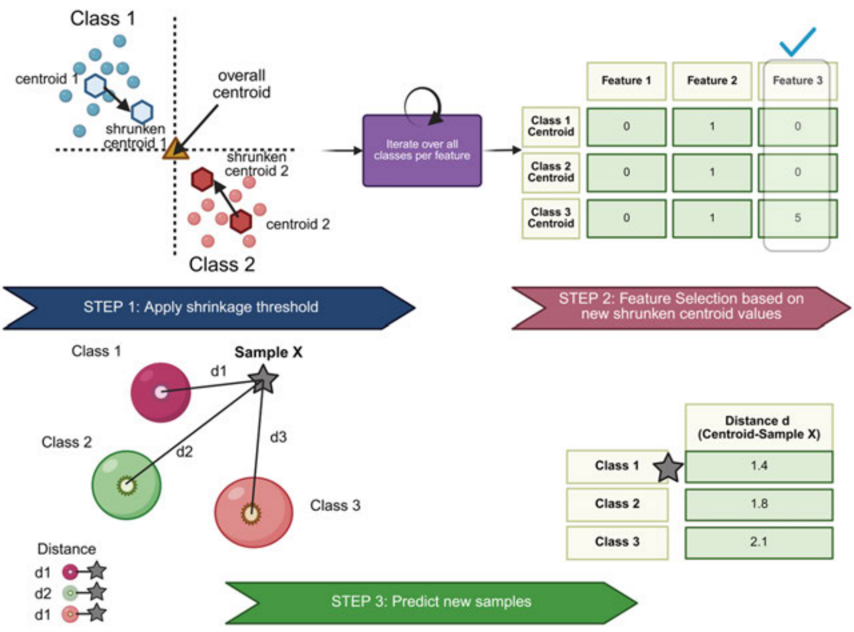


Figure 6. The Nearest Shrunken Centroid (NSC) classifier workflow, a core component of **Paper I**, operates in three steps. **STEP 1**: For each feature, a shrinkage threshold (soft thresholding hyperparameter) is applied, pushing the mean (centroid) of each class toward the overall centroid of the whole dataset. **STEP 2**: Features with low variability for all class centroids are excluded, keeping those with clear differences for at least one class. **STEP3**: The selected features are used to train a nearest centroid model that assigns new samples to the class with the nearest centroid in Euclidean distance. This approach removes noisy, non-informative features from a dataset and highlights those of potential clinical importance[135] (Created with Biorender.com).

Algorithms that support regression tasks include linear regression, decision trees, random forests, k-nearest neighbors, and support vector machines[137]. A notable exception is logistic regression, which, despite its name, functions as a classification model for binary or multinomial cases (multiple classes). Beyond generating discrete class predictions, logistic regression provides probability estimates for each class–a feature shared by many classification algorithms including NSCs.

Cox proportional hazards regression is a statistical method widely used in survival analysis, designed to evaluate the hazard ratio or relative risk of an event occurrence (e.g., relapse or death) over time while accounting for one or more covariates (univariate and multivariate, respectively). Cox regression does not directly yield probability estimates, but instead identifies covariates, features correlated with increased or decreased risk of events. Random survival forests and gradient boosted models[138] extend traditional survival analysis by incorporating ensemble trees specifically adapted for time-to-event datasets. These models can predict event risk based on molecular or/and clinical input data. Cox Regression and survival models were used in **Papers II-IV**.

Ultimately, by using class probabilities alongside model performance metrics, such as accuracy, sensitivity (recall), specificity, precision, and F1-score for classification tasks, and the concordance index (c-index) for survival analysis, the models can support robust and interpretable predictions. This is a critical step toward developing tools that may adapted for clinical decision-making.

Differential analyses

The increasing availability of high-throughput data from microarrays (e.g., 450k array) and NGS methods such as RNA-seq has led to the development of various computational tools for identifying disease biomarkers and potential therapeutic targets[139–141]. While ML-based methods demonstrate considerable power for performing such tasks, they often rely on large training datasets to perform reliably in high-dimensional feature spaces and necessitate external validation to ensure generalizability. In contrast, differential analysis methods offer a robust alternative, particularly for smaller yet complex datasets.

Among these, DESeq2[140] has become a widely used method for differential expression analysis (DEA) and was employed in **Paper III** (**Figure 7**) to identify genes differentially expressed between *ex vivo* drug-sensitive and drug-resistant patients. DESeq2 provides several advantages, including reproducibility, suitability for small sample sizes, and outlier detection that reduces false positives (Type I error). These characteristics allow DESeq2 to accommodate both straightforward comparative studies and complex multi-factorial study designs. Another software package *limma*[141], originally developed for microarray-based DEA, has been extended to accommodate RNA-seq data. Beyond gene expression, *limma* is also compatible with DNAm arrays, and was employed in **Paper III** for differential methylation analysis (DMA, **Figure 7**).

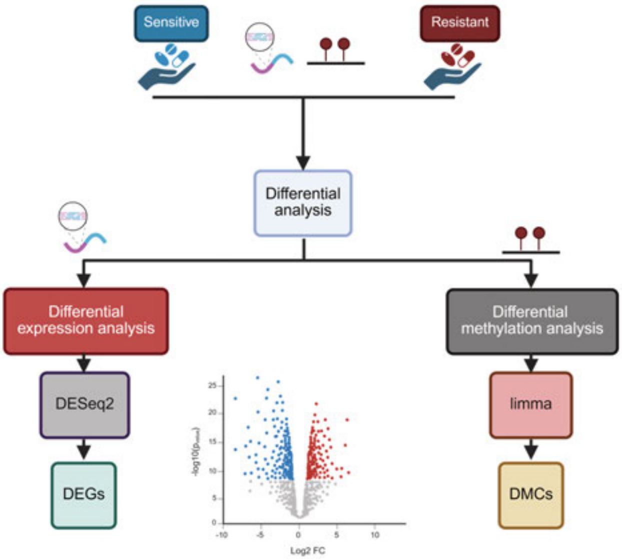


Figure 7. Differential analysis workflow followed in **Paper III**. DESeq2[140] and *limma*[141] R packages were utilized to identify differentially expressed genes (DEGs) and differentially methylated CpG sites (DMCs) that differentiate *ex vivo* drug-sensitive from drug-resistant patients (Created with Biorender.com).

While conventional statistical tests such as the non-parametric Mann-Whitney U test can handle simple two-group comparisons, they lack the flexibility required to model complex experimental designs or incorporate covariate adjustment, as DESeq2 and *limma* do. Both DESeq2 and *limma* are implemented as R/Bioconductor packages and are supported by extensive documentation and regular updates, making them accessible and reliable tools for the research community. Notably, both methods employ shrinkage techniques to stabilize variance estimates, with *limma* using empirical Bayes moderation to shrink gene-wise variance toward the common variance, and DESeq2 applying empirical Bayesian shrinkage to shrink gene-wise dispersion (within-group variability) estimates toward the predicted values. These approaches enhance statistical power and reduce false discoveries, particularly in studies with limited sample sizes, where outliers might otherwise skew results.

The final output of DESeq2 and *limma* are $\log_2$ fold change ($\log_2$FC) values, which quantify the logarithmic (base 2) difference in expression or methylation levels between two groups. This logarithmic transformation provides an intuitive interpretation: a $\log_2$FC of 1 indicates a two-fold increase in the treatment group relative to the controls, while a $\log_2$FC of -1 represents a two-fold decrease, meaning that the treatment group exhibits half the expression or methylation levels compared to the controls. The effect size values are computed for each feature (gene or CpG site) included the analysis.

Complementing the fold estimates, both methods provide statistical parameters, such as raw p-values and adjusted p-values, the latter of which account for multiple testing correction.

Data Integration

While diseases such as cancer are multifactorial in nature[142], studies that integrate multiple data modalities remain relatively scarce. This gap persists due to several challenges, including the lack of standardized integration frameworks, heterogeneous data pre-processing requirements, partially overlapped datasets, and computational complexity. Nonetheless, recent efforts have begun to uncover the molecular heterogeneity of hematological malignancies through a multiomic lens[143–147].

Integrating multiomic data offers a more comprehensive approach to increase understanding of the molecular dependencies and their associations with clinical variables, such as therapeutic responses. Integration strategies can vary from simple pairwise comparisons, as performed in **paper III**, to more complex frameworks capable of simultaneously analyzing multiple modalities, such as the one used in **paper IV**. The latter yields a latent space, a low-dimension representation of feature combinations across multiple data modalities.

Overall, data integration can support downstream analyses including pathway enrichment, multimodal network construction, ML-based predictions, and the discovery of clinically relevant molecular subgroups. The analyses can be performed by either using the latent integrative space or extracting top features (**Figure 8**).

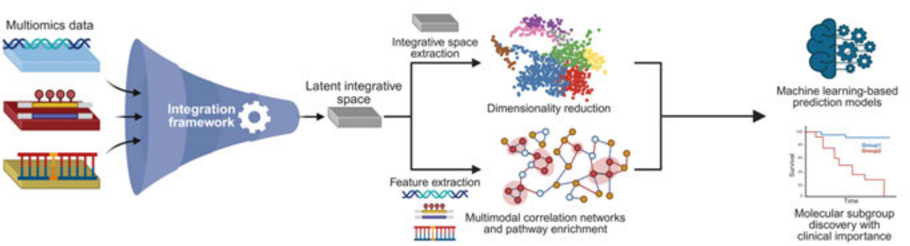


Figure 8. Multiomics integration as a framework for downstream analyses. Following integration, either the latent integrative space (top), or selected top features (bottom) can be leveraged to various downstream applications, including dimensionality reduction, multimodal network analysis and pathway enrichment, ML-based prediction models, and molecular subgroup discovery with clinical significance (bottom, Created with Biorender.com).

Among integrative frameworks, Multi-Omics Factor Analysis (MOFA)[144,145], is an unsupervised method applicable to both bulk and single-cell data. MOFA functions similarly to single-modality dimensionality reduction methods like PCA, but is designed to detect latent factors across multiple data types. Each MOFA-derived factor captures variation shared or unique to specific modalities, and ranks features with associated impact values (weights). These weights can be used to extract the most informative features per factor. In **Paper IV**, MOFA was applied to integrate molecular profiles, drug response measurements, and clinical parameters from BCP-ALL patients, enabling the functional interpretation and annotation of each factor by analyzing its top-weighted features.

Other integrative tools support both unsupervised and supervised components, enabling direct application of their outputs to prediction tasks such as disease classification [148–150]. These tools apply a variety of computational strategies, including probabilistic, similarity-based, correlation-based, multivariate, network/graph-based, and data fusion approaches[150]. They also differ in how they handle missing data within incomplete multiomics datasets. For example, MOFA excludes missing values during integration but can robustly impute them afterward by leveraging learned factor values and feature weights[144,145]. In contrast, Integrate Any Omics[148] builds modality-specific similarity networks and facilitates integration through matched samples that serve as bridges for information flow.

Integration models further differ in their ability to capture linear versus non-linear relationships among features. MOFA is a linear method unable to model non-linear interactions, instead offering interpretability and computational efficiency. This makes it well-suited for exploratory analysis of high-dimensional omics data, especially when compared to more complex and less interpretable approaches.

In summary, translational research requires time-efficient and cost-effective computational methods capable of generating clinically meaningful and interpretable insights, whether in the form of model predictions paired with feature importances or guidance for functional validation.

Aim of thesis

The central aim of this doctoral thesis is to map the multi-layered molecular landscape of childhood ALL using multiomics data from patients diagnosed and treated in the Nordic countries. By identifying molecular signatures across diverse biological layers and uncovering shared biological signals between them, this work seeks to improve our understanding of how these molecular patterns influence disease etiology, treatment response, and clinical outcomes. Collectively, the findings aim to support translational research and future precision diagnostics.

Summary of papers

Paper I

Title and reference: Multimodal classification of molecular subtypes in pediatric acute lymphoblastic leukemia (Krali,O., *et al.*, 2023 *npj Precision Oncology*)

Study Overview

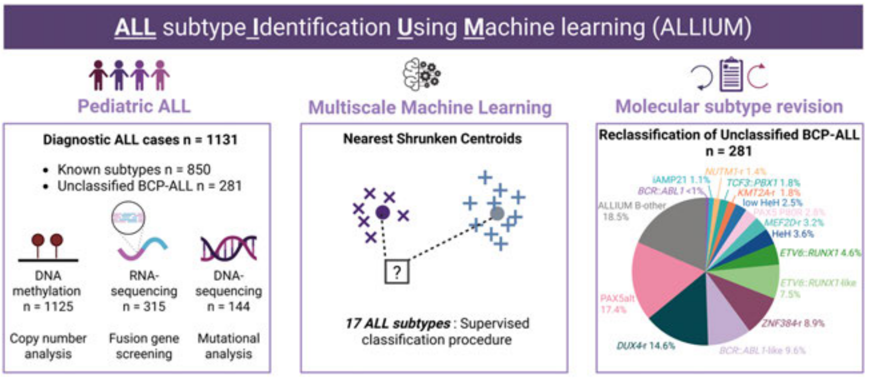


Figure 9. DNA methylation (DNAm, 450k arrays), gene expression (GEX, RNA-sequencing), and somatic mutation (WGS, targeted sequencing) data were generated from 1131 patients treated on the Nordic Society for Pediatric Hematology and Oncology (NOPHO) protocols diagnosed between 1996 and 2013. In total, the subtype of 281 of the BCP-ALL patients (24.8% of the entire ALL cohort) was unclassified at diagnosis. Molecular screening was performed based on a combination of cytogenetics, fusion gene screening, mutational analysis, and copy number analysis. Molecular screening resolved the subtype of 127 BCP-ALL patients. The remaining 154 patients were denoted "B-other". A supervised classification method (ALLIUM) was used to build subtype-specific models based on two modalities (DNAm and GEX) for 17 of the known molecular subtypes of ALL. ALLIUM re-classified the subtype of 102 B-other patients. This study expanded the scope of known subtypes across the entire cohort resulting in 1079 with known subtype (95.4% of the entire ALL cohort). The 52 patients remaining unclassified at the end of the study are referred to as ALLIUM B-other. From Krali, O. et al., 2023, *npj Precision Oncology*, reprinted with

permission from Springer Nature, under the terms of the Creative Commons Attribution 4.0 International License (CC BY).

Background and Aim

In ALL, molecular subtypes represent a key clinical variable, often guiding risk stratification. In a cohort of 1,331 patients diagnosed with ALL in the Nordic countries between 1996 and 2013, 850 patients were assigned to known subtypes at ALL diagnosis. The remaining 25% were cytogenetically unclassified at diagnosis due to lacking or inconclusive molecular findings from traditional cytogenetic-based diagnosis. The rapid adoption of NGS in the last decade enabled the discovery of new, non-canonical subtypes defined by distinct GEX patterns, expressed fusion genes, or specific mutations. Given the strong association between recurrent aberrations, DNAm, and GEX in ALL, we aimed to develop a supervised molecular classifier, *ALL subtype Identification Using Machine learning* (ALLIUM), to predict subtypes using either DNAm or GEX data.

Methods

We analyzed DNAm (450k arrays) and GEX (RNA-seq) data from 1,131 patients diagnosed and treated in the Nordic countries between 1996 and 2013. To resolve unclassified BCP-ALL cases (n = 281), we first integrated fusion gene screening, CNAs, and mutational data generated for this study and from our previous work in this cohort[28,151–155] mapping them to the 2022 ICC classification for ALL[20]. The annotated sample set was used to train and optimize two sets of subtype-specific (16 BCP-ALL and T-ALL) NSC models, with one set using DNAm and the other using GEX data–from 819 and 207 patients with known subtypes, respectively. The classifiers followed a hierarchical design, assigning group predictions before subtype predictions to reduce errors due to biological similarities. A control classifier was also developed to ensure that patients with low blast count would be separated from patients with no classification. The performance of ALLIUM was validated on internal hold-out and replication datasets, as well as three external datasets.

Results

ALLIUM selected 379 CpGs and 356 genes as the most informative signatures for defining the subtypes. Known ALL genes were among the subtype-specific signatures, such as *NUTM1* for *NUTM1*-r, *PBX1* for *TCF3*::*PBX1*, *MEF2C* for *MEF2D*-r, *CEBPA* for *ZNF384*-r, and CpG sites in *CBFA2T3* for the *ETV6*-group, the expression of *CDKN2A* and CpG sites in *AUTS2* for *PAX5*alt, and CpG sites in *ETV6*, *RUNX2*, and *IKZF1* for PAX5 P80R.

On both internal and external validation datasets, i.e., samples excluded from the classifier design, ALLIUM correctly classified 83-100% of the cases using DNAm and/or GEX. For patients with both data modalities available (n

= 242), 92% (n = 223) were correctly classified by both DNAm and GEX, demonstrating high biological concordance across data types. ALLIUM achieved its best performance in the subtypes *ZNF384-r*, *DUX4*-r, *MEF2D*-r, and *TCF3*::*PBX1*, while hypodiploidy (DNAm) and *BCR*::*ABL1*-like (DNAm) were among the most challenging. ALLIUM outperformed two widely used RNA-seq classifiers, ALLSorts[102] and ALLCatchR[100], for predicting HeH (86% vs. 71% and 86%), iAMP21 (100% vs. 33% and 67%), and *PAX5*alt (100% vs. 71% and 71%) in the hold out dataset. By integrating molecular screening with ALLIUM, 81.5% (229/281) of the previously unclassified BCP-ALL patients (**Figure 3**) were assigned to a distinct molecular subtype, reducing the proportion of patients labeled as B-other to only 4.6% of the total cohort (52/1131).

Discussion and future perspectives

We designed ALLIUM, a hierarchical classifier trained and supporting both DNAm and GEX data modalities, for molecular subtype classification in ALL using a small, yet biologically relevant set of signatures. Owning to its two independent modules, ALLIUM is flexible and can operate with either data modality. This is a unique feature of ALLIUM, which is not included in any over the other ALL subtyping approaches to date[71,100–103,105].

To facilitate broader implementation in both research and clinical settings, we subsequently implemented conformal prediction into ALLIUM[156]. This approach provides statistical guarantees to address uncertainty in subtype classification. When applied to ALLIUM's GEX-based predictions, the false negative rate on the known subtypes was reduced, as was the proportion of unclassified cases among patients with previously unknown subtype.

When we trained ALLIUM, we used RNA-seq generated within a relatively small, but well-characterized cohort of patients from NOPHO. This was a strength because the subtype membership of all the training samples in our classifier was well-established, although this came at the expense of a lack or underrepresentation of some rare subtypes. To date, at least five additional RNA-seq-based subtype classifiers for ALL have been proposed[100–103,105]. Most classifiers achieve similar excellent performances, however differences in diagnostic methods of the various training sets highlights the importance of large international collaborations to increase the sample sizes, particularly for rare subtypes, and to establish unified classifier frameworks, trained on harmonized datasets.

Extending ALL classification frameworks to other data types, such as DNA methylation from long-read sequencing, is a particularly interesting area for future research. A recent example of this concept is exemplified by the DNAm-based classifier ALMA[157], which demonstrated feasibility of applying an array-based classifier directly to long-read WGS DNA methylation calls. This exemplifies growing possibilities to extend pan-acute leukemia classification and the development of integrative classifiers that combine DNAm and GEX data, which will become more important moving forward in the clinical setting.

Paper II

Title and reference: Refining risk prediction in pediatric acute lymphoblastic leukemia through DNA methylation profiling (Mosquera Orgueira, A., Krali, O., *et al*., 2024 *Clinical Epigenetics*)

Study Overview

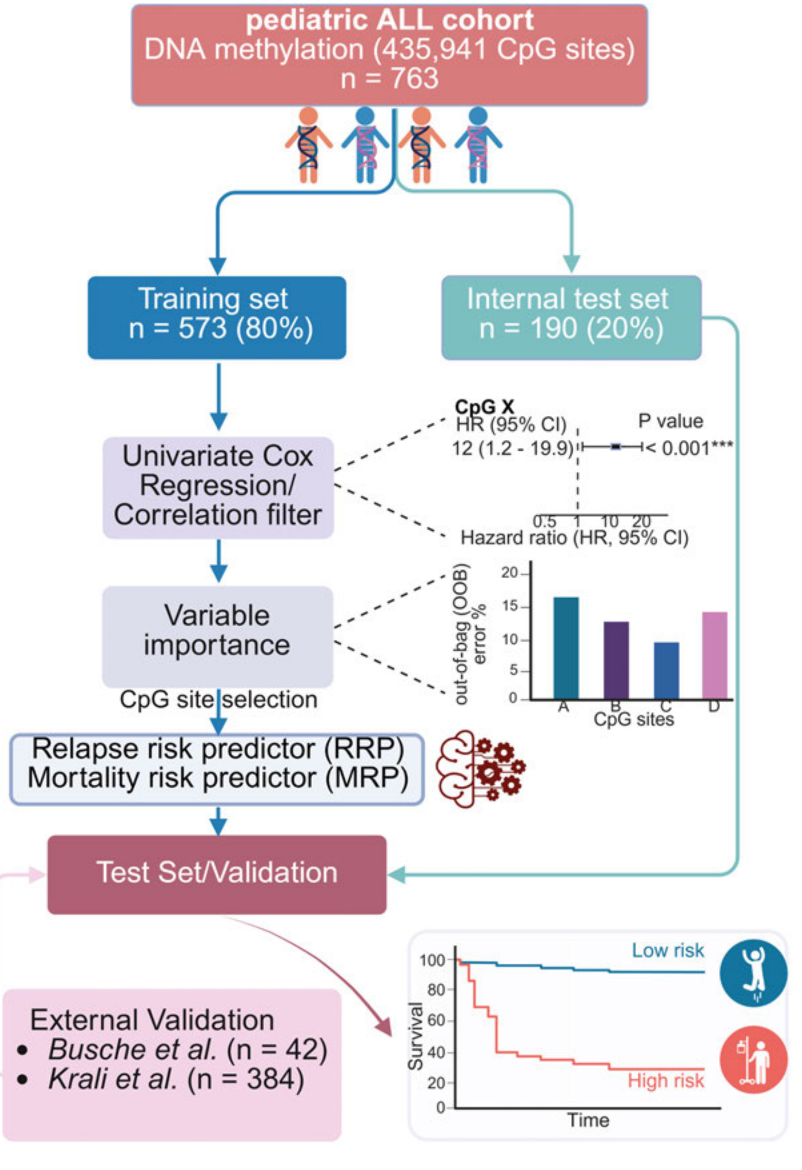


Figure 10. Graphical representation of the study design. The models were trained with data from 763 ALL patients, all of whom had previously been characterized by

genome-wide DNA methylation arrays. The dataset was partitioned into a training set (80% of the patients) and a test set (the remaining 20%). The training set was used to identify CpG sites with DNA methylation status associated with two key outcomes: relapse risk and mortality. The selected CpG sites were used to train Random Survival Forests models. Two models were generated: a Relapse Risk Predictor (RRP) and a Mortality Risk Predictor (MRP). The test set was utilized for internal validation. Finally, the models were further validated on two additional datasets. From Mosquera Orgueira, A., Krali, O., *et al*., 2024 *Clinical Epigenetics*, reprinted with permission from Springer Nature, under the terms of the Creative Commons Attribution 4.0 International License (CC BY).

Background and Aim

While DNAm profiling is not currently part clinical practice in ALL, its role as a disease biomarker has been extensively studied, largely own to its involvement in epigenetic regulation of normal hematopoiesis[158]. Specifically, in **Paper I**, we demonstrated that DNAm can be used for differentiating ALL subtypes at high resolution[159]. However, its application in risk stratification based on pretreatment signatures remains underexplored. Identifying molecular signatures associated with poor clinical outcomes, such as relapse and mortality, is essential to improve current stratification approaches. We developed two supervised ML models using DNAm data to predict the risk of relapse and mortality in pediatric ALL patients.

Methods

We trained two random survival forest models, a relapse risk predictor (RRP) and a mortality risk predictor (MRP), using DNAm (450k arrays) data and follow-up information from a training set of 573 ALL patients. Feature selection involved a combination of univariate Cox regression, correlation filtering, and model-derived variable importance scores. The final RRP and MRP incorporated selected CpG sites, time-to-event data including relapse free survival (RFS) and overall survival (OS), and event status (relapse or death). Model performance was evaluated on a test set (n = 190) and validated in two independent cohorts of pediatric ALL patients (n = 426).

Results

The models included 16 CpG sites for RRP and 53 CpG sites for MRP. Both models generalized well, achieving a c-index of 0.529-0.677 for RRP and 0.621-0.754 for MRP across internal and external validation cohorts. Interestingly, the MRP model also predicted the risk of relapse with comparable performance to the RRP (c-index: 0.62-0.643). Model performance improved when combined with clinical risk group data and surpassed traditional risk group-based models alone. Additionally, by stratifying patients into high- and low- groups based on their MRP scores, we observed that the resulting groups were often represented by high- and low-risk subtypes, respectively.

Discussion and future perspectives

We developed and validated two DNAm-based models, RRP and MRP, that outperformed traditional clinical risk group stratification and demonstrated added value when combined. Notably, some patients classified as high-risk by MRP belonged to canonical low-risk subtypes (e.g. HeH and *ETV6*::*RUNX1*), highlighting that subtype alone may not fully capture heterogeneity in treatment outcomes. These results suggest that DNAm profiling can refine risk prediction beyond cytogenetics.

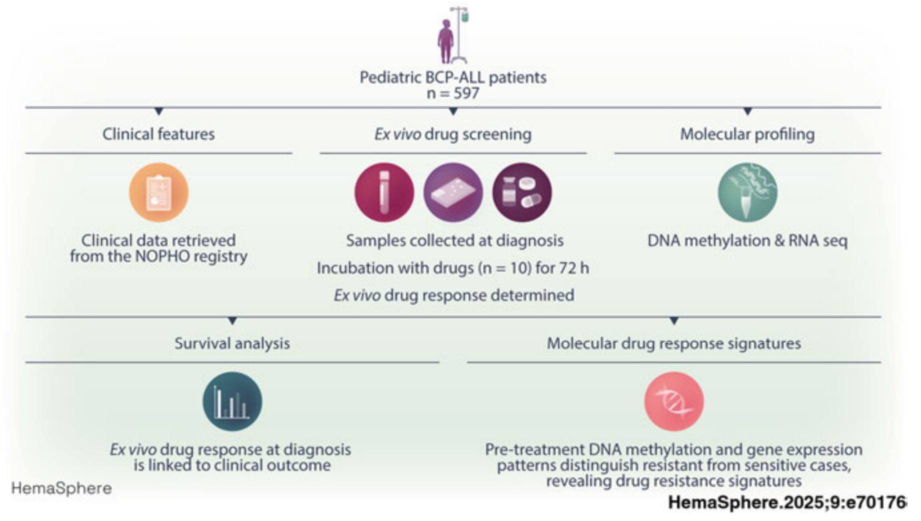
Similar efforts have shown that CIMP profiles can identify patients with inferior outcomes in T-ALL[11,65,66] and BCP-ALL[67]. In T-ALL, incorporating CIMP signatures with MRD status enhanced risk assessment[11,65,66], further supporting the value of combining molecular and clinical information for optimal risk stratification. In another example focusing on AML, ALMA not only enabled subtype classification but also performed risk prediction using a panel of 38 CpG sites[157]. In a previous study, we identified DNAm signatures that stratified patients within each of the three pediatric AML subtypes into two subgroups with distinct clinical outcomes[160]. Collectively, these studies underscore the potential of DNAm as a clinical biomarker.

Future implementations of RRP and MRP should include further exploration into the biological functions of the CpG sites, for instance through pathway analysis, to clarify mechanisms related to treatment responses and clinical outcomes. Integration with GEX data could reveal whether these DNAm signatures influence protein coding genes or reflect broader epigenetic programs. Future models tailored separately to BCP-ALL and T-ALL could help delineate lineage-specific risk predictors and further improve risk stratification. The introduction of additional DNAm signatures derived from epigenetic[161] and mitotic[68] clocks may enhance model performance.

Paper III

Title and reference: Ex Vivo Drug Responses and Molecular Profiles of 597 Pediatric Acute Lymphoblastic Leukemia Patients (Enblad, AP., *et al.*, 2025 *HemaSphere*)

Graphical Abstract



Background and Aim

While pediatric BCP-ALL patients generally have excellent outcomes, the molecular mechanisms underlying differing treatment responses and their prognostic implications remain inconclusive. Unravelling the biological determinants distinguishing treatment resistant and sensitive leukemic cells is critical for optimizing therapeutic strategies. This study investigated the association between *ex vivo* drug responses and clinical outcomes, while integrating drug response data with epigenomic and transcriptomic profiles in primary BPC-ALL patient samples.

Methods

We analyzed *ex vivo* drug response profiles (FMCA, n = 597) for ten treatment compounds, along with DNAm (450k and EPIC arrays, n = 437) and GEX data (RNA-seq, n = 119) from treatment-naïve samples. Patients were divided into sensitive, intermediate, and resistant response groups based the SI% values for each drug. Kaplan Meier survival analysis and multivariate Cox regression were used to assess the association between *ex vivo* response groups and clinical outcomes (RFS and OS). The Cox model incorporated relevant covariates, including sex and risk group to adjust for potential co-founding factors.

DMA and DEA were conducted comparing sensitive and resistant patients for each drug, adjusting for molecular subtypes. Overrepresentation analysis was employed to detect pathways in which differentially methylated CpG sites (DMCs) and differentially expressed genes (DEGs) were overrepresented. Transcription of the identified DEGs at cellular resolution was verified using single-cell RNA sequencing (scRNA-seq) data.

Results

Survival analyses demonstrated that for a majority of the drugs, *ex vivo* resistance was significantly associated with inferior RFS (log rank test $p < 0.05$), maintaining statistical significance after adjustment for clinical variables. Multidrug response groups were observed and were similarly linked to poor prognosis (RFS and OS, log rank test $p < 0.05$). Notably, the drug response groups did not align with the established clinical risk groups (standard, intermediate, and high risk).

Differential analyses identified 1,779 unique DMCs and 1,158 unique DEGs between *ex vivo* sensitive and resistant patients across the ten drugs. These signatures were enriched in biological pathways implicated in cancer and drug resistance, including heme metabolism, immune response, oncogenic signaling cascades (e.g., KRAS, TNF-alpha via NF-kB), cell cycle regulation, and transcriptional control mechanisms. T-cell receptor signaling and activation pathways were among the top-selected pathways, with single-cell validation confirming that the corresponding DEGs were expressed in leukemic blasts.

Discussion and future perspectives

This study highlights the clinical relevance of *ex vivo* drug profiling in pediatric BCP-ALL and its potential to refine risk stratification beyond current clinical classifications. When combined with longitudinal MRD measurements, this approach may improve the prediction of treatment resistance and therapy-related toxicities. Moreover, as previously demonstrated, *ex vivo* drug profiling can identify subgroups associated with clinical outcomes independently of MRD status[108]. Since MRD alone cannot indicate which therapeutic agents are optimal for treatment, *ex vivo* drug profiling represents a valuable complementary approach particularly in patients who exhibit poor responses[108].

Through integrated molecular characterization, we explored the differences between resistant and sensitive patients and identified molecular signatures that extend beyond traditional subtypes, offering new insights into the pharmacological landscape of BCP-ALL.

Future work may leverage ML-based approaches to prioritize drug combinations for functional screening. Comparisons between untreated and treated samples could uncover temporal resistance mechanisms, expanding our understanding of therapy response and molecular changes under treatment pressure. Finally, incorporating molecular data related to cell cycle and B-cell development could provide additional information relevant to resistance mechanisms[162].

Paper IV

Title and reference: An integrative molecular map of pediatric B-cell precursor acute lymphoblastic leukemia (Krali, O., *et al.*, Manuscript under revision)

Background and Aim

While the molecular landscape of pediatric BCP-ALL has been extensively characterized individually through single omics analyses, systematic integration of multiomics data with *ex vivo* drug responses, and clinical outcomes remains unexplored. Given the heterogeneity BCP-ALL, with its diverse chromosomal and molecular aberrations, high-dimensional multiomics data integration is a promising approach to increase our understanding of ALL biology, variable therapy responses and clinical outcomes. Here, we constructed an integrative molecular map of pediatric BCP-ALL by combining molecular profiles, drug response, and clinical data.

Methods

We applied MOFA to integrate somatic DNAm, GEX, SNVs, and *ex vivo* drug response (EVDR) data from 1,231 pediatric BCP-ALL patients. MOFA reduced the data into a low-dimensional space, represented by cross-modal elements (CMEs) that captured shared variability. The top contributed features for each CME were extracted and used for downstream correlation, network, and pathway analyses, as well as survival modelling to biologically annotate each CME.

Results

Ten distinct CMEs were identified, capturing either modality-specific or cross-modal patterns. Analysis of the top-contributing signatures within CME revealed molecular correlations associated with *ex vivo* drug resistance and clinical outcomes. Importantly, we detected drug/DNAm-based subgroups exhibiting poor prognosis within favorable subtypes (log-rank p-value 0.041), which remained significant after adjusting for risk groups and treatment protocols (Wald test p-value 0.039). Pathway enrichment analysis of the top CME-driven genes and CpG sites revealed involvement in fundamental biological processes including disease development, drug response, cellular regulation, and metabolism (FDR < 0.01). Integration of clinical risk group information with *ex vivo* responses to the top-selected drugs from three CMEs yielded a significant improvement in the classification performance of the baseline clinical model (FDR < 0.05).

Discussion and future perspectives

This study reveals molecular dependencies that transcend single-modality analyses, uncovering otherwise underappreciated relationships and clinically relevant subgroups. Together, these findings emphasize the added value of molecular data integration in refining risk stratification.

A potential limitation is the high cost of generating data across multiple -omics layers. However, recent initiatives are developing multi-omics profiling methods capable of analyzing multiple data modalities in a single run[163]. When these approaches are not feasible, the choice of modality depends on factors such as material availability, quality, and cost. As a result, clinical implementation of data integration remains challenging due to incomplete data modalities across patients. Nonetheless, recent advances in graph- and AI-based integration strategies[148] offer promising solutions for handling partial omics data.

Despite the drawbacks, multi-omics profiling and integration are expected to expand, driving the development of robust frameworks for prioritizing molecular signatures and drug candidates for functional validation and to further unravel the multi-layer complexity of ALL.

Conclusions and future directions

Conclusions

In summary, this thesis explores the multiomic molecular landscape of childhood ALL by leveraging molecular, *ex vivo* drug response, and clinical data from pediatric patients diagnosed and treated in the Nordic countries. Our ALL-based classifier, ALLIUM, enhanced subtype classification in pediatric ALL using a subset of CpG sites and genes. ALLIUM accurately predicted molecular subtypes in both well-characterized and ambiguous (B-other) cases, offering a fast and reliable tool for patient subtyping. Two predictive models, RRP and MRP, built on DNAm data, outperformed the baseline clinical risk group-based models–either individually or in combination, highlighting the added value of molecular profiling to improve risk stratification. *Ex vivo* resistance was associated with poorer clinical outcomes and distinct molecular signatures in treatment-naïve BCP-ALL cells, suggesting that some patients are characterized by molecular profiles linked to treatment resistance. These findings reveal a more intricate disease behavior that extends beyond subtype classification alone. Direct integration of multiple data modalities captured cross-modal signatures and multiomic patterns, revealing associations with dysregulated pathways and subgroups with potential clinical relevance.

Future directions

Ultimately, the four studies presented in the thesis have built a foundation for a data-driven approach to studying complex diseases such as pediatric ALL. The heterogeneity of the disease underscores the need for integrative frameworks that can leverage patient-specific data not only to benefit the patient but also to characterize other cases with similar molecular profiles. While each data modality offers valuable insights on its own, their integration enhances the impact of each. This work opens possibilities both for retrospective analysis in historical cohorts and prospective applications in up-front diagnostics, helping shape the future of precision medicine.

Acknowledgements

My PhD has come to an end, and while I am happy about this achievement, I cannot help but feel sad for leaving this journey behind. These years were intense yet the best years of my adult life. This would not have been possible had it not been for certain people that I am lucky to have in my life. I am grateful for each one of you, who guided me, made me feel welcome, motivated, and inspired me in every step.

First, I would like to thank my supervisor ***Associate Professor Jessica Nordlund*** for many reasons that will not be able to fit in here. Thank you for trusting me to be your first PhD student even when you did not know anything about me. When I was indecisive whether to stay in Sweden or not, you wrote to me: *The projects I have in mind for you are both ambitious and fun, and I would hate to have you (and my group!) miss out on this opportunity to do something really important together*, and it instantly felt *this is where I should be*. I owe you my present, as well as my future. You valued me from the first day, and that made me respect you back. A great supervisor is the one who earns respect by caring, not by spreading fear or by using superiority. During these years, you have supported me in every way possible and listened to me whenever I was frustrated with work. You cared about my wellbeing and my mental health, and gave me the push to have a life beyond my computer and coding. You helped me grow both scientifically and personally and I am grateful for that. You always encouraged my participation in courses and conferences. Our last conference in Milan is one of my sweetest memories. You are such a kind, inspiring, and humble person, and I am sure that this group will continue to grow day by day, as you also bring together people who can coexist peacefully.

As for my co-supervisors, ***professor Arja Harila*** and ***professor Merja Heinäniemi***, I am so happy that you were part of my scientific journey. Both of you have provided me knowledge, data, advice, and smiles, and of course great collaborations. You are great people and I have learned a lot from you. Special thanks to Merja, who hosted me in her lab at the University of Eastern

Finland in Kuopio, and beyond science, she taught me how to ride a sledge on the frozen lake.
I am also grateful to my examiner ***Professor Tove Fall*** for all the support I received throughout my PhD years.

I would like to express my gratitude to my opponent ***Professor Jonathan Bond*** and the defence committee ***Associate Professor Carreras Puigvert***, ***Dr. Anna Robelius***, and ***Associate Professor Nick Tobin***. I would also like to thank ***Associate Professor Maliheh Mehrshad***, who kindly agreed to become the extra committee member for my defense. Thank you all for accepting to be part of this process and review my work.

I could not leave out the amazing ***molecular precision medicine*** (MPM) group. Pandemic hit hard during the first 1-2 years, and I did not have the opportunity to meet you in real life since the beginning, but even through Zoom you were always so nice and welcoming. Things changed drastically after my halftime, as we started spending more time together. I am leaving this group not only with lessons, knowledge, and memories, but also with friends. I am thankful for the beautiful pottery class that was arranged as a surprise for my bachelorette party. You served my childhood memories on a silver plate.

To my fellow PhD partners in crime:

Anna Pia thank you for your clinical knowledge and all ethical approvals you have written for the group, as well as for being eager to code and learn from scratch. I also thank you for your patience and the great care I received from you when I was sick in both conferences we attended in Utrecht and San Diego. I am thankful for the great collaboration in the ALL-FMCA project, as we call it. All these moments brought us closer, so you can imagine how emotional I got the first time you called me your friend, because I feel the same. Thank you so much for being the master mind behind my bachelorette party and all the gyros and souvlaki memes we may have shared.

Mariya thank you for your emotional support–you taught me how not to be apologetic all the time, and take control my impostor syndrome–and for the MINDSET book, which literally changed my life. I am writing a novel thanks to it! Thank you for your coding expertise and the great collaborations we have shared. I can be calm that the classifier project will be in great hands when I depart.

To our postdocs:

Maria, the moment you came for the postdoc interview, I knew that you were a great fit for the group. We have so much in common–except for being Latinas. You have also been a great roommate in all of our conference trips, where we had such a good time together. Thank you for all the beautiful memories we shared, and all the moments you patiently sat and explained biology to me. You inspire me a lot and have learned a lot from you. Special thanks for finding the cozy Thai place, we visit from time to time.

Birong and ***Dea***, even though we have not spent years working together, your positivity showed from day one, and has been enough for me to open up to you. You are always willing to help out with anything I may need and I am grateful for that. I am so sad that I will not spend more time next to you, but I know for sure that we will keep meeting in the future.

To our former master students–my sunshines:

Julia, it has been a pleasure being your supervisor for your master thesis, but mostly your friend. You took your work seriously, but you were also nice and chill. We interacted daily and you were always punctual and dedicated. Whoever is going to employ you, they will be happy to have you. If they complain let me know–I will haunt them forever. I am thankful for meeting you and for all the life and coding lessons I have learned from you. I will miss your little step sounds when you were coming over to ask for feedback. And thank you for my little crochet hamster.

Hanna, you are an amazing person with a great skillset. You are gentle and kind, hence, I think everyone would like to have you in their group. You are very dedicated to your lab work, but you are always keen to know more about the current advancements in bioinformatics. I am thankful that I get to know people with growth mindsets like yours. Thank you for being my beta reader for my novels and for starting our little book club. I would not have survived my days without my baklava.

Emma, who has more Greek friends than I have, is the only one who calls me Olgica, which reminds me of home–it means little Olga if someone wonders.

She is also a reminder of my journey, as both our backgrounds are not purely computational. I witnessed Emma evolving day by day. Her passion about data is profound and that showed from her Master's thesis. The little office became her second home, but she never gave up. She is a true example of a fighter. Her efforts did not go unnoticed as she landed a job as a Bioinformatician. I am thankful that we crossed paths in academia and had such a positive person around during my final year of the PhD.

To the rest of the MPM group:

Henrik, we met years ago, as I started working with Tasos at the same facility as you did at SciLifeLab Solna. You can imagine how happy I felt when you applied for a position in the MPM group. I vouched for you since the beginning, as I knew your work ethos. You are a very valuable member of the group, who is always willing to help out even if your schedule is packed. I could not be more grateful. You are the one who planted the seed about climbing and we have been there quite a few times. Thank you for being the best climbing instructor and for the Greek songs you put during our climbing sessions. I may know less than 20% of them, as I have lived abroad for years, but the music gives me enough motivation to boulder from one side of the wall to the other with one go.

Amanda thank you for all the insightful discussions and your advice about the postdoc positions. Thank you for teaching us so many things about horse asthma and showing that animal research is equally important as the human research. Our round tables during lunch have been very inspiring and I wish I had more time left to enjoy some more.

Anders, you have been a great help with the bioinformatics analysis and finding data in our servers. It takes one data person to know another data person, and I know that having access to the group's data folders is both a blessing and a curse. I really hope you will not get too frustrated if you need to find information from my project folders, but if you do, I made a special folder on Argos for that purpose.

I would also like to thank our former colleague ***Gustav***, who opened my eyes to the word of making great scientific figures and his expertise on gene expression analysis. I would also like to thank another former member, ***Kim*** for all the great talk and for the lovely pie you made when you invited us in Gävle.

Special thanks to my colleagues from the University of Eastern Finland and Tampere University, ***Sanni***, ***Laura***, ***Konstantin***, ***Mari***, ***Buddika***, ***Jonne***, ***Tiina***, and ***Andrea*** for being the best companions during my time in Kuopio, as well as during most international conferences I attended.

To our national and international collaborators ***Gudmar Lönnerholm***, ***Josefine Palle***, ***Claes Andersson***, ***Tom Erkers***, ***Adrián Mosquera Orgueira***, ***Akram Ghantous***, and ***Iñaki Martín-Subero***, whose help and contribution was pivotal for my PhD progress, a wholehearted thank you.

Finally, yet importantly, I would like to express my gratitude to the ***funding agencies*** Barncancerfonden, Cancerfonden, Göran Gustafssons Stiftelser, and Vetenskapsrådet for funding our research and my conference travels, as well as the ***Nordic Society of Hematology and Oncology*** (NOPHO), the ***NOPHO biobank***, the ***patients***, and their ***families***, and ***all people*** who worked hard to generate the data and publish the previous studies, as without their contribution these studies would not be feasible.

Many thanks to ***Karthik*** (the memes master) and ***Anita*** for the best biryani I have ever tasted along with the tears I shed over a kilo of red onions, for the legendary memes that were created after our talks about biryani, and the rest of the ***SNP&SEQ*** platform for the insightful round table discussions. I may have been a ghost for the first years of my PhD, but you were always very welcoming.

I could never forget my great friend, and former supervisor and mentor, ***Tasos***, as he is the reason, I work with Artificial Intelligence today. Similar to Jessica who shaped my present and future, you shaped the past that brought me to this path. I would not be here today without you and for that, I am grateful! Despite I was clueless about GPUs, as I thought that CUDA was gouda, I still managed to learn how to work with one and who Geoffrey Hinton was before he received the Nobel prize.

To my friends, ***Mariangela***, ***Christina***, ***Thano***, ***Ria***, ***Theo***, ***Vasia***, and ***Petrus*** thank you for all the beautiful moments we shared together, including trips, weddings, baby showers, for your failed attempts to teach me ping pong (Mariangela) or to keep up with the Pilates schedule (Ria).

Of course, I would never forget about my little friends, for being such great companions. Every word, every smile, every game means to me more than

you will ever know–at least until you grow up and decide to pursue an academic career. One of you gave me an important list as a present with the following sequence–play, work, play, eat, and have fun. Adults tend to forget that life is more than work. I will treasure your advice for life.

A shout-out to my ***book characters***–you have to read the books to learn who they are–who accompanied me almost every evening this past year, and gave my voice a chance to be heard outside academia. Being a PhD student can be very tough and emotionally draining. I remember that I could not stop thinking about coding even when I was out trying to enjoy my evening. Writing my novels gave me a new sense of purpose and distracted me from the PhD stress. Cheers to them and cheers to the possibility of publishing my books one day!

To my Swedish family, ***Christian***, ***Noreen***, ***Liam***, ***Julia***, and ***Emilia*** I owe a big thanks and a lot of hugs. What started as a babysitting job more than 9 years ago, evolved as something so much more than this. You have always treated me more of a family than an employee. Thank you for everything you have done for me, all life lessons, and for having me in your lives past my babysitting duties.

To the ***Krali-Ricardo Da Costa*** family I owe a huge thanks for helping me grow to the person I am today. Life was not always easy, but you taught me how to chase my dreams and never give up. My ***father Spiros*** taught me to be tough and determined, and my lovely ***mother Deja*** how to be kind and respectful, as well as have empathy and think about others. Both of you taught me to love and care for nature, as it was part of our lives, and this love for nature led to this PhD. Some people might not know it, but my field of expertise was Ecology and I fell in love with data when analyzing my plant inventories. Thank you for all these things and so much more! I want to thank my siblings, ***Andreas*** and ***Giannis***, with whom I may have had a rocky relationship, for all the moments–good and tough. I will never forget when both of you taught me how to play football or how Giannis defended me every time someone bullied me or when Andreas brought me blue roses on my 30th birthday. To my uncle ***Pantelis***, who always says that I take my brains after him, a big thanks, because he may be true. I hope I will continue making you proud. To my aunts ***Katerina*** and ***Chrysoula***, a wholehearted thanks for being my extra mother figures and for all your nice treats. To my godmother ***Rita***, thank you for being here since I was a little girl. My early summers are forever inked inside my mind, as well as your amazing food and lemonade you offer me every time, I visit you in Athens. I will always buy you the chocolate pralines,

even if you complain that you do not want me to spend money. I would like to thank the most recent additions to the family, my two beautiful ***nieces***, who have been our pride and joy. Last but not least, I want to thank my husband's family–***my new family***, for all their support, kindness, hospitality, and of course the food they give us whenever we return to Sweden. After all, I cannot produce qualitive work without eating products from our parents' gardens.

And since a lot of years have passed and I am almost 33, a lot of people have been around and supported me. Everyone is equally important, but if I write about more people, this part will turn longer than my summary. So, I would like to thank all of my teachers, university professors, classmates, old friends, and other family members for their great contribution to what I call the school of life–my life.

To my husband and best friend ***Andreas***, who moved all the way from Greece to give us a shot, a BIG thank you for your unwavering support during the difficult days, even when I, could not stand myself. We have built a beautiful life together, and I could not think of a better partner in crime like yourself. No words can express my gratitude or my feelings for you. But I remember something you told me once. *You said that you would break up with me if I moved back to Greece, if that meant leaving my dreams behind*. Life may require sacrifices, but you never made me choose. My dreams matter to you as much as your own. *And you matter to me as much as my dreams matter to me and so much more.* This summer was the highlight of our common life as we decided to drive from Sweden to Greece, get married, and return back for my PhD defense. You are as crazy as I am!

Closing I would like to make a special dedication to all the people we have lost in our family over the years–both young and old. The pain remains the same. To my paternal grandparents ***Andreas*** and ***Olga***, to my maternal grandparents ***Zoze*** and ***Dejanira***, to my uncle ***Nikos***, my godfather ***Giannis***, my first cousin ***Nikos***, and my uncle ***Dimitris***–each of you have shaped the person I am today. You left a big hole in my heart, but your memories and life lessons are forever carved in my mind and soul. "*Να 'ταν τα νιάτα δυο φορές, τα γηρατειά καμία*", my grandma Olga used to sing to me as we sat at the balcony ("*If only youth came twice, and old age not at all*"). I could not forget about ***grandpa (πάππος) Leonidas***, who passed away earlier this year. You may not have been my blood, but you embraced me as a granddaughter. Thank you for all your advice and the beautiful moments we shared. I will never forget our family meals or your ravani dessert. *May your souls rest in peace, and rest assured that as long as this thesis exists, your names will never be forgotten…*

References

1. Steliarova-Foucher, E. *et al.* International incidence of childhood cancer, 2001–10: a population-based registry study. *Lancet Oncol.* **18**, 719–731 (2017).
2. Marcotte, E. L., Spector, L. G., Mendes-de-Almeida, D. P. & Nelson, H. H. The Prenatal Origin of Childhood Leukemia: Potential Applications for Epidemiology and Newborn Screening. *Front. Pediatr.* **9**, 639479 (2021).
3. Schmiegelow, K. *et al.* Long-term results of NOPHO ALL-92 and ALL-2000 studies of childhood acute lymphoblastic leukemia. *Leukemia* **24**, 345–354 (2010).
4. Toft, N. *et al.* Results of NOPHO ALL2008 treatment for patients aged 1–45 years with acute lymphoblastic leukemia. *Leukemia* **32**, 606–615 (2018).
5. Iacobucci, I. & Mullighan, C. G. Genetic Basis of Acute Lymphoblastic Leukemia. *J. Clin. Oncol.* **35**, 975–983 (2017).
6. Oskarsson, T. *et al.* Relapsed childhood acute lymphoblastic leukemia in the Nordic countries: prognostic factors, treatment and outcome. *Haematologica* **101**, 68–76 (2016).
7. Pölönen, P. *et al.* The genomic basis of childhood T-lineage acute lymphoblastic leukaemia. *Nature* **632**, 1082–1091 (2024).
8. Pölönen, P., Mullighan, C. G. & Teachey, D. T. Classification and risk stratification in T-lineage acute lymphoblastic leukemia. *Blood* **145**, 1464–1474 (2025).
9. Buckley, M., Yeung, D. T., White, D. L. & Eadie, L. N. T-cell acute lymphoblastic leukaemia: subtype prevalence, clinical outcome, and emerging targeted treatments. *Leukemia* **39**, 1294–1310 (2025).
10. Quist-Paulsen, P. *et al.* T-cell acute lymphoblastic leukemia in patients 1–45 years treated with the pediatric NOPHO ALL2008 protocol. *Leukemia* **34**, 347–357 (2020).
11. Schäfer Hackenhaar, F. *et al.* CpG island methylator phenotype classification improves risk assessment in pediatric T-cell acute lymphoblastic leukemia. *Blood* **145**, 2161–2178 (2025).
12. Biondi, A. *et al.* Imatinib after induction for treatment of children and adolescents with Philadelphia-chromosome-positive acute lymphoblastic leukaemia (EsPhALL): a randomised, open-label, intergroup study. *Lancet Oncol.* **13**, 936–945 (2012).
13. Pieters, R. *et al.* A treatment protocol for infants younger than 1 year with acute lymphoblastic leukaemia (Interfant-99): an observational study and a multicentre randomised trial. *The Lancet* **370**, 240–250 (2007).
14. Nyvold, C. *et al.* Precise quantification of minimal residual disease at day 29 allows identification of children with acute lymphoblastic leukemia and an excellent outcome. *Blood* **99**, 1253–1258 (2002).

15. Borowitz, M. J. *et al.* Clinical significance of minimal residual disease in childhood acute lymphoblastic leukemia and its relationship to other prognostic factors: a Children's Oncology Group study. *Blood* **111**, 5477–5485 (2008).
16. Inaba, H. & Pui, C.-H. Immunotherapy in pediatric acute lymphoblastic leukemia. *Cancer Metastasis Rev.* **38**, 595–610 (2019).
17. Gupta, S. *et al.* Blinatumomab in Standard-Risk B-Cell Acute Lymphoblastic Leukemia in Children. *N. Engl. J. Med.* **392**, 875–891 (2025).
18. Toft, N. *et al.* Toxicity profile and treatment delays in NOPHO ALL2008-comparing adults and children with Philadelphia chromosome-negative acute lymphoblastic leukemia. *Eur. J. Haematol.* **96**, 160–169 (2016).
19. Heyman, M. A Treatment Study Protocol for Participants 0-45 Years With Acute Lymphoblastic Leukaemia. (2020).
20. Arber, D. A. *et al.* International Consensus Classification of Myeloid Neoplasms and Acute Leukemias: integrating morphologic, clinical, and genomic data. *Blood* **140**, 1200–1228 (2022).
21. Pui, C.-H. *et al.* Childhood Acute Lymphoblastic Leukemia: Progress Through Collaboration. *J. Clin. Oncol.* **33**, 2938–2948 (2015).
22. Nordkamp, L. O., Mellink, C., van der Schoot, E. & van den Berg, H. Karyotyping, FISH, and PCR in Acute Lymphoblastic Leukemia: Competing or Complementary Diagnostics? *J. Pediatr. Hematol. Oncol.* **31**, 930–935 (2009).
23. Ohsaka, A. *et al.* Spectral karyotyping and fluorescence in situ hybridization analyses identified a novel three-way translocation involving inversion 16 in therapy-related acute myeloid leukemia M4eo. *Cancer Genet. Cytogenet.* **184**, 113–118 (2008).
24. Lilljebjörn, H. & Fioretos, T. New oncogenic subtypes in pediatric B-cell precursor acute lymphoblastic leukemia. *Blood* **130**, 1395–1401 (2017).
25. Berglund, E. *et al.* A Study Protocol for Validation and Implementation of Whole-Genome and -Transcriptome Sequencing as a Comprehensive Precision Diagnostic Test in Acute Leukemias. *Front. Med.* **9**, 842507 (2022).
26. Mullighan, C. G. Genomic Characterization of Childhood Acute Lymphoblastic Leukemia. *Semin. Hematol.* **50**, 314–324 (2013).
27. Ryan, S. L. *et al.* Whole genome sequencing provides comprehensive genetic testing in childhood B-cell acute lymphoblastic leukaemia. *Leukemia* (2023) doi:10.1038/s41375-022-01806-8.
28. 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).
29. Yasuda, T. *et al.* Recurrent DUX4 fusions in B cell acute lymphoblastic leukemia of adolescents and young adults. *Nat. Genet.* **48**, 569–574 (2016).
30. Gu, Z. *et al.* Genomic analyses identify recurrent MEF2D fusions in acute lymphoblastic leukaemia. *Nat. Commun.* **7**, 13331 (2016).
31. Ohki, K. *et al.* Clinical and molecular characteristics of *MEF2D* fusion-positive B-cell precursor acute lymphoblastic leukemia in childhood, including a novel translocation resulting in *MEF2D-HNRNPH1* gene fusion. *Haematologica* **104**, 128–137 (2019).
32. Dickerson, K. M. *et al.* ZNF384 Fusion Oncoproteins Drive Lineage Aberrancy in Acute Leukemia. *Blood Cancer Discov.* **3**, 240–263 (2022).
33. 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).
34. Shago, M., Abla, O., Hitzler, J., Weitzman, S. & Abdelhaleem, M. Frequency and outcome of pediatric acute lymphoblastic leukemia with *ZNF384* gene

rearrangements including a novel translocation resulting in an *ARID1B/ZNF384* gene fusion. *Pediatr. Blood Cancer* **63**, 1915–1921 (2016).
35. Boer, J. M. *et al.* Favorable outcome of NUTM1-rearranged infant and pediatric B cell precursor acute lymphoblastic leukemia in a collaborative international study. *Leukemia* **35**, 2978–2982 (2021).
36. 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).
37. Cario, G. *et al.* BCR-ABL1-like acute lymphoblastic leukemia in childhood and targeted therapy. *Haematologica* **105**, 2200–2204 (2020).
38. Tasian, S. K., Loh, M. L. & Hunger, S. P. Philadelphia chromosome–like acute lymphoblastic leukemia. *Blood* **130**, 2064–2072 (2017).
39. Zaliova, M. *et al. ETV6/RUNX1* -like acute lymphoblastic leukemia: A novel B-cell precursor leukemia subtype associated with the CD27/CD44 immunophenotype: ZALIOVA et al. *Genes. Chromosomes Cancer* **56**, 608–616 (2017).
40. Bastian, L. *et al.* PAX5 biallelic genomic alterations define a novel subgroup of B-cell precursor acute lymphoblastic leukemia. *Leukemia* **33**, 1895–1909 (2019).
41. Marincevic-Zuniga, Y. *et al.* PAX5-ESRRB is a recurrent fusion gene in B-cell precursor pediatric acute lymphoblastic leukemia. *Haematologica* **101**, e20–e23 (2016).
42. Passet, M. *et al.* PAX5 P80R mutation identifies a novel subtype of B-cell precursor acute lymphoblastic leukemia with favorable outcome. *Blood* **133**, 280–284 (2019).
43. Edwards, J. R. *et al.* Chromatin and sequence features that define the fine and gross structure of genomic methylation patterns. *Genome Res.* **20**, 972–980 (2010).
44. Jeziorska, D. M. *et al.* DNA methylation of intragenic CpG islands depends on their transcriptional activity during differentiation and disease. *Proc. Natl. Acad. Sci.* **114**, (2017).
45. Gardiner-Garden, M. & Frommer, M. CpG Islands in vertebrate genomes. *J. Mol. Biol.* **196**, 261–282 (1987).
46. Jones, P. A. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat. Rev. Genet.* **13**, 484–492 (2012).
47. Moore, L. D., Le, T. & Fan, G. DNA Methylation and Its Basic Function. *Neuropsychopharmacology* **38**, 23–38 (2013).
48. Loyfer, N. *et al. A Human DNA Methylation Atlas Reveals Principles of Cell Type-Specific Methylation and Identifies Thousands of Cell Type-Specific Regulatory Elements*. http://biorxiv.org/lookup/doi/10.1101/2022.01.24.477547 (2022) doi:10.1101/2022.01.24.477547.
49. Smith, Z. D. & Meissner, A. DNA methylation: roles in mammalian development. *Nat. Rev. Genet.* **14**, 204–220 (2013).
50. Illingworth, R. *et al.* A Novel CpG Island Set Identifies Tissue-Specific Methylation at Developmental Gene Loci. *PLoS Biol.* **6**, e22 (2008).
51. Zeng, J., Nagrajan, H. K. & Yi, S. V. Fundamental diversity of human CpG islands at multiple biological levels. *Epigenetics* **9**, 483–491 (2014).
52. Lock, L. F., Takagi, N. & Martin, G. R. Methylation of the Hprt gene on the inactive X occurs after chromosome inactivation. *Cell* **48**, 39–46 (1987).
53. Bachman, K. E. *et al.* Histone modifications and silencing prior to DNA methylation of a tumor suppressor gene. *Cancer Cell* **3**, 89–95 (2003).

54. Mutskov, V. & Felsenfeld, G. Silencing of transgene transcription precedes methylation of promoter DNA and histone H3 lysine 9. *EMBO J.* **23**, 138–149 (2004).
55. Challen, G. A. *et al.* Dnmt3a is essential for hematopoietic stem cell differentiation. *Nat. Genet.* **44**, 23–31 (2012).
56. Figueroa, M. E. *et al.* Leukemic IDH1 and IDH2 Mutations Result in a Hypermethylation Phenotype, Disrupt TET2 Function, and Impair Hematopoietic Differentiation. *Cancer Cell* **18**, 553–567 (2010).
57. Flavahan, W. A., Gaskell, E. & Bernstein, B. E. Epigenetic plasticity and the hallmarks of cancer. *Science* **357**, eaal2380 (2017).
58. Nguyen, C. *et al.* Susceptibility of Nonpromoter CpG Islands to De Novo Methylation in Normal and Neoplastic Cells. *JNCI J. Natl. Cancer Inst.* **93**, 1465–1472 (2001).
59. Hetzel, S. *et al.* Acute lymphoblastic leukemia displays a distinct highly methylated genome. *Nat. Cancer* **3**, 768–782 (2022).
60. Nordlund, J. *et al.* Genome-wide signatures of differential DNA methylation in pediatric acute lymphoblastic leukemia. *Genome Biol.* **14**, r105 (2013).
61. Wahlberg, P. *et al.* DNA methylome analysis of acute lymphoblastic leukemia cells reveals stochastic *de novo* DNA methylation in CpG islands. *Epigenomics* **8**, 1367–1387 (2016).
62. Takeuchi, S. *et al.* Clinical significance of aberrant DNA methylation in childhood acute lymphoblastic leukemia. *Leuk. Res.* **35**, 1345–1349 (2011).
63. Wong, N. C. *et al.* A distinct DNA methylation signature defines pediatric pre-B cell acute lymphoblastic leukemia. *Epigenetics* **7**, 535–541 (2012).
64. Issa, J.-P. CpG island methylator phenotype in cancer. *Nat. Rev. Cancer* **4**, 988–993 (2004).
65. Borssén, M. *et al.* Promoter DNA Methylation Pattern Identifies Prognostic Subgroups in Childhood T-Cell Acute Lymphoblastic Leukemia. *PLoS ONE* **8**, e65373 (2013).
66. Borssén, M. *et al.* DNA Methylation Adds Prognostic Value to Minimal Residual Disease Status in Pediatric T-Cell Acute Lymphoblastic Leukemia. *Pediatr. Blood Cancer* **63**, 1185–1192 (2016).
67. Borssén, M. *et al.* DNA methylation holds prognostic information in relapsed precursor B-cell acute lymphoblastic leukemia. *Clin. Epigenetics* **10**, 31 (2018).
68. Duran-Ferrer, M. *et al.* The proliferative history shapes the DNA methylome of B-cell tumors and predicts clinical outcome. *Nat. Cancer* **1**, 1066–1081 (2020).
69. Gabbutt, C. *et al.* Evolutionary dynamics of 1,976 lymphoid malignancies predict clinical outcome. Preprint at https://doi.org/10.1101/2023.11.10.23298336 (2023).
70. Milani, L. *et al.* DNA methylation for subtype classification and prediction of treatment outcome in patients with childhood acute lymphoblastic leukemia. *Blood* **115**, 1214–1225 (2010).
71. Nordlund, J. *et al.* DNA methylation-based subtype prediction for pediatric acute lymphoblastic leukemia. *Clin. Epigenetics* **7**, 11 (2015).
72. Lee, S.-T. *et al.* Epigenetic remodeling in B-cell acute lymphoblastic leukemia occurs in two tracks and employs embryonic stem cell-like signatures. *Nucleic Acids Res.* **43**, 2590–2602 (2015).
73. Dirks, R. A. M., Stunnenberg, H. G. & Marks, H. Genome-wide epigenomic profiling for biomarker discovery. *Clin. Epigenetics* **8**, 122 (2016).

74. De Abreu, A. R. *et al.* Comparison of current methods for genome-wide DNA methylation profiling. *Epigenetics Chromatin* **18**, 57 (2025).
75. Sigurpalsdottir, B. D. *et al.* A comparison of methods for detecting DNA methylation from long-read sequencing of human genomes. *Genome Biol.* **25**, 69 (2024).
76. Liu, Y. *et al.* Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. *Nat. Biotechnol.* **31**, 142–147 (2013).
77. Shen, J. *et al.* Exploring genome-wide DNA methylation profiles altered in hepatocellular carcinoma using Infinium HumanMethylation 450 BeadChips. *Epigenetics* **8**, 34–43 (2013).
78. Touleimat, N. & Tost, J. Complete pipeline for Infinium ® Human Methylation 450K BeadChip data processing using subset quantile normalization for accurate DNA methylation estimation. *Epigenomics* **4**, 325–341 (2012).
79. Reinius, L. E. *et al.* Differential DNA Methylation in Purified Human Blood Cells: Implications for Cell Lineage and Studies on Disease Susceptibility. *PLoS ONE* **7**, e41361 (2012).
80. Aryee, M. J. *et al.* Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* **30**, 1363–1369 (2014).
81. Dedeurwaerder, S. *et al.* Evaluation of the Infinium Methylation 450K technology. *Epigenomics* **3**, 771–784 (2011).
82. Pidsley, R. *et al.* Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. *Genome Biol.* **17**, 208 (2016).
83. Sandoval, J. *et al.* Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. *Epigenetics* **6**, 692–702 (2011).
84. Stirzaker, C., Taberlay, P. C., Statham, A. L. & Clark, S. J. Mining cancer methylomes: prospects and challenges. *Trends Genet.* **30**, 75–84 (2014).
85. Wang, Z., Wu, X. & Wang, Y. A framework for analyzing DNA methylation data from Illumina Infinium HumanMethylation450 BeadChip. *BMC Bioinformatics* **19**, 115 (2018).
86. Daenekas, B. *et al.* Conumee 2.0: enhanced copy-number variation analysis from DNA methylation arrays for humans and mice. *Bioinformatics* **40**, btae029 (2024).
87. Marzouka, N. *et al.* CopyNumber450kCancer: baseline correction for accurate copy number calling from the 450k methylation array. *Bioinformatics* **32**, 1080–1082 (2016).
88. Nurk, S. *et al.* The complete sequence of a human genome. *Science* **376**, 44–53 (2022).
89. Levine, A. J. & Puzio-Kuter, A. M. The Control of the Metabolic Switch in Cancers by Oncogenes and Tumor Suppressor Genes. *Science* **330**, 1340–1344 (2010).
90. Anastasiadou, E., Jacob, L. S. & Slack, F. J. Non-coding RNA networks in cancer. *Nat. Rev. Cancer* **18**, 5–18 (2018).
91. Statello, L., Guo, C.-J., Chen, L.-L. & Huarte, M. Gene regulation by long non-coding RNAs and its biological functions. *Nat. Rev. Mol. Cell Biol.* **22**, 96–118 (2021).
92. Moarii, M., Boeva, V., Vert, J.-P. & Reyal, F. Changes in correlation between promoter methylation and gene expression in cancer. *BMC Genomics* **16**, 873 (2015).

93. Bumgarner, R. Overview of DNA Microarrays: Types, Applications, and Their Future. *Curr. Protoc. Mol. Biol.* **101**, (2013).
94. Kukurba, K. R. & Montgomery, S. B. RNA Sequencing and Analysis. *Cold Spring Harb. Protoc.* **2015**, pdb.top084970 (2015).
95. Byron, S. A., Van Keuren-Jensen, K. R., Engelthaler, D. M., Carpten, J. D. & Craig, D. W. Translating RNA sequencing into clinical diagnostics: opportunities and challenges. *Nat. Rev. Genet.* **17**, 257–271 (2016).
96. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10**, 57–63 (2009).
97. Li, J.-F. *et al.* Transcriptional landscape of B cell precursor acute lymphoblastic leukemia based on an international study of 1,223 cases. *Proc. Natl. Acad. Sci.* **115**, (2018).
98. Mitelman, F., Johansson, B. & Mertens, F. Fusion genes and rearranged genes as a linear function of chromosome aberrations in cancer. *Nat. Genet.* **36**, 331–334 (2004).
99. Zaliova, M. *et al.* Genomic landscape of pediatric B-other acute lymphoblastic leukemia in a consecutive European cohort. *Haematologica* **104**, 1396–1406 (2019).
100. Beder, T. *et al.* The Gene Expression Classifier ALLCatchR Identifies B-cell Precursor ALL Subtypes and Underlying Developmental Trajectories Across Age. *HemaSphere* **7**, e939 (2023).
101. Mäkinen, V.-P., Rehn, J., Breen, J., Yeung, D. & White, D. L. Multi-Cohort Transcriptomic Subtyping of B-Cell Acute Lymphoblastic Leukemia. *Int. J. Mol. Sci.* **23**, 4574 (2022).
102. Schmidt, B. *et al.* ALLSorts: an RNA-Seq subtype classifier for B-cell acute lymphoblastic leukemia. *Blood Adv.* **6**, 4093–4097 (2022).
103. Gu, A. *et al.* TALLSorts: a T-cell acute lymphoblastic leukemia subtype classifier using RNA-seq expression data. *Blood Adv.* **7**, 7402–7406 (2023).
104. Brown, L. M. *et al.* The application of RNA sequencing for the diagnosis and genomic classification of pediatric acute lymphoblastic leukemia. *Blood Adv.* **4**, 930–942 (2020).
105. Hu, Z. *et al.* MD-ALL: an integrative platform for molecular diagnosis of B-acute lymphoblastic leukemia. *Haematologica* (2023) doi:10.3324/haematol.2023.283706.
106. Van Outersterp, I. *et al.* Tyrosine kinase inhibitor resistance in de novo *BCR* :: *ABL1* –positive BCP-ALL beyond kinase domain mutations. *Blood Adv.* **8**, 1835–1845 (2024).
107. Iacobucci, I. & Papayannidis, C. SOHO State of the Art Updates and Next Questions | Approach to BCR::ABL1-Like Acute Lymphoblastic Leukemia. *Clin. Lymphoma Myeloma Leuk.* **25**, 13–22 (2025).
108. Lee, S. H. R. *et al.* Pharmacotypes across the genomic landscape of pediatric acute lymphoblastic leukemia and impact on treatment response. *Nat. Med.* **29**, 170–179 (2023).
109. Liebers, N. *et al.* Ex vivo drug response profiling for response and outcome prediction in hematologic malignancies: the prospective non-interventional SMARTrial. *Nat. Cancer* **4**, 1648–1659 (2023).
110. Peterziel, H. *et al.* Drug sensitivity profiling of 3D tumor tissue cultures in the pediatric precision oncology program INFORM. *Npj Precis. Oncol.* **6**, 94 (2022).
111. Frismantas, V. *et al.* Ex vivo drug response profiling detects recurrent sensitivity patterns in drug-resistant acute lymphoblastic leukemia. *Blood* **129**, e26–e37 (2017).

112. Frost, B. *et al.* Increased *in vitro* cellular drug resistance is related to poor outcome in high-risk childhood acute lymphoblastic leukaemia. *Br. J. Haematol.* **122**, 376–385 (2003).
113. Kaspers, G. J. L. *et al.* In Vitro Cellular Drug Resistance and Prognosis in Newly Diagnosed Childhood Acute Lymphoblastic Leukemia. *Blood* **90**, 2723–2729 (1997).
114. Riss, T. L. *et al.* Cell Viability Assays. in *Assay Guidance Manual* (eds. Markossian, S. et al.) (Eli Lilly & Company and the National Center for Advancing Translational Sciences, Bethesda (MD), 2004).
115. Den Boer, M. L. *et al.* Patient Stratification Based on Prednisolone-Vincristine-Asparaginase Resistance Profiles in Children With Acute Lymphoblastic Leukemia. *J. Clin. Oncol.* **21**, 3262–3268 (2003).
116. Holleman, A. *et al.* Gene-Expression Patterns in Drug-Resistant Acute Lymphoblastic Leukemia Cells and Response to Treatment. *N. Engl. J. Med.* **351**, 533–542 (2004).
117. Gocho, Y. *et al.* Network-based systems pharmacology reveals heterogeneity in LCK and BCL2 signaling and therapeutic sensitivity of T-cell acute lymphoblastic leukemia. *Nat. Cancer* **2**, 284–299 (2021).
118. Lönnerholm, G. *et al.* In vitro cellular drug resistance adds prognostic information to other known risk-factors in childhood acute lymphoblastic leukemia. *Leuk. Res.* **35**, 472–478 (2011).
119. Lindhagen, E., Nygren, P. & Larsson, R. The fluorometric microculture cytotoxicity assay. *Nat. Protoc.* **3**, 1364–1369 (2008).
120. Dada, E. G. *et al.* Machine learning for email spam filtering: review, approaches and open research problems. *Heliyon* **5**, e01802 (2019).
121. Vijh, M., Chandola, D., Tikkiwal, V. A. & Kumar, A. Stock Closing Price Prediction using Machine Learning Techniques. *Procedia Comput. Sci.* **167**, 599–606 (2020).
122. Rasouli, K., Hsieh, W. W. & Cannon, A. J. Daily streamflow forecasting by machine learning methods with weather and climate inputs. *J. Hydrol.* **414–415**, 284–293 (2012).
123. OpenAI *et al.* GPT-4 Technical Report. Preprint at https://doi.org/10.48550/ARXIV.2303.08774 (2023).
124. Brown, T. B. *et al.* Language Models are Few-Shot Learners. Preprint at http://arxiv.org/abs/2005.14165 (2020).
125. Park, D. J. *et al.* Development of machine learning model for diagnostic disease prediction based on laboratory tests. *Sci. Rep.* **11**, 7567 (2021).
126. Uddin, S., Khan, A., Hossain, M. E. & Moni, M. A. Comparing different supervised machine learning algorithms for disease prediction. *BMC Med. Inform. Decis. Mak.* **19**, 281 (2019).
127. Li, Z., Wang, Y., Ganan-Gomez, I., Colla, S. & Do, K.-A. A machine learning-based method for automatically identifying novel cells in annotating single-cell RNA-seq data. *Bioinformatics* **38**, 4885–4892 (2022).
128. Pasquini, G., Rojo Arias, J. E., Schäfer, P. & Busskamp, V. Automated methods for cell type annotation on scRNA-seq data. *Comput. Struct. Biotechnol. J.* **19**, 961–969 (2021).
129. Fortino, V. *et al.* Machine-learning–driven biomarker discovery for the discrimination between allergic and irritant contact dermatitis. *Proc. Natl. Acad. Sci.* **117**, 33474–33485 (2020).
130. Sarker, I. H. Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Comput. Sci.* **2**, (2021).

131. Inamdar, R., Sundarr, S. K., Khandelwal, D., Sahu, V. D. & Katal, N. A comprehensive review on safe reinforcement learning for autonomous vehicle control in dynamic environments. *E-Prime - Adv. Electr. Eng. Electron. Energy* **10**, 100810 (2024).
132. Maćkiewicz, A. & Ratajczak, W. Principal components analysis (PCA). *Comput. Geosci.* **19**, 303–342 (1993).
133. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. Preprint at https://doi.org/10.48550/ARXIV.1802.03426 (2018).
134. McInnes, L., Healy, J., Saul, N. & Großberger, L. UMAP: Uniform Manifold Approximation and Projection. *J. Open Source Softw.* **3**, 861 (2018).
135. Tibshirani, R., Hastie, T., Narasimhan, B. & Chu, G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl. Acad. Sci.* **99**, 6567–6572 (2002).
136. Nordlund, J. *et al.* Digital gene expression profiling of primary acute lymphoblastic leukemia cells. *Leukemia* **26**, 1218–1227 (2012).
137. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
138. Sebastian, P. scikit-survival: A Library for Time-to-Event Analysis Built on Top of scikit-learn. *J. Mach. Learn. Res.* **21**, 1–6 (2020).
139. Rosati, D. *et al.* Differential gene expression analysis pipelines and bioinformatic tools for the identification of specific biomarkers: A review. *Comput. Struct. Biotechnol. J.* **23**, 1154–1168 (2024).
140. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
141. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47–e47 (2015).
142. Hanahan, D. Hallmarks of Cancer: New Dimensions. *Cancer Discov.* **12**, 31–46 (2022).
143. Leo, I. R. *et al.* Integrative multi-omics and drug response profiling of childhood acute lymphoblastic leukemia cell lines. *Nat. Commun.* **13**, 1691 (2022).
144. Argelaguet, R. *et al.* Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. *Mol. Syst. Biol.* **14**, e8124 (2018).
145. Argelaguet, R. *et al.* MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biol.* **21**, 111 (2020).
146. Schroeder, M. P. *et al.* Integrated analysis of relapsed B-cell precursor Acute Lymphoblastic Leukemia identifies subtype-specific cytokine and metabolic signatures. *Sci. Rep.* **9**, 4188 (2019).
147. Isobe, T. *et al.* Multi-omics analysis defines highly refractory RAS burdened immature subgroup of infant acute lymphoblastic leukemia. *Nat. Commun.* **13**, 4501 (2022).
148. Ma, S. *et al.* Moving towards genome-wide data integration for patient stratification with Integrate Any Omics. *Nat. Mach. Intell.* **7**, 29–42 (2025).
149. Rohart, F., Gautier, B., Singh, A. & Lê Cao, K.-A. mixOmics: An R package for ‘omics feature selection and multiple data integration. *PLOS Comput. Biol.* **13**, e1005752 (2017).
150. Subramanian, I., Verma, S., Kumar, S., Jere, A. & Anamika, K. Multi-omics Data Integration, Interpretation, and Its Application. *Bioinforma. Biol. Insights* **14**, 117793221989905 (2020).
151. Lindqvist, C. M. *et al.* The Mutational Landscape in Pediatric Acute Lymphoblastic Leukemia Deciphered by Whole Genome Sequencing. *Hum. Mutat.* **36**, 118–128 (2015).

152. Lindqvist, C. M. *et al.* Deep targeted sequencing in pediatric acute lymphoblastic leukemia unveils distinct mutational patterns between genetic subtypes and novel relapse-associated genes. *Oncotarget* **7**, 64071–64088 (2016).
153. Nordlund, J. *et al.* Refined detection and phasing of structural aberrations in pediatric acute lymphoblastic leukemia by linked-read whole-genome sequencing. *Sci. Rep.* **10**, 2512 (2020).
154. Sayyab, S. *et al.* Mutational patterns and clonal evolution from diagnosis to relapse in pediatric acute lymphoblastic leukemia. *Sci. Rep.* **11**, 15988 (2021).
155. Marincevic-Zuniga, Y. *et al.* Transcriptome sequencing in pediatric acute lymphoblastic leukemia identifies fusion genes associated with distinct DNA methylation profiles. *J. Hematol. Oncol.J Hematol Oncol* **10**, 148 (2017).
156. Lysenkova Wiklander, M., Zachariah, D., Krali, O. & Nordlund, J. Error Reduction in Leukemia Machine Learning Classification With Conformal Prediction. *JCO Clin. Cancer Inform.* e2400324 (2025) doi:10.1200/CCI-24-00324.
157. Marchi, F. *et al.* Epigenomic diagnosis and prognosis of Acute Myeloid Leukemia. *Nat. Commun.* **16**, 6961 (2025).
158. Nordlund, J. & Syvänen, A.-C. Epigenetics in pediatric acute lymphoblastic leukemia. *Semin. Cancer Biol.* **51**, 129–138 (2018).
159. Krali, O. *et al.* Multimodal classification of molecular subtypes in pediatric acute lymphoblastic leukemia. *Npj Precis. Oncol.* **7**, 131 (2023).
160. Krali, O. *et al.* DNA Methylation Signatures Predict Cytogenetic Subtype and Outcome in Pediatric Acute Myeloid Leukemia (AML). *Genes* **12**, 895 (2021).
161. Horvath, S. & Raj, K. DNA methylation-based biomarkers and the epigenetic clock theory of ageing. *Nat. Rev. Genet.* **19**, 371–384 (2018).
162. Iacobucci, I. *et al.* Multipotent lineage potential in B cell acute lymphoblastic leukemia is associated with distinct cellular origins and clinical features. *Nat. Cancer* (2025) doi:10.1038/s43018-025-00987-2.
163. Iacobucci, I., Witkowski, M. T. & Mullighan, C. G. Single-cell analysis of acute lymphoblastic and lineage-ambiguous leukemia: approaches and molecular insights. *Blood* **141**, 356–368 (2023).

Acta Universitatis Upsaliensis

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine 2179

Editor: The Dean of the Faculty of Medicine

A doctoral dissertation from the Faculty of Medicine, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine. (Prior to January, 2005, the series was published under the title "Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine".)

Distribution: publications.uu.se
urn:nbn:se:uu:diva-563468

ACTA UNIVERSITATIS
UPSALIENSIS
2025