

<https://doi.org/10.1038/s43856-026-01568-9>

An integrative molecular map of pediatric B-cell precursor acute lymphoblastic leukemia

Check for updates

Olga Kralli ^{1,2}, Anna Pia Enblad^{1,2,3}, Julia Sulyaeva^{1,2}, Dea Gogishvili^{1,2}, Anders Lundmark^{1,2}, Arja Harila³, Claes Andersson¹, Tom Erkers^{4,5}, Merja Heinäniemi ⁶, Gudmar Lönnnerholm³ & Jessica Nordlund ^{1,2} ✉

Abstract

Background The molecular landscape of pediatric B-cell precursor acute lymphoblastic leukemia (BCP-ALL) has been extensively characterized through single-modality studies. However, the interplay between molecular modalities and their collective influence on treatment response and outcomes remains poorly understood.

Methods We integrated genomic, epigenomic, transcriptomic, and ex vivo drug response data from 1231 patients diagnosed with BCP-ALL. Using Multi-Omics Factor Analysis, we identified signatures explaining key aspects of the integrative molecular landscape, referred to as cross-modal elements (CMEs). The CME-derived signatures were introduced into pathway and intermodal network analyses, while their impact on patient outcomes was assessed through survival modeling.

Results Pathway and network analyses annotate the resulting integrative CMEs, linking them to key biological processes, including disease development, cellular regulatory processes, metabolic pathways, and drug response. By leveraging correlations between DNA methylation and ex vivo response to doxorubicin, we stratify patients with hyperdiploidy into subgroups that differ in relapse-free survival. These signatures are independent of clinical variables. Survival models incorporating CME-selected ex-vivo drug responses combined with clinical data improve risk prediction compared to clinical models alone (FDR < 0.05), demonstrating the potential of integrative multiomics in refining risk stratification.

Conclusions Our study highlights the importance of multimodal data integration in BCP-ALL to provide biological insights with potential relevance for precision medicine.

Plain language summary

Pediatric BCP-ALL is the most common childhood cancer. While most children are cured today, some still relapse, die from the disease, or experience severe treatment-related side effects. To better understand why outcomes differ, we analyzed data from 1,231 children with BCP-ALL, integrating multiple layers of information from their leukemia cells. This approach revealed biological patterns related to leukemia subtype, cell growth, and treatment response. We also found that including information on how leukemia cells respond to treatment improved relapse prediction compared with using clinical data alone. Together, our findings show that combining different types of data can reveal subtle differences between patients and help us better understand the complex biology of childhood leukemia.

Over the past decades, the molecular phenotypes underlying pediatric B-cell precursor acute lymphoblastic leukemia (BCP-ALL) have been extensively studied¹. Strong associations between genetic subtypes and corresponding molecular phenotypes have been well established²⁻⁴. These associations extend beyond recurrent somatic single-nucleotide variants (SNVs)² to include quantitative signatures derived from DNA methylation (DNAm)⁵⁻⁷, gene expression (GEX)⁸⁻¹⁰, protein expression¹¹⁻¹³, and metabolomics data¹⁴. Collectively, these studies have significantly expanded our understanding of ALL pathogenesis, enabling improved risk stratification and therapy monitoring, which have contributed to the remarkable cure rates achieved

in recent years¹. Despite cure rates exceeding 90% for most BCP-ALL patients¹⁵⁻¹⁷, significant challenges remain. Relapsed patients have poor outcomes, with survival rates of ~50%^{17,18}, while others suffer severe side effects and morbidities resulting from overtreatment¹⁶.

Most studies focus on single molecular entities or modality comparisons, despite the fact that molecular drivers of BCP-ALL, like those in other cancers^{19,20}, are likely to operate across multiple interconnected molecular levels²¹. Several studies have successfully integrated more than one data modality^{11,22-27}. Meanwhile, ex vivo drug response (EVDR) profiling is emerging as a promising precision medicine tool, providing direct

¹Department of Medical Sciences, Uppsala University, Uppsala, Sweden. ²SciLifeLab, Uppsala University, Uppsala, Sweden. ³Department of Women's and Children's Health, Uppsala University, Uppsala, Sweden. ⁴Department of Oncology-Pathology, Karolinska Institutet, Stockholm, Sweden. ⁵SciLifeLab, Stockholm, Sweden. ⁶Institute of Biomedicine, School of Medicine, University of Eastern Finland, Kuopio, Finland. ✉e-mail: jessica.nordlund@medsci.uu.se

assessment of how patient-derived cells respond to established and experimental therapies^{28–31}. Bridging the gap between single-modality and integrative, multimodal approaches is essential for deepening our understanding of ALL pathogenesis and accelerating precision therapies.

Several computational methods for inter-modal integration have been developed to better understand complex disease patterns in cancer^{22,25–27,32,33}. Among these, Multi-Omics Factor Analysis (MOFA) is an unsupervised framework that integrates diverse data modalities to identify functional multiomics patterns²². In chronic lymphocytic leukemia (CLL), MOFA revealed latent factors linked with multimodal heterogeneity, identifying previously overlooked pathways, such as oxidative stress response, as key contributors to cellular processes, disease development and progression, and clinical outcomes²². In Myelodysplastic syndrome (MDS), MOFA identified potential prognostic markers and protective mechanisms associated with patient outcome²⁶. In ALL, multiomics integration has been applied to cell-lines¹¹ and data from T-ALL patients³⁷, however, such integration has yet to be applied to data from primary BCP-ALL patient samples.

This study provides an integrative molecular map of pediatric BCP-ALL, combining SNVs, DNAm, GEX, and EVDR data from 1231 patients. By correlating these data with clinical outcomes, we identify cross-modal elements (CMEs) that capture shared heterogeneity, which we interpret in the context of BCP-ALL pathophysiology. Intermodal network analysis detects DNAm signatures with potential for patient stratification, revealing a subgroup of patients with inferior outcomes. Adding CME information to risk models demonstrates its potential for linking functional and molecular signatures to improve risk stratification.

Methods

Patients

Treatment-naïve peripheral blood or bone marrow aspirates were obtained from 1231 pediatric patients with BCP-ALL at the time of diagnosis with blast counts >80%. The children were diagnosed between 1992–2013 and treated according to the Nordic Society of Pediatric Hematology and Oncology (NOPHO) NOPHO ALL-92 ($n = 303$), ALL-2000 ($n = 504$), ALL-2008 ($n = 367$), EsPh-ALL ($n = 17$), or Interfant ($n = 40$) treatment protocols^{16,34–36}. The median age of the patients at diagnosis was 4.5 years (interquartile range, IQR, 2.9–8.4), and the median follow-up time was 17 years (IQR, 13–21). Cytological analysis, immunophenotyping, and cytogenetics of pretreatment leukemic cells enabled the establishment of the molecular diagnosis of BCP-ALL. Molecular subtypes were revised to the current international consensus classification (ICC) as previously described³⁷. Guardians and/or patients provided written informed consent. The study complied with the Declaration of Helsinki and was approved by the NOPHO Scientific Committee (Study #56) and the Regional Ethics Review Authority (reference numbers: DNR 2007/023, 2010/416, 2013/237, 2014/482) in Uppsala, Sweden.

DNA methylation (DNAm)

The methylation status was interrogated using the Infinium HumMeth450K BeadChip assay (450k array, $n = 1012$ patients) or the MethylationEPIC v2.0 BeadChip assay (EPIC array, $n = 55$ patients). The data were processed as previously described^{5,37,38}. Probes located on the X and Y chromosomes and probes known to be affected by underlying genetic variation were filtered out⁵. Only probes overlapping the two array types (450k array and EPIC) were retained, resulting in 372,264 CpG sites across 1067 patients for downstream analysis. For copy number analysis, the methylumi and conumee³⁹ 2.0 R packages were used.

RNA sequencing (GEX)

RNA sequencing (RNA-seq) GEX data were retrieved from GSE227832 and processed as previously described³⁷. Only protein-coding genes were retained. Additionally, genes located in X and Y chromosomes, as well as ribosomal, mitochondrial, and scaffold genes were removed. The final GEX dataset contained protein-coding 18,928 genes in 295 patients for

downstream analysis. The *limma*⁴⁰ R package was used to perform differential gene expression analysis.

Single nucleotide variants (SNVs)

Mutational data (somatic SNVs) from 128 patients were retrieved from previous studies^{37,41–44}. In short, the somatic SNV data were generated either from an 872-cancer gene Haloplex panel^{41,42} ($n = 125$) or/and from whole genome sequencing (WGS, $n = 16$)^{41,43,44}. For the patients with RNA-seq data available ($n = 295$) the variant alleles *PAX5* p.Pro80Arg, *IKZF1* p.Asn159Tyr, and *ZEB2* p.His1038Arg were specifically interrogated as previously described³⁷. In total, SNV data across 529 genes and 128 patients were available for downstream analyses.

Ex vivo drug response (EVDR)

Ex vivo drug response to a panel of ten treatment compounds was assessed using the Fluorometric Microculture Cytotoxicity Assay (FMCA) as previously described⁴⁵. The data represent the fraction of surviving leukemic cells after 72 hours of incubation with each drug at empirically selected concentrations (Supplementary Data S1). These concentrations were chosen to yield survival index (SI) values that capture inter-sample variability³⁸. The SI was calculated as the mean fluorescence signal from wells containing leukemic cells with intact plasma membranes after drug exposure, divided by the mean fluorescence signal from wells containing untreated leukemic cells (control), after subtracting the background signal from blank wells (medium only). This ratio was multiplied by 100 to obtain SI%⁴⁵.

The drugs belong to the glucocorticoid (dexamethasone and prednisolone), topoisomerase II inhibitor (amsacrine, doxorubicin, mitoxantrone, and etoposide), antimetabolite (cytarabine and thioguanine), enzyme (L-asparaginase), and vinca alkaloid (vincristine) classes. In total, EVDR data for 857 patients across 10 drugs were available for downstream analyses.

Multi-omics factor analysis

The data containing all four data modalities (DNAm, GEX, SNV, and EVDR) were initially split into training ($n = 923$) and test ($n = 308$) datasets in a stratified manner to retain the data modality proportions of the entire cohort. Patients treated on different NOPHO treatment protocols were evenly distributed across the two datasets. MOFA was employed using the *mofa2* R library²² on the training dataset. We renamed the MOFA factors as CMEs. To run MOFA, we selected a subset of CpG sites ($n = 5000$) and genes ($n = 10,000$) with the highest variance in the training dataset, and included all of the available SNV ($n = 529$) and EVDR ($n = 10$) features. This variance-based feature selection follows MOFA recommendations to reduce dimensionality and exclude low-variance features. Because RNA-seq data were available for only 24% of patients compared to 87% for DNA methylation, we included a larger number of genes to balance the relative contribution of the transcriptomic modality and prevent overrepresentation of methylation data. The training dataset was used to create a MOFA object and run the analysis, without imputation. The train options were kept as default apart from the convergence mode (medium instead of slow), the number of maximum iterations (700 instead of 1000) and the variance explained threshold (2% instead of none). A seed parameter of 42 was applied when running the analysis. MOFA generates CME values at a patient level and feature weights at a feature level for all CMEs. We extracted the features weights scaled from -1 to 1 at a CME and modality level, where -1 represented the highest negative impact, 1 the highest positive impact, and 0 no impact on each CME. For the GEX and DNAm modalities specifically, we selected the top 1000 genes and CpG sites (based on the highest absolute weights) per CME for downstream analyses. MOFA was used to impute missing values in the training set. A separate MOFA model was built for the test set using the same parameters as the training model, applied to impute missing values independently.

Annotation of CpG sites

The GREAT 4.0.4 annotation tool⁴⁶ was used to annotate CpG sites to genes in proximity based on the distance from a transcription start site (TSS) extended by 5 kb upstream and 1 kb downstream, and up to 10 kb extension in both directions, including regulatory domains curated from literature. For each CME and weight (positive, negative), a bed file was generated containing the start and stop coordinates, and the chromosomal location of each CpG site. A bed file was prepared for the background CpG sites ($n = 5000$ CpG sites).

Pathway analysis

Over-representation analysis (ORA) was performed on the top-ranked CpG sites and genes derived from the MOFA model trained on the training dataset, with absolute weight >0.6 for each CME using the `enrichr()/enrich()` modules from the `gseapy` Python library⁴⁷. The positively and negatively weighted features were analyzed separately. The background was set as the top 5000 annotated CpG sites (1735 gene annotations) and 10,000 ENSEMBL genes (9998 unique gene names), which were used as the input features for MOFA. Each CME/weight gene list was tested on gene sets from gene ontology (GO) biological processes (BP), the molecular signatures database (MSigDB) Hallmark, and seven custom gene lists. The custom lists comprised subtype classifiers, as ALLIUM DNAm (272 annotated CpGs) and GEX (356 genes)³⁷, ALLSorts⁴⁸ (1003 genes) and ALLCatchR⁴⁹ (2,802 genes), cell cycle genes from Seurat⁵⁰ (97 genes), and cell state genes from bulk (145 genes) and single cell RNA-seq (574 genes) ALL samples⁵¹.

Pathway networks and clusters were generated using the enrichment map function of `gseapy`⁴⁷. Pathways with ≤ 1 overlapping gene, as well as CMEs with ≤ 2 pathways in present, were excluded from the analysis. To annotate the resulting clusters, we performed text mining by tokenizing the pathways into individual words using `CountVectorizer` from `scikit-learn`⁵² using the 20 most frequent words per cluster.

Network analysis

The `igraph` R library⁵³ was used to build inter-modal networks. For each CME, the most impactful features (absolute weight >0.6) were retrieved for both training and test datasets. Networks were generated from the training data, and the reproducibility of central dependencies was assessed in the test data. The features with positive and negative weights were analyzed separately. Pairwise correlation matrices (DNAm-GEX, DNAm-EVDR, GEX-EVDR) were generated using the non-imputed data, ensuring the analysis reflects real correlations without bias due to imputation. For each CME, an absolute correlation coefficient cut-off (0.2–0.6) was applied to remove weak correlations among features. The correlation matrices were combined to create an adjacency matrix, which was used as input to create the network graphs. Self-loops, as well as unconnected nodes, were removed from the networks. A random seed 1 was employed for each plot to control the stochasticity of the graphical representations. Finally, the resulting networks were used to extract communities (`cluster_edge_betweenness`) and hub features. Each network consisted of one or multiple clusters (communities) of features. Within these networks, the features with > 3 connections (correlations) were assigned as central points (hubs).

Survival models

Gradient boosted Cox proportional hazards loss survival models⁵⁴ were built using the MOFA-imputed training set in three configurations. A CME feature-based (CMEf, top features with absolute weight >0.6) model, a baseline model trained on clinical risk groups, and a combined model incorporating both CMEf and baseline features. In addition, single-omics (DNAm, GEX, and EVDR) and the CMEf-all features model were constructed. These models were also evaluated combined with the baseline features. This resulted in a total of 83 different models: one clinical baseline model, two CMEf-all models (with and without clinical data), and 20 models each based on top-ranked features from the multiomic (CMEf), DNAm, GEX, and EVDR data, each evaluated alone and in combination with clinical risk grouping.

The survival models link the covariates to the time of an event, which, in our case, corresponds to event free survival (EFS). They constitute a type of regression model, which generates risk scores in relation to an event (event vs. no event). In our cohort, 227 and 81 events occurred in the training and test sets, respectively (relapse, death in complete remission, induction failure, secondary malignant neoplasm, and resistant disease as defined in the protocol in question). EFS is calculated from the time at diagnosis to the occurrence of an event or, for patients who did not experience an event, to the date of the last follow-up.

Each survival model comprised 100 regression trees. Repeated stratified k -fold cross-validation (CV) was performed⁵² using 3×5 repeated splits. A stratified setup retains the same event distribution for the inner training and validations CV sets, as in the outer training dataset ($n = 923$). A CV setup mitigates overfitting, as each model is trained on the inner training sets and is validated on the left-out (CV) data. The CV scores were extracted to perform statistical analysis using BH-adjusted Wilcoxon signed-rank test to assess differences in model performance between the baseline and the molecular models. The hyperparameter space included the model's learning rate (0.01), the number of minimum samples to split a tree node (5), the maximal depth of the tree (5), and the maximal number of features for the best node tree split (square root of #features). As the survival models predict risk scores, Harrell's concordance index (c-index) was used to evaluate their performance. The c-index takes into account the magnitude of the assigned risk in relation to the time that an event has occurred. For instance, a model with a high c-index assigns higher scores to patients who experience an event at a shorter time span. In addition, the models were further validated using the MOFA-imputed test set.

Statistics and reproducibility

All statistical analyses were performed using Python (version 3.8.5) or R (version 4.2.3).

Where applicable, Pearson's correlation coefficients were used to assess relationships between two variables, with corresponding p -values adjusted for multiple testing using the Benjamini-Hochberg procedure (FDR < 0.05).

Enrichment analyses (ORA) relied on hypergeometric tests to obtain significantly enriched pathways (FDR < 0.01).

Pairwise group comparisons were made using two-sided Mann-Whitney U tests. For comparisons involving more than two groups, Dunn's test was applied. Multiple testing correction was performed using the BH method (FDR < 0.05).

Kaplan–Meier survival curves were compared using log-rank test (p -value < 0.05). Multivariable Cox proportional hazards models were fitted to estimate hazard ratios, and statistical significance was assessed using the Wald test (p -value < 0.05). Fisher's exact test was used to assess differences in proportions of good- or poor-risk patients between the low- and high-doxorubicin response clusters (p -value < 0.05).

Two-sided Wilcoxon signed-rank tests were used to evaluate cross validation performance between the baseline clinical model and the CME-based models, followed by BH correction for multiple testing (FDR < 0.05).

Sample sizes varied depending on data availability across molecular modalities and are reported in the corresponding Results sections, figures, or figure legends. No experimental or technical replicates were included in this study. All analyses were conducted using predefined statistical thresholds as described above.

Results

Cohort demographics and data overview

Pre-treatment peripheral blood samples or bone marrow aspirates were retrieved from 1231 children diagnosed with BCP-ALL between 1992 and 2013, who were enrolled in the consecutive NOPHO ALL-92 ($n = 303$), ALL-2000 ($n = 504$), ALL-2008 ($n = 367$), EsPh-ALL ($n = 17$), or Interfant treatment protocols ($n = 40$)^{16,34–36}.

The omics data were obtained from previous studies: Genome-wide DNAm profiling data were obtained from 1,067 patients using Illumina 450k ($n = 1012$)³⁷ or EPIC ($n = 55$) arrays³⁸. RNA-seq data were available

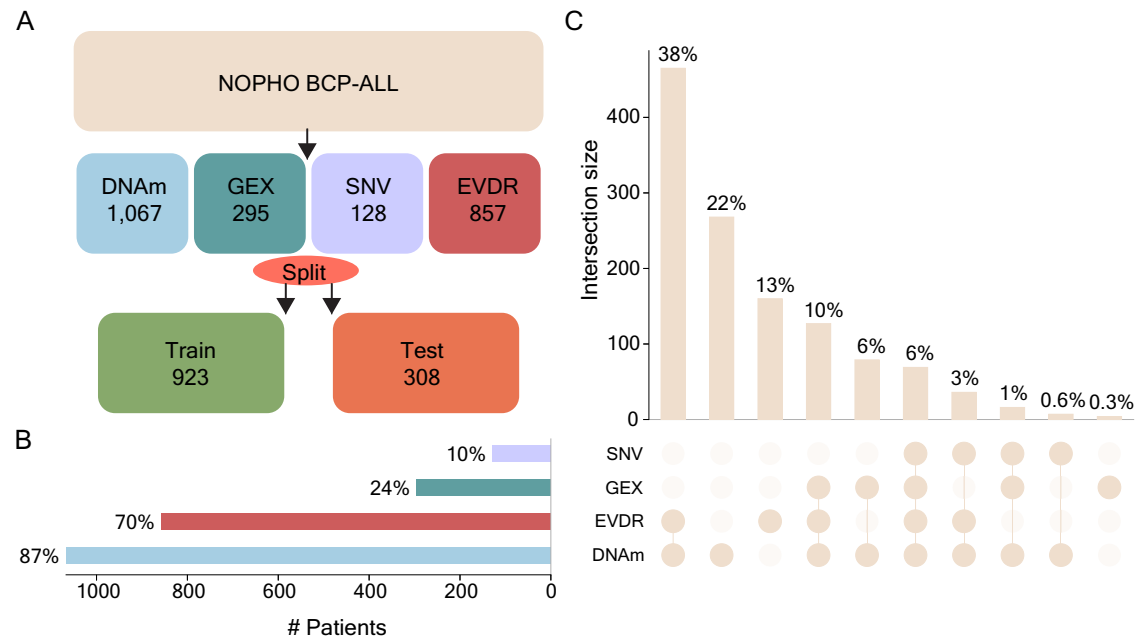


Fig. 1 | Pediatric BCP-ALL samples and data modality overview. **A** DNA methylation (DNAm), gene expression (GEX), somatic single-nucleotide variants (SNVs), and ex vivo drug response data (EVDR) from 1231 pediatric BCP-ALL

patients. The patients were split into training (75%) and test sets (25%). **B** Percentage-wise sample distribution across each modality. **C** Upset plot showing the matched samples between two, three, all, or none of the data modalities.

from 295 patients³⁷. Somatic SNVs were obtained using a 872-cancer gene Haloplex panel ($n = 125$ patients)^{41,42}, WGS ($n = 16$ patients)^{41,43,44}, or by targeted analysis for single-nucleotide variants *PAX5* p.Pro80Arg, *IKZF1* p.Asn159Tyr, and *ZEB2* p.His1038Arg³⁷ ($n = 295$ patients). EVDR data for 10 drugs were generated with the FMCA⁴⁵ for 857 patients (Fig. 1A).

The dataset was split into training ($n = 923$, 75%) and test ($n = 308$, 25%) datasets stratified to balance the representation of DNAm, GEX, SNV, and EVDR data (Fig. 1A, B, Fig. S1). The percentage of matched samples varied, with 65% of samples appearing in at least two modalities, 20% in at least three modalities, and 6% in all four modalities (Fig. 1C).

Cross-modal elements are associated with clinical features

We applied MOFA²² on the training dataset ($n = 923$ patients) using four data modalities: DNAm, GEX, SNV, and EVDR (Fig. S1, Supplementary Data S2). For DNAm and GEX, we retained the 5000 and 10,000 most variable features, respectively, while all dimensions were included for SNV ($n = 529$) and EVDR ($n = 10$) data (Fig. 2A). This analysis generated a matrix (Z) of CME values per patient and a feature weight matrix (W^1, \dots, W^M) for each CME and data modality. MOFA identified 10 CMEs using a total variance explained cut-off of 2%, explaining 55% of variance in DNAm, 48% in GEX, 24% in EVDR, and 0% in SNV data (Fig. 2B). CMEs were categorized as either modality-specific (e.g., CME 3–4, 8, and 10) or inter-modal (e.g., CME 1–2, 5–7, and 9, Fig. 2C), based on a variance explained cut-off of >1%. For modality-specific CMEs, CME 3 explained 17% of the variance in the DNAm data, while CMEs 4 (14%), 8 (8%), and 10 (3%) explained the most variance in the GEX data. Inter-modal CMEs (1–2 and 5–7) captured shared variability across DNAm, GEX, and EVDR data, while CME 9 captured variability shared between DNAm (6%) and GEX data (2%). None of the CMEs captured variability in the SNV data, likely due to a combination of the small number of recurrent SNVs in ALL⁵⁵ and the sparsity of our dataset⁴². Consequently, SNVs were excluded from downstream analyses.

To evaluate model robustness and CME orthogonality, pairwise correlations between CMEs were calculated, revealing no significant codependence between CMEs (Fig. 2D, absolute $\rho < 0.25$). Next, we assessed correlations between CMEs and clinical covariates, including sex, age at diagnosis, outcome, treatment protocol, risk group, and molecular subtype (Fig. 2E). No strong association was observed for sex

and outcomes across CMEs (absolute $\rho < 0.15$, Fig. 2E, Supplementary Data S3). CMEs 4 and 9 were correlated with age at diagnosis ($\rho = -0.34$, $FDR < 0.001$ and $\rho = 0.40$, $FDR < 0.001$, Fig. 2E, Supplementary Data S3). Although it is well established that age at diagnosis is correlated with molecular subtype⁵⁶, no significant correlation was observed between CME 4 and any of the molecular subtypes or risk groups (Fig. 2E). However, CME 9 was positively correlated with the PAX5alt subtype ($\rho = 0.46$, $FDR < 0.001$, Fig. 2E). Additionally, patients treated according to the infant protocol Interfant were negatively associated with CME 9 ($\rho = -0.26$, $FDR < 0.001$, Fig. 2E). Molecular BCP-ALL subtypes were strongly correlated with CME 1 and 2, underscoring the potential of the first CMEs to capture known biology. Specifically, high hyperdiploidy (HeH, $\rho = 0.77$, $FDR < 0.001$) and *ETV6::RUNX1* ($\rho = -0.55$, $FDR < 0.001$) were correlated with CME 1, and *ETV6::RUNX1* ($\rho = -0.73$, $FDR < 0.001$) with CME 2 (Fig. 2F, Supplementary Data S3).

While CMEs 1, 2, 4, and 9 showed significant associations with known clinical parameters, including molecular subtype, age, treatment protocol, and risk group, the remaining CMEs exhibited more subtle and complex interactions across modalities and clinical features, warranting further exploration.

CMEs are enriched for cellular functions and regulatory networks

To better understand the molecular variables contributing to the CMEs, we first focused specifically on the top-ranked features (absolute weight > 0.6) in the DNAm and GEX datasets (Supplementary Data S4). This resulted in 1040 genes and 2681 CpG sites across the ten CMEs (Supplementary Data S5, S6). The CpG sites were annotated to genes using the GREAT annotation tool⁴⁶, resulting in 1027 genes for functional enrichment analysis (Supplementary Data S6). Of these, 64 top-ranked CpG sites were annotated to 36 top-ranked genes within the same CME, indicating a possible relationship between methylation level and transcriptional activity for these genes. Notably, this set included genes well-known to distinguish ALL subtypes and/or are involved in B-cell development: *BIRC7*, *DSC3*, *IGF2BP1*, *TCFL5*, *S100A16*, *PCDH9*, *IRF8*, *HPS4* and *PLVAI*^{37,48,49,51} (Supplementary Data S7).

To identify biological pathways enriched within each CME, we performed ORA using the CME-associated genes. Because the overlap

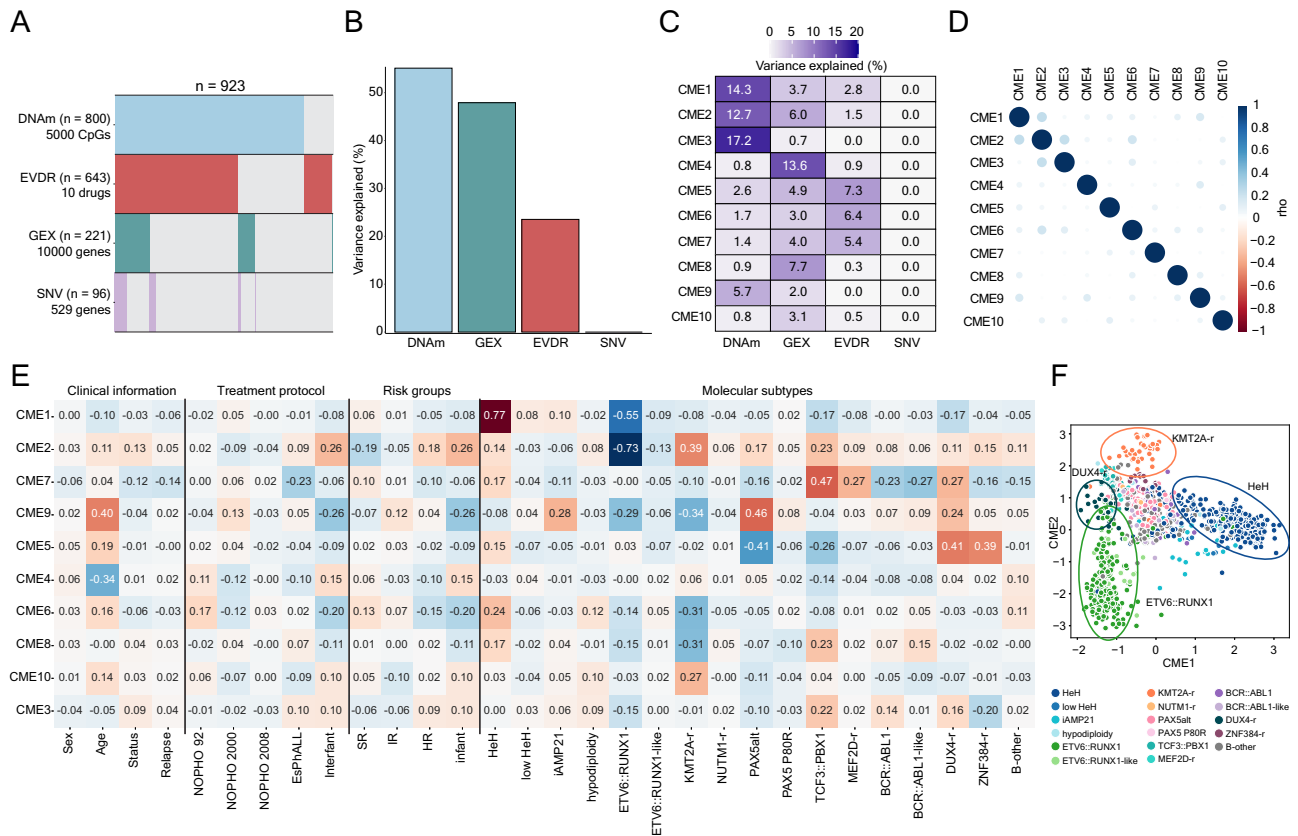


Fig. 2 | Multi-Omics Factor analysis (MOFA) across 923 BCP-ALL patients. **A** Patient distribution per modality in the training dataset (n = 923). Missing data are indicated by light gray color. **B** Percentage of variance explained (R²) by all cross-modal elements (CMEs) per data modality. **C** R² by CME for each data modality. **D** Pairwise correlation between CMEs (Pearson's correlation coefficient, rho). **E** CME correlation (rho) with clinical variables (i.e., sex, age, status (alive vs.

deceased), and relapse (no relapse vs. relapse)), NOPHO treatment protocols, clinical risk groups (SR: standard risk, IR: intermediate risk, HR: high risk), and molecular subtypes. **F** Scatterplot showing the association between CME 1 (x-axis) and CME 2 (y-axis), color-coded by molecular subtype. Subtypes with defined boundaries are highlighted with ellipses.

between genes and annotated CpG sites was minimal, ORA was performed separately for each modality. This analysis revealed enrichment in 195 pathways and seven curated gene lists built based on prior knowledge^{37,48–51} (FDR < 0.01, Supplementary Data S8). Next, we built pathway networks based on the degree of similarity (overlap coefficient > 0) between pathways by CME. Pathways with overlapping genes (> 1 genes) were merged into a single “pathway cluster” and pathways without overlapping genes remained unclustered (Supplementary Data S8). To annotate each “pathway cluster”, we tokenized each pathway into words using CountVectorizer, retaining the 20 most frequent terms. This text mining approach resulted in ten distinct annotations: *molecular ALL subtypes*, *molecular ALL subtypes including B-cell development*, *cell cycle regulation*, *cell cycle regulation including B-cell development*, *immune response*, *metabolic pathways*, *signaling pathways*, *T-cell pathways*, *transcriptional regulation*, and *mRNA processing* (Figs. 3A, S2, Supplementary Data S8). Corroborating our previous observations, the genes contributing to CMEs 1 and 2 were significantly enriched to *molecular ALL subtypes* (Fig. S2). CME 2, which explained 12.7% of the variance in the DNAm modality, was also enriched for *immune response* (Fig. 3B). CME 6, which accounted for 6.4% of the explained variance in the EVDR modality, was enriched for *cell cycle regulation* (Fig. 3C). CME 8, contributing to 7.7% of the variance captured in the GEX data modality, was enriched in *cell cycle regulation*, *metabolic pathways*, and *mRNA processing* (Fig. 3D).

Inter-modal interactions within CMEs

To further explore inter-modal interactions within CMEs, we analyzed GEX, DNAm, and EVDR data from the training (n = 923) and test sets (n = 308)

separately, focusing on the top-ranked genes, CpG sites, and drugs (Supplementary Data S6, S7 and S9). For each CME, we constructed networks independently for positive and negative weights, which were organized into “communities” comprising clusters of related features with central “hub” features. Community size ranged from 2–10 clusters, and the number of hubs per CME ranged from 1–16 (Table 1, Supplementary Data S10, S11). Interactions across the three data modalities were observed in both positive and negative weight networks (Fig. S3, S4, Supplementary Data S10, S11).

For example, CME 2, which strongly correlated with ALL molecular subtypes (Fig. 2E), contained two DNAm communities with etoposide and doxorubicin as hubs (Fig. 4A). Unsupervised hierarchical clustering of DNAm from patients in the training set (n = 800), revealed two clusters of low and high DNAm levels (Fig. 4B). Increased DNAm levels were associated with higher survival indexes to both drugs (Mann-Whitney U test p-value < 0.001, Fig. 4C). Key CpG sites were annotated to ALL subtype-specific genes *IGF2BP1* (n = 1), *TCFL5* (n = 2), *TMED6* (n = 1), *BIRC7* (n = 2), and *DLGAP2* (n = 1)³⁷, and to the immune response gene *CSF2RB* (n = 1). Notably, the expression of *IGF2BP1*, *TCFL5*, *TMED6*, and *BIRC7* were also among the most influential features associated with CME 2 (Supplementary Data S7). Subtype distribution analysis of the low and high methylation clusters for both hubs revealed distinct patterns (Fig. 4D). The low-methylation etoposide-associated cluster was predominantly composed of *ETV6::RUNX1*-positive patients (83.9%, n = 188), indicating that this subtype is closely associated with the methylation status of the CpG sites in the etoposide hub.

In contrast, the doxorubicin hub spanned ALL subtypes, but with distinct subtype distribution. The low-methylation group was enriched for subtypes with generally favorable prognosis, including *ETV6::RUNX1*

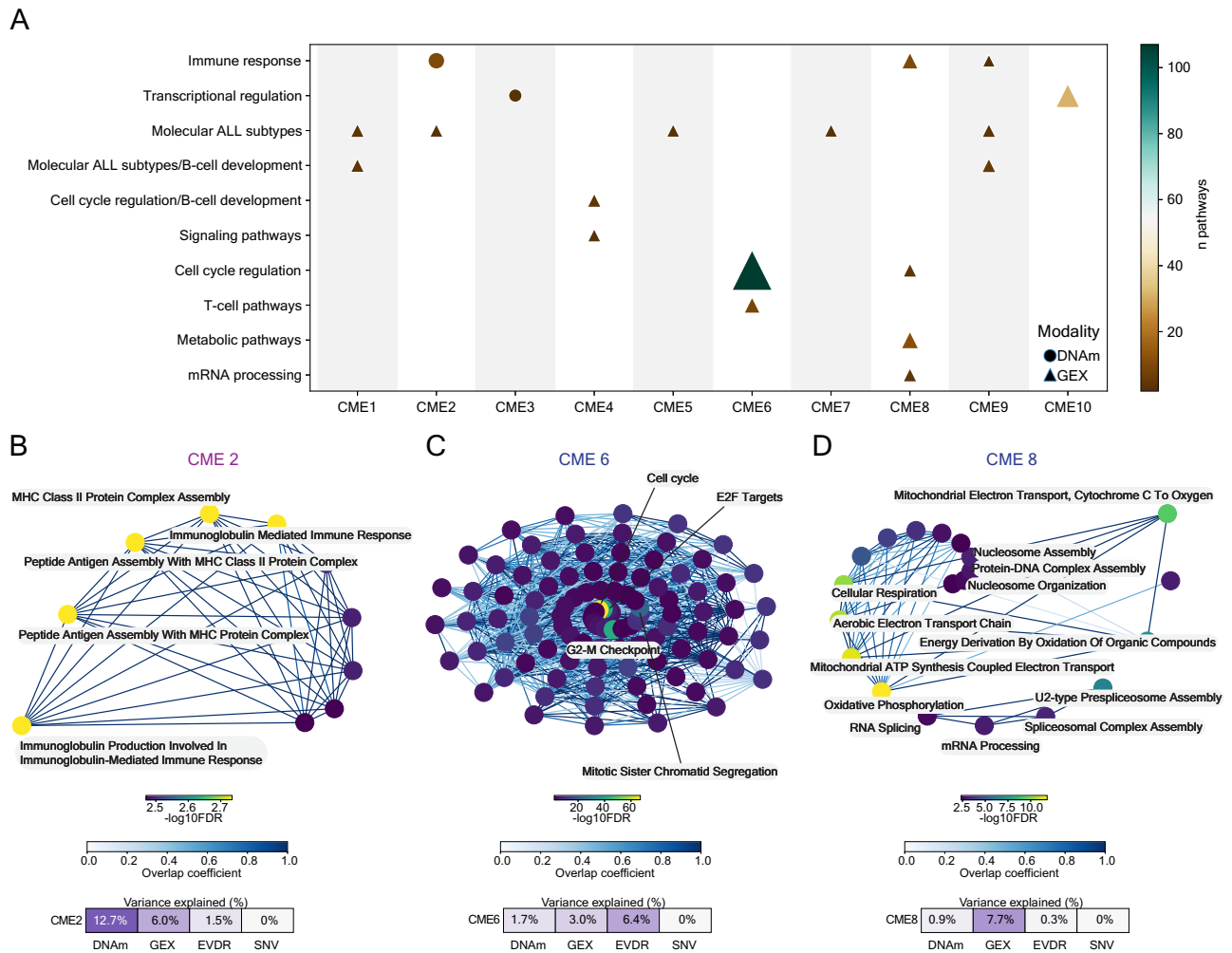


Fig. 3 | Biological pathways associated with CMEs. **A** Significantly enriched pathways (FDR < 0.01, y-axis) in features with negative or positive weights by CME (x-axis). The pathway clusters enriched in DNAm are denoted with circles and GEX with triangles. **B** Network of the immune response cluster for CpG sites with positive weights in CME 2. **C** Network of the cell cycle cluster for genes with negative weights

in CME 6. **D** Network of the combined cell cycle, metabolic pathways, mRNA processing clusters for genes with negative weights for CME 8. The pathway nodes are color-coded based on $-\log_{10}$ FDR score, and edge connectivity is color-coded based on the overlap coefficient.

(82.1%, $n = 184$), *ETV6::RUNX1*-like (66.7%, $n = 10$), high hyperdiploidy (HeH, 69.2%, $n = 166$), and low HeH (100%, $n = 6$). In comparison, subtypes associated with poorer outcomes, such as *KMT2A-r* (100%, $n = 39$), *BCR::ABL1* (77.8%, $n = 14$), and *BCR::ABL1*-like (85.7%, $n = 18$), predominantly clustered within the high-methylation group. Surprisingly, the high-methylation cluster, associated with higher SI% after doxorubicin exposure, also contained about one third of HeH patients (30.8%, $n = 74$). For these patients, all CpG sites defining the hub ($n = 17$) were differentially methylated between the low- and high-methylation clusters, with 11 sites demonstrating > 20% methylation differences (Fig. 4E). Kaplan-Meier analysis confirmed a significant difference in relapse-free survival (RFS) between these HeH subclusters (log-rank p -value = 0.041, Fig. 4F). Specifically, HeH patients in the high-methylation doxorubicin-associated cluster exhibited both higher SI% and inferior survival outcome, an effect that remained significant after adjustment for risk group and treatment protocol (Wald test p -value = 0.039, HR = 2.34, 95% CI 1.04–5.2, Fig. 4F). Importantly, these patients did not differ in baseline characteristics such as age, white blood cell count (Mann–Whitney U test p -value > 0.05), or MRD response at day 29 ($n = 78$, Mann–Whitney U test p -value > 0.05).

While HeH is typically associated with a favorable prognosis, our findings align with previous reports describing a subset of high-risk HeH patients with poor outcomes^{57,58}. Using the karyotype-based risk stratification proposed by Enshaei et al.⁵⁹, we classified HeH patients into good-risk

(+17 and +18, or +17 or +18 without +5 or +20) and poor-risk (all other karyotype patterns). We did not observe a significant difference in the proportion of good- and poor-risk patients between the low- and high-doxorubicin response clusters (Fisher’s exact test, p -value = 0.22). Furthermore, differential gene expression analysis comparing the high- ($n = 8$) and low-doxorubicin ($n = 24$) response HeH subclusters identified three differentially expressed genes, *CLIP4* (\log_2 Fold Change (\log_2 FC) 1.14, FDR 0.035), *ABCB1* (\log_2 FC 1.2, FDR 0.049), and *CYBB* (\log_2 FC -2.64 , FDR 0.043).

In a second example, in CME 8, low RNA expression of six histone genes (*H3C2*, *H2AC12*, *H2AC17*, *H2BC17*, *H2AC14* and *H2BC13*) was associated with higher SI% after exposure to vincristine (Fig. S5). Stratifying patients into three groups based on the expression levels of these genes revealed significant differences in SI% after ex vivo vincristine exposure, with the low-expression group showing higher SI% (Dunn’s test, BH-adjusted p -value < 0.01, Fig. S5). In contrast, in CME 10, doxorubicin response was central and was associated with the expression of *FOSL1*, *FOSB*, *JUNB*, *TRIB1*, *LMNA*, and *ZFP36* (Fig. S5, Supplementary Data S10). Stratification of patients into three clusters based on the expression of these genes revealed significantly lower SI% after ex vivo doxorubicin exposure in the low-expression group (Dunn’s test BH p -value < 0.001, Fig. S5).

Finally, we validated the inter-modal interactions using the test dataset ($n = 308$ patients), achieving 28.6–100% concordance for CME-specific

Table 1 | Hub features per CME for positive and negative weights

CMEs	Positive CMEs Hubs	Negative CMEs Hubs
CME1	cg00593243(<i>DUSP1</i>); <i>ITPRIPL2</i> ; cg13280914; <i>CAMK1</i> ; <i>RBM47</i> ; cg01578875; <i>IL6R</i>	<i>CXXC5</i> ; <i>KCNQ5</i> ; <i>BEST3</i> ; <i>CCNJL</i> ; <i>RHOBTB1</i>
CME2	Etoposide; Doxorubicin	<i>ARHGEF4</i> ; <i>HAP1</i> ; <i>EPHA7</i> ; cg03717315(<i>SEMA4F</i>); <i>KCNN1</i> ; cg16206460(<i>PGLYRP2</i> ; <i>RASAL3</i>); <i>BIRC7</i> ; <i>PCLO</i> ; <i>LOXHD1</i> ; cg26218983(<i>ZC3H12C</i>); cg07372034; <i>IGF2BP1</i> ; cg16016176(<i>ADPRH</i>); <i>CLIC5</i> ; <i>DSC3</i> ; cg02851793(<i>SRGN</i>)
CME3	Asparaginase; Dexamethasone	<i>FBXO41</i> ; <i>FLT3</i>
CME4	Dexamethasone; <i>ACSL1</i> ; cg12150931(<i>ZNF385A</i>); Vincristine; <i>GABARAPL1</i> ; <i>BHLHE40</i> ; cg18425731(<i>ERMN</i>)	<i>VPREB1</i> ; cg27392771; cg20752878(<i>ASGR1</i>); <i>CYTL1</i> ; cg02225720(<i>HASPIN</i>); <i>RAG1</i> ; cg20559385; cg06958535(<i>LAX1</i>)
CME5	cg12150931(<i>ZNF385A</i>); <i>ATP9A</i> ; cg12761788(<i>ENDOU</i>); <i>SALL4</i> ; <i>DAPK1</i> ; <i>GATA3</i> ; <i>TTC28</i> ; <i>CXCL2</i>	<i>LRRC14B</i> ; cg27022853; cg05569131(<i>RAB44</i>); <i>CAMK2D</i> ; <i>PHLDB2</i> ; <i>TTN</i> ; <i>LEF1</i>
CME6	cg06452129; <i>GPRIN3</i> ; cg16997486; <i>GIMAP5</i> ; <i>TRAT1</i> ; <i>TC2N</i> ; cg16977751; <i>LTB</i> ; <i>GIMAP7</i> ; cg24394336	cg08097359
CME7	cg17183174(<i>COMMD6</i>); cg15354065(<i>DGKA</i>); <i>LDLRAD3</i> ; <i>GLDC</i> ; cg00524374	<i>AHR</i> ; <i>ANTXR2</i> ; cg14429979; cg00915974; cg11952493(<i>MPP7</i>); cg16783349; <i>SAMSN1</i> ; cg08015762
CME8	<i>CIITA</i>	cg07311994; Vincristine; <i>LGALS1</i> ; <i>RGS2</i> ; <i>GPR183</i> ; cg04450037
CME9	cg24618492(<i>EPN2</i>); cg25789861(<i>DSC3</i>); <i>MAPKBP1</i> ; cg09144073(<i>STXBP5</i>); cg02312409(<i>RNF217</i>); cg19478500(<i>RNF217</i>); cg19832521(<i>NOVA1</i>); <i>CD9</i> ; cg03017520(<i>DSC3</i>); <i>GALNT2</i> ; cg13434989(<i>EDNRB</i>); <i>NIBAN3</i> ; <i>PKIG</i>	cg16747164; cg05754179; cg08478016; <i>KCNN1</i> ; <i>DSC2</i> ; cg01984854; <i>BASP1</i> ; <i>IGF2BP1</i> ; <i>PLCB4</i> ; <i>DSC3</i> ; <i>PRKAR2B</i>
CME10	cg17604985	Amsacrine; Doxorubicin

A hub is defined as a feature with > 3 connections to features that belong to another data modality.

hubs (Supplementary Data S12, S13). Specifically, we confirmed associations within CME 2 (etoposide, 42/70 CpG sites; doxorubicin, 5/17 CpG sites), CME 8 (vincristine, *H3C2* and *H2BC13*), and CME 10 (doxorubicin, *FOSB*, *JUNB* and *ZFP36*).

Functional annotation of CMEs

Next, we integrated our key findings and summarized the interactions, mechanisms, and pathways associated with each CME (Fig. 5). This map provides a unified and comprehensive view of CME-driven associations. Key annotations related to BCP-ALL subtypes or B-cell development were linked to several CMEs (1, 2, 4, 5, 7, and 9), underscoring their central roles in the disease. Immune-related responses were also prominent across CMEs 2, 6, 8, and 9, encompassing diverse mechanisms such as Major Histocompatibility Complex (MHC) class responses (CME 2), T-cell pathways (CME 6), humoral responses (CME 8), and interleukin-mediated responses (CME 9). Ex vivo drug responses were mapped to several CMEs (2, 3, 4, 6, 8, and 10). Notably, CME 4 captured variability in glucocorticoid (dexamethasone and prednisolone), thioguanine, and vincristine responses across GEX and DNAm data (Fig. S3). Cell cycle regulation emerged as a central driver across CMEs 4, 6, and 8. CME 8, in particular, was characterized by regulation of metabolic pathways and histone-mediated vincristine response.

Distinct processes were identified for specific CMEs, such as CME 10, which was uniquely associated with cell proliferation and apoptosis. This CME featured AP-1 complex genes (*FOSL1*, *FOSB*, and *JUNB*), which were positively linked to response to topoisomerase inhibitors, such as doxorubicin and amsacrine.

CMEs reveal the prognostic value of ex vivo drug responses

To evaluate the prognostic utility of CMEs beyond baseline clinical risk groups, we implemented gradient boosted survival models and evaluated their performance using the CV c-index⁵⁴ across the 3 × 5 repeated stratified folds, as well as the c-index on the independent test dataset. For models incorporating molecular and drug response data, we applied the MOFA imputation step to generate complete feature sets on the complete training set, and the training was performed using the top features (absolute weight > 0.6).

Pairwise CV score comparisons between the clinical baseline model and 82 CME-based models identified three models with significantly

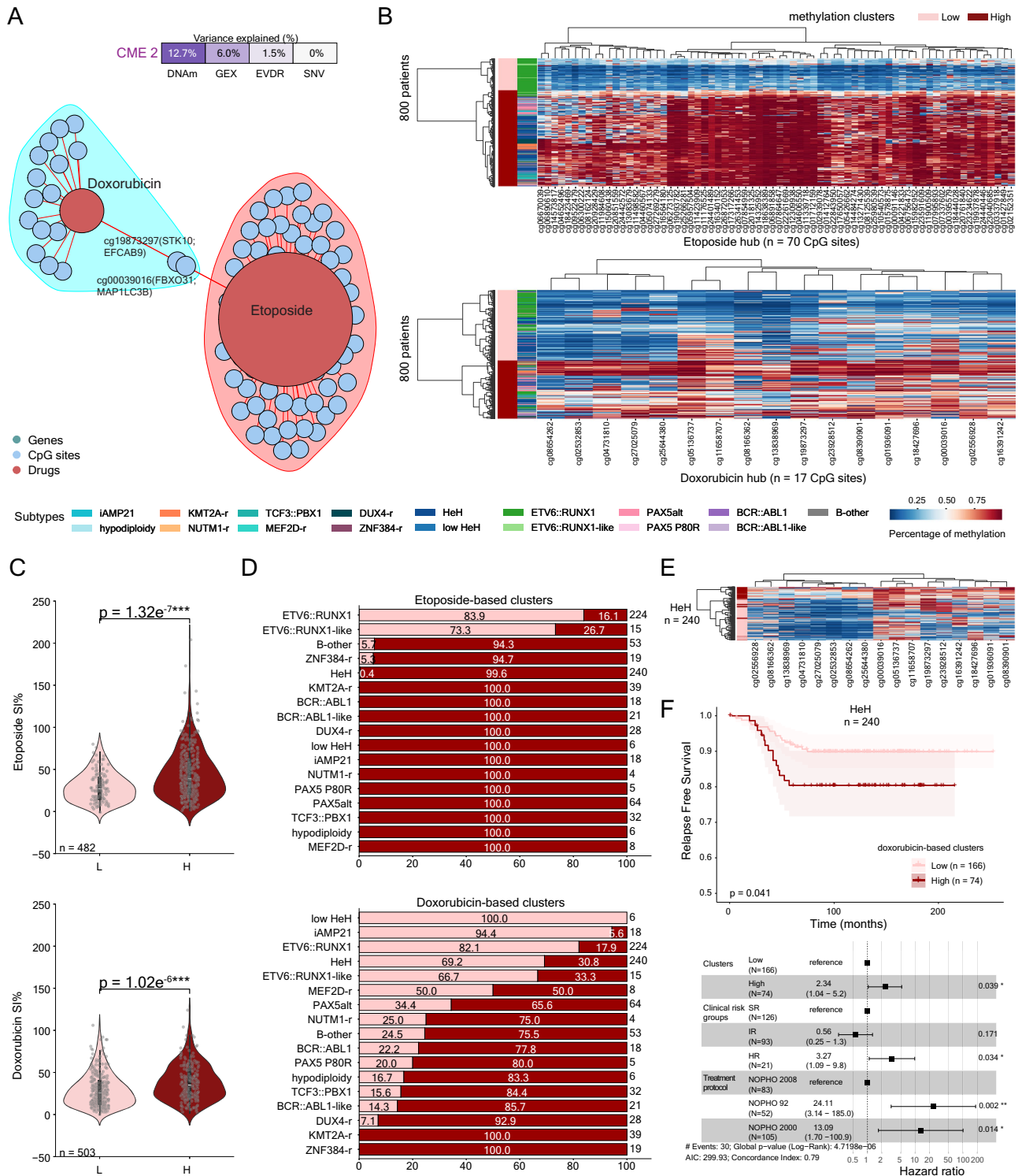
improved performance (FDR < 0.05, Supplementary Data S14, Fig. 6A). In all three cases, combining CME features (i.e., ex vivo drug responses) with clinical data improved prognostic accuracy. Specifically, the top performing models derived from CME 2 (FDR 0.01, mean CV c-index 0.64, 95% CI 0.62–0.66), CME 7 (FDR 0.02, mean CV c-index 0.64, 95% CI 0.62–0.66), and CME 8 (FDR 0.005, CV c-index 0.62, 95% CI 0.60–0.65), all outperforming the baseline clinical model (mean CV c-index 0.59, 95% CI 0.56–0.61) and generalized well to the test dataset (Fig. 6B, Supplementary Data S14). Specifically, these models achieved higher c-index (CME 2: 0.663, CME 7: 0.667, and CME 8: 0.714) compared to the clinical model alone (0.656) on the test dataset.

Top drugs contributing to CME 2 were cytarabine, etoposide, and doxorubicin. Notably, in the previous section, we linked ex vivo response to doxorubicin to the DNA methylation levels of 17 CpG sites, which further stratified patients by relapse-free survival (Fig. 4). CME 7 is one of the intermodal CMEs, spanning across all data modalities, yet the top features of this CME were ex vivo response to dexamethasone, prednisolone, doxorubicin, etoposide, thioguanine, and vincristine, explaining most of the variance in the EVDR data (5.4%). Finally, CME 8 was primarily driven by vincristine, L-asparaginase, and cytarabine. Of particular interest, vincristine SI% was negatively associated with the expression levels of six histone genes (Fig. 4).

Discussion

Risk stratification for pediatric BCP-ALL remains imperfect, reflecting both the biological diversity of the disease and our incomplete understanding of its molecular drivers. To address this challenge, we integrated DNA methylation, transcriptional, and drug response data from 1,231 patients to construct an integrated molecular map of pediatric BCP-ALL. This analysis identified ten distinct CMEs that captured key aspects of leukemia biology and functional drug responses.

Although our integrative CMEf and CMEf-all models did not generally outperform the clinical baseline, our framework proved effective for uncovering informative feature combinations. Our data demonstrated that survival models incorporating CME-derived drug response features (CMEs 2, 7, and 8) yielded added prognostic value. In this context, integrative multi-omics analyses primarily supported feature prioritization rather than directly enhancing the predictive performance of survival models.



From a pathophysiological perspective, our data highlighted that CMEs 1 and 2 were significantly correlated with the molecular subtype. This is not unexpected, as gene expression and DNA methylation biomarkers are driven by the known, subtype-specific heterogeneity of BCP-ALL^{1,2,10,60}. However, in the inter-modal network analysis for CME 2, we noted an unexpectedly high proportion of patients with low-risk molecular subtypes (e.g., HeH) stratified, using DNA-methylation data, in a subgroup of patients with higher SI% when exposed to topoisomerase inhibitors (i.e., doxorubicin). These patients exhibited significantly inferior outcomes, independent of other clinical variables. Differential

gene expression analysis between the two HeH subgroups revealed upregulation of *CLIP4* and *ABCB1* and downregulation of *CYBB* in patients with a hypermethylation pattern and high doxorubicin SI%. *ABCB1*, in particular, which encodes the efflux pump P-glycoprotein, is a key drug-efflux transporter and well-known driver of anthracycline resistance⁶¹, suggesting a drug-refractory cellular state in this group. These results further support the presence of molecular subgroups within the HeH subtype that exhibit differential clinical outcomes, consistent with previous studies^{57-59,62}. Importantly, this study demonstrates that integrative multiomics analysis can uncover previously overlooked

Fig. 4 | Inter-modal networks and correlations in the train dataset (n = 923 samples). **A** Inter-modal networks for features with positive weights and Pearson's absolute correlation coefficient (ρ) > 0.2 in CME 2. The variance explained by CME 2 is shown at the top of the panel. Positive correlations are denoted with red. The size of each correlation (edge) is based on the correlation value, while the size of each circle (vertex) represents the number of connections + a weight of 10. **B** Heatmaps of the DNA methylation levels for CpG sites (x-axis) across 800 patient samples (y-axis) mapped to the etoposide (top) and doxorubicin hubs (bottom) ordered by unsupervised hierarchical clustering. The methylation clusters and the molecular subtypes are shown as annotation bars on the y-axis. **C** Violin plots of the distribution of surviving cells (SI%, y-axis) after etoposide (top) and doxorubicin (bottom) treatment, colored by DNA methylation cluster group (low-high, x-axis). Two-sided Mann-Whitney U test *p*-values: *** < 0.001, ** < 0.01, * < 0.05, ns: non-significant. L: low, H: high. **D** Molecular subtype distribution in the low and high

methylation clusters across 800 patients with DNAm data available on the training dataset in the etoposide hub-based clusters (top) and in the doxorubicin hub-based clusters (bottom). **E** Heatmap of the DNA methylation levels for CpG sites (x-axis) across 240 patient samples with high hyperdiploid (HeH) subtype (y-axis) ordered by unsupervised hierarchical clustering and color-coded by the low- and high-doxorubicin-based methylation clusters. **F** Kaplan-Meier survival curves for the doxorubicin hub-based clusters, illustrating relapse-free survival (RFS, y-axis) over time (x-axis) for HeH patients (n = 240) with available DNAm data (top). The log-rank test *p*-value access the difference between the two clusters and is denoted on the bottom left corner. The error bands define the upper and lower 95% confidence intervals. Cox proportional hazards regression adjusted for treatment protocol and clinical risk group (bottom). The forest plot shows the hazard ratio with the corresponding 95% confidence intervals for relapse in the doxorubicin hub-based clusters.

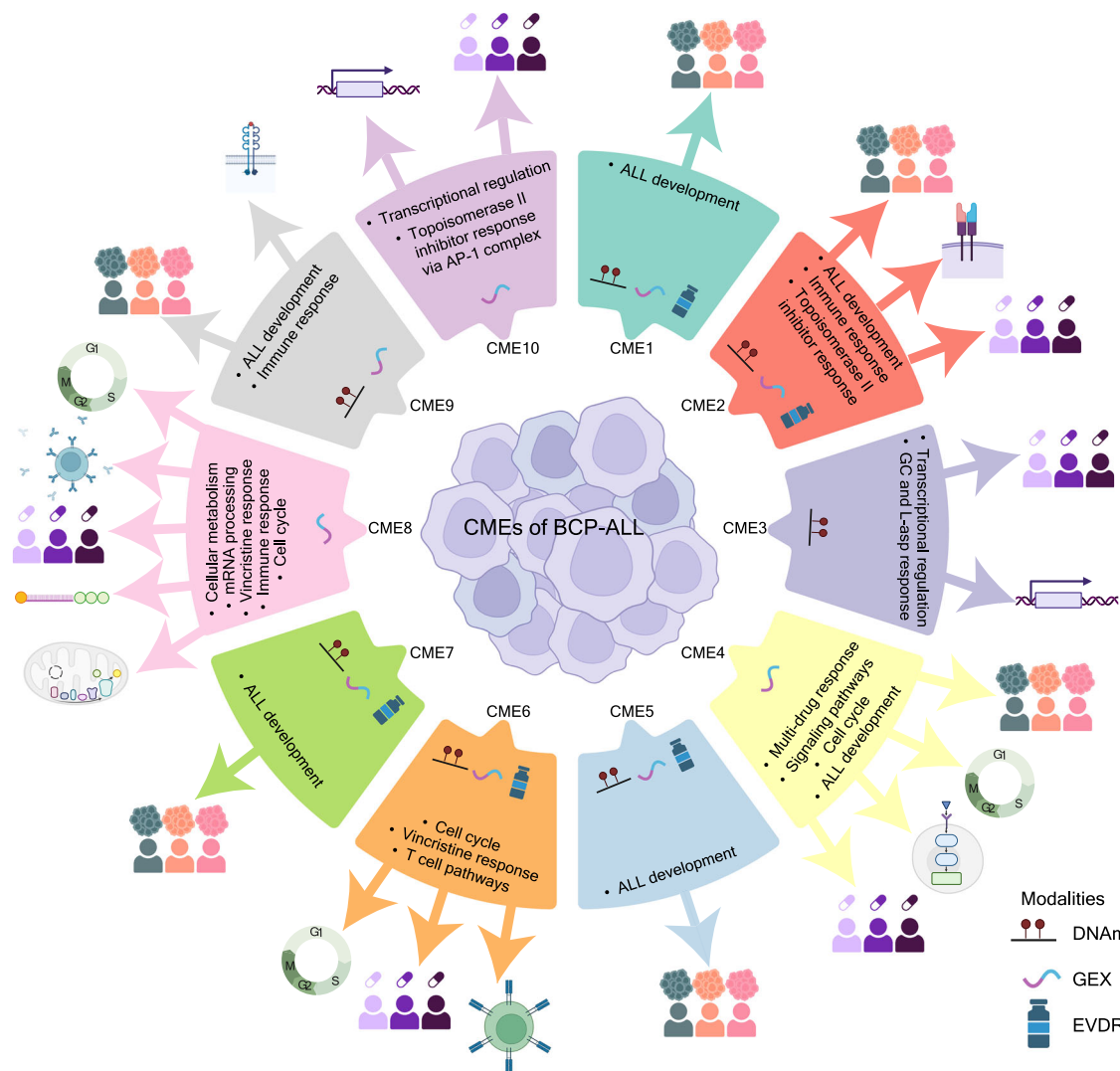


Fig. 5 | Graphical functional annotation map across CMEs. The data modalities with >1% variance explained by MOFA are highlighted in the center of each CME. CG: glucocorticoids, L-asp: L-asparaginase, DNAm: DNA methylation, GEX: Gene

expression, EVDR: ex vivo drug response. The figure was created in Adobe Illustrator, but the icons were adapted from Biorender.com.

molecular interactions influencing doxorubicin response. In line with our observations, Lee et al.³⁰ reported variable responses to anthracyclines among HeH patients, where daunorubicin-resistant cases demonstrated sensitivity to other compounds such as trametinib, venetoclax, and ibrutinib, suggesting potential alternative treatment options.

Furthermore, in the inter-modal network analysis of CME 2, of the 70 hub-specific CpG sites in the etoposide hub, only ten overlapped with the ALLIUM³⁷ ALL subtype differentiating CpGs, while none were found on the doxorubicin hub. This finding suggests that drug responses may be influenced by other factors beyond the molecular subtypes, including

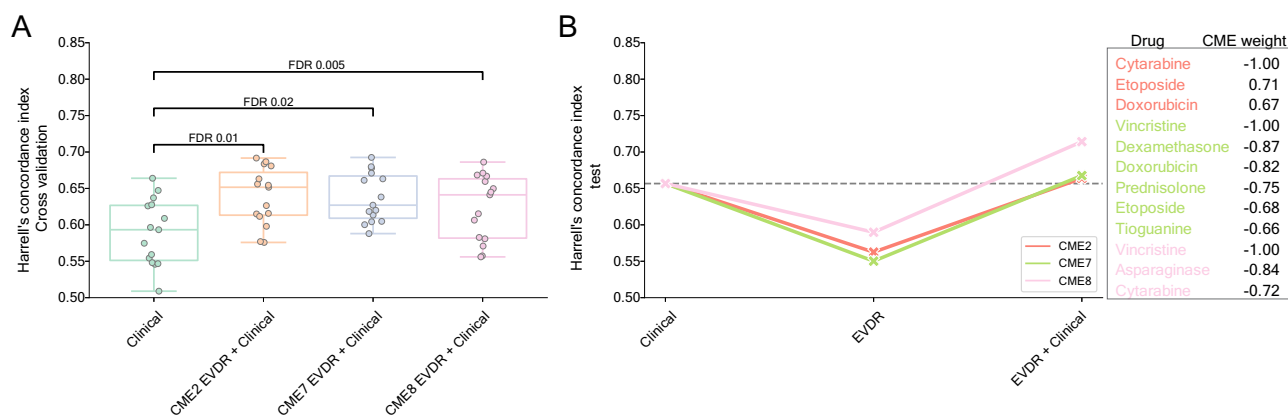


Fig. 6 | Survival model performance evaluation. **A** Harrell's concordance index (c-index, y-axis) for each model (x-axis) across individual CMEs on the 3×5 repeated stratified cross-validated sets. The whiskers demonstrate the distribution of the data points beyond the lower and upper quartiles. Two-sided Wilcoxon signed-rank test Benjamini-Hochberg (BH) adjusted *p*-values are shown above each boxplot.

B c-index (y-axis) for the clinical baseline model, the CME-derived EVDR models, and combined models (x-axis) on the test dataset (*n* = 308). The grey dotted line represents the c-index (0.657) for the clinical baseline model. The top drugs color-coded by CME and their impact (weights) on the respective CMEs are provided as a figure legend.

epigenetic modification, given the exclusive connection of the hubs to CpG sites.

Cell cycle regulation plays a dual role in chemotherapy, either acting synergistically with or counteracting drug efficacy⁶³. The drugs used in this study exhibit diverse mechanisms of action, with some targeting all cell cycle stages (cell cycle non-specific agents) and others affecting only actively dividing cells (cell cycle-specific agents). For instance, alkylating agents, anthracyclines, and topoisomerase inhibitors like doxorubicin, amsacrine, and etoposide are non-specific agents, effective against cells irrespective of their cell cycle phase⁶³. In contrast, antimetabolites (e.g., cytarabine, thioguanine) and vinca alkaloids (e.g., vincristine) are cell cycle-specific, limiting their effects to actively dividing cells^{63,64}. Here, cell cycle regulation emerged as a main feature of CMEs 6 and 8, where cell cycle-dependent drugs served as central hubs. Notably, in CME 8, higher expression of genes encoding histone proteins (markers of proliferating cells) was associated with lower SI % following vincristine exposure. Our findings underscore the role of core histone protein expression in mediating drug sensitivity in ALL. These results align with prior studies showing that slow or non-cycling cancer cells are associated with therapy resistance⁶⁵. For example, ALL samples with higher cell proliferation rates showed distinct sensitivity profiles to cell cycle-related drugs in ex vivo xenografts models⁶⁶ and cell cycle checkpoint regulation is critical in ALL treatment response⁶⁷. Potential strategies include combining vincristine with agents active in quiescent cells, such as BCL2 inhibitors (venetoclax)⁶⁸. CME 6 and 8, along with their combined survival models, did not outperform the clinical model, suggesting that while cell-cycle-specific features are biologically relevant, they may not capture the full variability required to enhance model performance. However, the EVDR-combined survival model for CME 8 outperformed the baseline model, suggesting a link between ex vivo drug responses and clinical outcomes.

Alterations in cellular metabolism, oxidative phosphorylation, and cellular respiration pathways are established mechanisms of chemoresistance and potential therapeutic targets in acute leukemia⁶⁹. CME 8 was enriched to metabolic and cell cycle-related pathways, showcasing evidence of the relationship between metabolism and cell cycle, as previously described⁷⁰. Deciphering this bidirectional relationship may allow the use of drugs targeting metabolic pathways, which can in turn affect cell cycle progression^{70,71}.

The top-ranked genes in CME 10 were enriched in transcriptional regulation pathways, including TNF-alpha signaling via NF-kB, p53, apoptosis, and hypoxia. Notably, the expression levels of *FOSL1*, *FOSB*, *JUNB*, *TRIB1*, *LMNA* and *ZFP36* were positively correlated with response to doxorubicin. These genes are components or regulators of the Activator Protein 1 (AP-1) transcriptional complex⁷²⁻⁷⁶. AP-1 is known to play a

central role in transcriptional regulation, cell proliferation, apoptosis, and metastasis. Our findings suggest that patients with high AP-1 gene expression may exhibit reduced suitability for treatment with topoisomerase inhibitors like doxorubicin. Although, topoisomerase inhibitors and glucocorticoids are mechanistically distinct, previous studies showed that JUN knockdown mediates resistance to glucocorticoids in T-ALL⁷³, highlighting the broader role of AP-1 components as potential therapeutic targets in ALL.

Multi-Omics integration approaches, such as MOFA²², are powerful tools for vertically integrating and reducing the dimensionality of complex multimodal datasets. They employ various strategies, including probabilistic, multivariate, network-based, similarity-based, fusion, or correlation-based methods³³. The number of available tools continues to grow, yet systematic benchmarking remains limited. Although benchmarking MOFA against other integration tools was beyond the scope of this study, our choice was guided by the need for an unsupervised method capable of handling partially overlapping data, while remaining computationally efficient and interpretable. Despite the limitations of using a linear approach such as MOFA, interpreting the results within a biological context remains essential.

MOFA provided a holistic exploration of multimodal datasets, while also identifying individual modalities associated with specific biological aspects. In addition, as the authors of MOFA demonstrated by masking the data, the method effectively handles missing data (within or between assays) through imputation²². This mitigated the risk of reduced sample size due to modality imbalance and maximized the statistical power of our survival analyses. By incorporating all available patient samples (923 vs. 147 in the train set and 308 instead of 49 in the test set), we were able to build robust survival models, a crucial step for developing reliable, predictive tools, which can generalize effectively across diverse cohorts. However, when the proportion of missing samples is high (e.g., GEX-76%), the imputation step may not be effective for this modality. Therefore, we recommend integrating all modalities to mitigate the effect of missing data. One limitation of MOFA is the lack of a method for projecting CMEs onto external datasets. To address this, we used MOFA as a proxy to identify the most influential features (genes, CpG sites, and drugs) for downstream analyses, which were validated on the test data. In the context of other diseases, Iperi et al.⁷⁷ and Pekayvaz et al.⁷⁸, and for healthy blood profiling, de Visser et al.⁷⁹ successfully applied MOFA and validated their findings in other datasets, focusing on either multiple or single data modalities. Finally, given the associations between drugs and GEX profiles in CMEs 8 and 10, exploring these relationships at the proteomic level may yield additional insights given the significant role of the proteome in reflecting the impact of mutations in ALL¹³ and its potential as a source of druggable proteins¹¹.

As our study is based on a retrospective cohort, the long follow-up time is a major strength. However, treatment protocols have evolved over time. End-of-induction (EOI) minimal residual disease (MRD) assessment, now routinely used for risk stratification, was only available for 348 of the patients in our dataset. The limited number of patients stratified by MRD data prevented its broader inclusion in our analyses, as selective incorporation could introduce bias. However, we did not observe any association between CMEs and the main treatment protocols (NOPHO ALL-92, –2000, or –2008) except for infants under one year of age treated according to the Interfant protocol. This distinction aligns with the unique molecular and age-based stratification criteria of Interfant³⁵, which were expected to be reflected in the CME data. The patients included in this study also preceded the implementation of copy number alteration-based stratification⁸⁰. While our findings remain relevant for understanding leukemia biology and treatment responses, future studies should strive to fully integrate MRD and copy number analyses to align with contemporary risk stratification strategies.

Conclusion

Our multiomics integration study revealed how individual and combined data modalities contribute to key biological processes and clinical outcomes in pediatric BCP-ALL. This study underscores the potential of multiomics profiling to improve risk estimation and disease stratification in BCP-ALL. By reducing the complexity of large datasets to a few clinically significant factors, we identified overarching molecular signatures with prognostic relevance. These findings highlight the value of integrative, factor-based approaches for advancing clinical, basic, and translational research in BCP-ALL.

Data availability

The GEX count matrix was retrieved from GSE227832. DNA methylation data were available from their original studies, under controlled access via GSE49031, <https://doi.org/10.17044/scilifelab.22303531>⁸¹ and <https://doi.org/10.17044/scilifelab.26096371>⁸² (<https://figshare.scilifelab.se/>). The processed EVDR data are available in Supplementary Data S15. Source data underlying the analyses in the main Figures are available in Supplementary Data S16. The raw FASTQ and IDAT files are not shared publicly due to confidentiality and ethical restrictions. Any data inquiries may be submitted to the corresponding author and will be reviewed within four weeks. Access may be granted to qualified researchers subject to a data use agreement restricting use of the data to approved research purposes, prohibiting attempts to re-identify participants, and requiring compliance with applicable ethical and data protection regulations.

Code availability

All R and Python scripts, and the environment requirements to reproduce the analyses, are openly available at our GitHub repository https://github.com/Molmed/Krali_2026_MultiOmics⁸³. All figures can be reproduced using the provided Supplementary Data files and the workflow scripts.

Received: 16 April 2025; Accepted: 18 March 2026;
Published online: 11 April 2026

References

1. Walter, W., Iacobucci, I. & Meggendorfer, M. Diagnosis of acute lymphoblastic leukaemia: an overview of the current genomic classification, diagnostic approaches, and future directions. *Histopathol. His.* **86**, 134–145 (2024).
2. 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).
3. Alaggio, R. et al. The 5th edition of the World Health Organization Classification of Haematolymphoid Tumours: Lymphoid Neoplasms. *Leukemia* **36**, 1720–1748 (2022).

4. Duffield, A. S., Mullighan, C. G. & Borowitz, M. J. International Consensus Classification of acute lymphoblastic leukemia/lymphoma. *Virchows Arch.* **482**, 11–26 (2023).
5. Nordlund, J. et al. Genome-wide signatures of differential DNA methylation in pediatric acute lymphoblastic leukemia. *Genome Biol.* **14**, r105 (2013).
6. 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).
7. Turati, V. A. et al. Chemotherapy induces canalization of cell state in childhood B-cell precursor acute lymphoblastic leukemia. *Nat. Cancer* **2**, 835–852 (2021).
8. 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).
9. Lilljebjörn, H. & Fioretos, T. New oncogenic subtypes in pediatric B-cell precursor acute lymphoblastic leukemia. *Blood* **130**, 1395–1401 (2017).
10. Gu, Z. et al. PAX5-driven subtypes of B-progenitor acute lymphoblastic leukemia. *Nat. Genet.* **51**, 296–307 (2019).
11. Leo, I. R. et al. Integrative multi-omics and drug response profiling of childhood acute lymphoblastic leukemia cell lines. *Nat. Commun.* **13**, 1691 (2022).
12. Yang, M. et al. Proteogenomics and Hi-C reveal transcriptional dysregulation in high hyperdiploid childhood acute lymphoblastic leukemia. *Nat. Commun.* **10**, 1519 (2019).
13. Lorentzian, A. C. et al. Targetable lesions and proteomes predict therapy sensitivity through disease evolution in pediatric acute lymphoblastic leukemia. *Nat. Commun.* **14**, 7161 (2023).
14. Fu, J. et al. Metabolic profiling reveals metabolic features of consolidation therapy in pediatric acute lymphoblastic leukemia. *Cancer Metab.* **11**, 2 (2023).
15. Iacobucci, I. & Mullighan, C. G. Genetic Basis of Acute Lymphoblastic Leukemia. *J. Clin. Oncol.* **35**, 975–983 (2017).
16. Toft, N. et al. Results of NOPHO ALL2008 treatment for patients aged 1–45 years with acute lymphoblastic leukemia. *Leukemia* **32**, 606–615 (2018).
17. Oskarsson, T. et al. Relapsed childhood acute lymphoblastic leukemia in the Nordic countries: prognostic factors, treatment and outcome. *Haematologica* **101**, 68–76 (2016).
18. Rheingold, S. R. et al. Determinants of survival after first relapse of acute lymphoblastic leukemia: a Children’s Oncology Group study. *Leukemia* **38**, 2382–2394 (2024).
19. Zhu, D. et al. The interaction between DNA methylation and tumor immune microenvironment: from the laboratory to clinical applications. *Clin. Epigenet.* **16**, 24 (2024).
20. Hanahan, D. Hallmarks of Cancer: New Dimensions. *Cancer Discov.* **12**, 31–46 (2022).
21. Kimura, S. & Mullighan, C. G. Molecular markers in ALL: Clinical implications. *Best. Pract. Res. Clin. Haematol.* **33**, 101193 (2020).
22. Argelaguet, R. et al. Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. *Mol. Syst. Biol.* **14**, e8124 (2018).
23. 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).
24. Isobe, T. et al. Multi-omics analysis defines highly refractory RAS burdened immature subgroup of infant acute lymphoblastic leukemia. *Nat. Commun.* **13**, 4501 (2022).
25. Esplin, E. D. et al. Multiomic analysis of familial adenomatous polyposis reveals molecular pathways associated with early tumorigenesis. *Nat. Cancer* **5**, 1737–1753 (2024).
26. Gerlevik, S. et al. Identification of novel myelodysplastic syndromes prognostic subgroups by integration of inflammation, cell-type

- composition, and immune signatures in the bone marrow. *eLife* **13**, RP97096 (2024).
27. Yan, Z. et al. Multi-omics integration reveals potential stage-specific druggable targets in T-cell acute lymphoblastic leukemia. *Genes Dis.* **11**, 100949 (2024).
 28. Bottomly, D. et al. Integrative analysis of drug response and clinical outcome in acute myeloid leukemia. *Cancer Cell* **40**, 850–864.e9 (2022).
 29. 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).
 30. 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).
 31. Malani, D. et al. Implementing a Functional Precision Medicine Tumor Board for Acute Myeloid Leukemia. *Cancer Discov.* **12**, 388–401 (2022).
 32. 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).
 33. Subramanian, I., Verma, S., Kumar, S., Jere, A. & Anamika, K. MultiomicS Data Integration, Interpretation, And Its Application. *Bioinforma. Biol. Insights* **14**, 117793221989905 (2020).
 34. 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).
 35. 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. *Lancet* **370**, 240–250 (2007).
 36. 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).
 37. Krali, O. et al. Multimodal classification of molecular subtypes in pediatric acute lymphoblastic leukemia. *Npj Precis. Oncol.* **7**, 131 (2023).
 38. Enblad, A. P. et al. Ex vivo drug responses and molecular profiles of 597 pediatric acute lymphoblastic leukemia patients. *HemaSphere* **9**, e70176 (2025).
 39. Daenekas, B. et al. Conumee 2.0: enhanced copy-number variation analysis from DNA methylation arrays for humans and mice. *Bioinformatics* **40**, btac029 (2024).
 40. Ritchie, M. E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47–e47 (2015).
 41. Lindqvist, C. M. et al. The mutational landscape in pediatric acute lymphoblastic leukemia deciphered by whole genome sequencing. *Hum. Mutat.* **36**, 118–128 (2015).
 42. 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).
 43. 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).
 44. Sayyab, S. et al. Mutational patterns and clonal evolution from diagnosis to relapse in pediatric acute lymphoblastic leukemia. *Sci. Rep.* **11**, 15988 (2021).
 45. Lindhagen, E., Nygren, P. & Larsson, R. The fluorometric microculture cytotoxicity assay. *Nat. Protoc.* **3**, 1364–1369 (2008).
 46. McLean, C. Y. et al. GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.* **28**, 495–501 (2010).
 47. Fang, Z., Liu, X. & Peltz, G. GSEAPy: a comprehensive package for performing gene set enrichment analysis in Python. *Bioinformatics* **39**, btac757 (2023).
 48. Schmidt, B. et al. ALLSorts: an RNA-Seq subtype classifier for B-cell acute lymphoblastic leukemia. *Blood Adv.* **6**, 4093–4097 (2022).
 49. 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).
 50. Tirosh, I. et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **352**, 189–196 (2016).
 51. Iacobucci, I. et al. Multipotent lineage potential in B cell acute lymphoblastic leukemia is associated with distinct cellular origins and clinical features. *Nat. Cancer* **6**, 1242–1262 (2025).
 52. Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
 53. Csárdi, G. et al. igraph for R: R interface of the igraph library for graph theory and network analysis. Zenodo <https://doi.org/10.5281/ZENODO.7682609> (2024).
 54. 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).
 55. For The St. Jude Children's Research Hospital–Washington University Pediatric Cancer Genome Project et al. The landscape of somatic mutations in infant MLL-rearranged acute lymphoblastic leukemias. *Nat. Genet.* **47**, 330–337 (2015).
 56. Inaba, H. & Mullighan, C. G. Pediatric acute lymphoblastic leukemia. *Haematologica* **105**, 2524–2539 (2020).
 57. Purvis, K. et al. Outcomes in patients with ETV6::RUNX1 or high-hyperdiploid B-ALL treated in the St. Jude Total Therapy XV/XVI studies. *Blood* **145**, 190–201 (2025).
 58. Harrison, C. J. Lower-intensity therapy for good-risk B-ALL. *Blood* **145**, 144–146 (2025).
 59. Enshaei, A., Vora, A., Harrison, C. J., Moppett, J. & Moorman, A. V. Defining low-risk high hyperdiploidy in patients with paediatric acute lymphoblastic leukaemia: a retrospective analysis of data from the UKALL97/99 and UKALL2003 clinical trials. *Lancet Haematol.* **8**, e828–e839 (2021).
 60. Nordlund, J. & Syvänen, A.-C. Epigenetics in pediatric acute lymphoblastic leukemia. *Semin. Cancer Biol.* **51**, 129–138 (2018).
 61. Sajid, A., Rahman, H. & Ambudkar, S. V. Advances in the structure, mechanism and targeting of chemoresistance-linked ABC transporters. *Nat. Rev. Cancer* **23**, 762–779 (2023).
 62. Mosquera Orgueira, A. et al. Refining risk prediction in pediatric acute lymphoblastic leukemia through DNA methylation profiling. *Clin. Epigenet.* **16**, 49 (2024).
 63. Sun, Y., Liu, Y., Ma, X. & Hu, H. The influence of cell cycle regulation on chemotherapy. *Int. J. Mol. Sci.* **22**, 6923 (2021).
 64. Mohammadgholi, A., Rabbani-Chadegani, A. & Fallah, S. Mechanism of the Interaction of Plant Alkaloid Vincristine with DNA and Chromatin: Spectroscopic Study. *DNA Cell Biol.* **32**, 228–235 (2013).
 65. Basu, S., Dong, Y., Kumar, R., Jeter, C. & Tang, D. G. Slow-cycling (dormant) cancer cells in therapy resistance, cancer relapse and metastasis. *Semin. Cancer Biol.* **78**, 90–103 (2022).
 66. Frisimantas, V. et al. Ex vivo drug response profiling detects recurrent sensitivity patterns in drug-resistant acute lymphoblastic leukemia. *Blood* **129**, e26–e37 (2017).
 67. Malyukova, A. et al. Sequential drug treatment targeting cell cycle and cell fate regulatory programs blocks non-genetic cancer evolution in acute lymphoblastic leukemia. *Genome Biol.* **25**, 143 (2024).
 68. Palmisiano, N. D. et al. A phase 1 trial of venetoclax in combination with liposomal vincristine in patients with relapsed or refractory B-cell or T-cell acute lymphoblastic leukemia: Results from the ECOG-ACRIN EA9152 protocol. *eJHaem* **5**, 951–956 (2024).
 69. Chen, C. et al. Oxidative phosphorylation enhances the leukemogenic capacity and resistance to chemotherapy of B cell acute lymphoblastic leukemia. *Sci. Adv.* **7**, eabd6280 (2021).
 70. Diehl, F. F., Sapp, K. M. & Vander Heiden, M. G. The bidirectional relationship between metabolism and cell cycle control. *Trends Cell Biol.* **34**, 136–149 (2024).

71. Luengo, A., Gui, D. Y. & Vander Heiden, M. G. Targeting Metabolism for Cancer Therapy. *Cell Chem. Biol.* **24**, 1161–1180 (2017).
72. Bejjani, F., Evanno, E., Zibara, K., Piechaczyk, M. & Jariel-Encontre, I. The AP-1 transcriptional complex: Local switch or remote command? *Biochim. Biophys. Acta BBA - Rev. Cancer* **1872**, 11–23 (2019).
73. Zhang, Z. et al. JUN mediates glucocorticoid resistance by stabilizing HIF1a in T cell acute lymphoblastic leukemia. *iScience* **26**, 108242 (2023).
74. Gendelman, R. et al. Bayesian Network Inference Modeling Identifies TRIB1 as a novel regulator of cell-cycle progression and survival in cancer cells. *Cancer Res.* **77**, 1575–1585 (2017).
75. Ivorra, C. et al. A mechanism of AP-1 suppression through interaction of c-Fos with lamin A/C. *Genes Dev.* **20**, 307–320 (2006).
76. Canzoneri, R. et al. Identification of an AP1-ZFP36 Regulatory Network Associated with Breast Cancer Prognosis. *J. Mammary Gland Biol. Neoplasia* **25**, 163–172 (2020).
77. Iperi, C. et al. Integration of multi-omics analysis reveals metabolic alterations of B lymphocytes in systemic lupus erythematosus. *Clin. Immunol.* **264**, 110243 (2024).
78. Pekayvaz, K. et al. Multiomic analyses uncover immunological signatures in acute and chronic coronary syndromes. *Nat. Med.* **30**, 1696–1710 (2024).
79. De Visser, C. et al. Comprehensive multi-omics profiling of a healthy human cohort. Preprint at <https://doi.org/10.1101/2024.11.07.622407> (2024).
80. Moorman, A. V. et al. A novel integrated cytogenetic and genomic classification refines risk stratification in pediatric acute lymphoblastic leukemia. *Blood* **124**, 1434–1444 (2014).
81. Krali, O. et al. Multimodal classification of molecular subtypes in pediatric acute lymphoblastic leukemia. 0 Bytes Uppsala University <https://doi.org/10.17044/SCILIFELAB.22303531> (2024).
82. Enblad, A. P. et al. Ex-vivo Drug Responses and Molecular Genetic Profiles of 597 Pediatric ALL Patients. 0 Bytes Uppsala University <https://doi.org/10.17044/SCILIFELAB.26096371> (2024).
83. Krali, O. & Sulyaeva, J. Molmed/Krali_2026_Multiomics: 2026-03-09 Initial release. Zenodo <https://doi.org/10.5281/ZENODO.18923664> (2026).

Acknowledgements

This work was supported by grants from the Swedish Research Council (2019-01976 to JN), the Swedish Cancer Society (CAN2022-2395 to JN), and the Swedish Childhood Cancer Foundation (PR2022-0082 and HFT2023-0011 to JN, TJ2020-0039 to AH). We thank the SciLifeLab National Genomics Infrastructure, SNP&SEQ Technology Platform, which is funded by the Swedish Research Council and the Knut and Alice Wallenberg Foundation, for assistance with data generation. The data handling was enabled by resources provided by the Swedish National Infrastructure for Computing (SNIC) and the National Academic Infrastructure for Supercomputing in Sweden (NAISS). SNIC and NAISS are partially funded by the Swedish Research Council through grant agreement no. 2022-06725. We especially thank the ALL patients who contributed samples to this study.

Author contributions

J.N. and O.K. conceived the study. O.K., J.S., and A.L. analyzed the data. J.S. ran MOFA. O.K. performed downstream analyses and generated the figures and tables. GL and AH provided clinical material, data, and expertise. C.A. provided the FMCA data and expertise. JN provided funding. O.K., A.P.E., J.S., and J.N. wrote the manuscript. D.G. provided expertise in survival modelling and pathway analysis. T.E. and M.H. critically reviewed the manuscript and provided expertise in data integration. All authors read and approved the final version.

Funding

Open access funding provided by Uppsala University.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at

<https://doi.org/10.1038/s43856-026-01568-9>.

Correspondence and requests for materials should be addressed to Jessica Nordlund.

Peer review information *Communications Medicine* thanks Thomas Beder and the other anonymous reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2026