Supplementary Information

Methylome profiling enhances risk stratification and reveals context-dependent chemo-resistance mechanisms in acute myeloid leukemia

This PDF file includes:

Supplementary Methods Data processing, DMR/DMP and DEG analysis

DNA methylation raw intensity data were processed using the ChAMP pipeline (v2.30.0).[1] Quality control was performed with champ.QC, followed by comprehensive filtering to ensure data reliability. Low-quality probes with detection p-values > 0.01, bead counts < 3, SNP-related probes, multi-hit probes, and sex chromosome probes were filtered out. Samples with more than 10% failed probes were excluded to ensure high data quality. Beta values were normalized using the Beta-Mixture Quantile normalization (BMIQ) method implemented in champ.norm to correct for technical bias between type I and type II probes. Batch effects were identified through singular value decomposition analysis and corrected using ComBat.[2] For differentially methylated probe (DMP) analysis, we applied a linear modeling approach with empirical Bayes moderation from the limma package. DMPs were defined using a Benjamini-Hochberg (BH)-adjusted p-value threshold of 0.05 and an absolute beta value difference > 0.2 between comparison groups. Differentially methylated regions (DMRs) were identified using the Bumphunter algorithm with a minimum of 3 CpGs per region and a maximum gap of 1000bp between adjacent CpGs. Functional annotation of DMPs and DMRs was performed using genomic location and CpG island relation from the IlluminaHumanMethylationEPICanno.ilm10b4.hg19 annotation package. RNAsequencing raw reads underwent quality control using FastQC followed by adapter and low-quality base trimming with Trim Galore (q-score cutoff = 25, minimum length = 36bp). Clean reads were aligned to the GRCh19 reference genome using STAR (v2.7.11).[3] Transcript abundance was quantified using RSEM (v1.3.3) with fragmentlength estimation enabled, and gene-level counts were determined using featureCounts.[4] Differential gene expression (DEG) analysis was performed using DESeq2 with BH-adjusted p-value < 0.05 and absolute log2 fold-change > 1 as significance thresholds.[5]

Motif enrichment analysis

DMRs were converted to BED format, with each region extended by 200bp in both directions. We used HOMER (v4.11)[6] for motif discovery with the command 'findMotifsGenome.DMRs.bed hg19 ./motif_output -size given -gc -mask', which generates GC-content matched background sequences to control for nucleotide composition biases common in CpG-dense regions, while '-mask' excluded repetitive regions from analysis to prevent spurious motif detection.

Myeloid development signature

We utilized 3 public datasets, GSE35069, GSE50797, GSE63409 to develop a signature of normal hematopoietic development based unions of DMPs called from hematopoietic stem and progenitor cell (HSPC) vs each blood cell types on cutoff of logFC > 0.2 & BH adj.P.Val < 0.05, which includes 49,665 probes in total. A signature representing for myeloid lineage is developed using top 5,000 probes by comparing methylation level of HSPCs vs monocytes. Probes were classified into distinct aberrant

methylation events based on the following criteria: "Aberrant Hypermethylation" (AML samples vs HSPC difference exceeds monocyte vs HSPC difference by more than 0.2, where monocyte vs HSPC difference > -0.2), "Aberrant Hypomethylation" (AML samples vs HSPC difference is more than 0.2 below monocyte vs HSPC difference, with AML samples vs HSPC difference < -0.2), and "Failed Hypomethylation" (AML samples vs HSPC difference exceeds monocyte vs HSPC difference by more than 0.2, where monocyte vs HSPC difference < -0.2). Probes not meeting these criteria were classified as "No change". For direct RR vs CR comparison, criteria as: "Aberrant Hypermethylation" (beta value > 0.5 in RR and difference compared to CR > 0.2), "Aberrant Hypomethylation" (beta value < 0.5 in RR and difference compared to CR < -0.2), and "Failed Hypomethylation" (beta value > 0.5 in RR while beta value < 0.5 in CR, with difference > 0.2). Probes not meeting these criteria were classified as "No Change".

Chromatin states were defined using ChromHMM25 with a 15-state model based on six histone modifications (H3K4me3, H3K4me1, H3K27ac, H3K27me3, H3K36me3, H3K9me3) in three HSPC samples from the Roadmap Epigenomics Project (E035, E036, E050). The resulting model captures distinct chromatin signatures corresponding to promoters, enhancers, transcribed regions, and repressed domains. For enrichment analysis, we used LOLA26 to test for significant overlap between methylation patterns and chromatin states, using a hypergeometric test with BH correction for multiple testing (q -value < 0.05). Enrichment scores were calculated as log2 odds ratios of observed versus expected overlap.

Model training and cross-validation

We evaluated machine learning algorithms including Support Vector Machines (SVM) with radial basis function kernel, Random Forest, Extreme Gradient Boosting (XGBoost), and Elastic Net regression. For each algorithm, hyperparameters were optimized through exhaustive grid search (e.g., 2,187 combinations for XGBoost and120 for Elastic Net) using 100-fold bootstrap cross-validation with 90% sampling per iteration. In each bootstrap iteration, 90% of training samples were randomly selected with replacement for model fitting, with the remaining 10% serving as validation. Model selection was based on mean AUROC across all bootstrap iterations. The independent test cohort was completely held out from all training, feature selection, and hyperparameter optimization steps, ensuring unbiased evaluation of generalization performance.

Feature selection and importance analysis

For initial feature selection, we select from top 20% significantly differentially methylated CpG probes between RR and CR groups (two-sample t-test with BH

correction for multiple testing (adjusted P < 0.05)) and rank based on their significance. To identify CpG probes most predictive of treatment resistance, we implemented a permutation-based feature importance analysis using the trained SVM classifier. For each feature i in our model, importance was quantified as:

$$I_i = \text{AUC}_{\text{full}} - \text{AUC}_{\text{i-permuted}}$$

where AUC_{full} is the performance with intact data and $\text{AUC}_{\text{i-permuted}}$ is the performance after randomly shuffling values of feature i while preserving all other features. To ensure statistical robustness, we performed five independent permutations per feature and calculated the mean importance I_i and standard deviation σ_i. Features were ranked by their standardized score:

$$Z_i = \frac{\bar{I}_i}{\sigma_i}$$

Features with |Z| > 2.0 were considered as significant contributors to classification performance. This threshold corresponds to P < 0.05, identifying methylation markers with discriminative power beyond random chance.

Methylation-specific polymerase chain reaction (MSP)

AM cell line KG1, THP1, HL60 and MOLM-14 were obtained from The American Type Culture Collection (ATCC, Manassas, VA, USA). Genomic DNA from AML cell lines was isolated by DNeasy Blood and Tissue Kit (Hilden, Germany) and chemically modified by using an EZ DNA Methylation-Gold™ Kit according to the manufactory's instruction (Zymo Research, Irvine, CA, USA). The methylation status of PTX4 promoter region was determined by MSP. The PTX4 promoter region is located approximately1 kb up stream and up to 200 bp downstream of the transcription initiation site.

PTX4 M-forward primer sequence: 5'- GTTGTTAATTTAAAAATTCGATTGC-3';

PTX4 M-reverse primer sequence: 5'- CAATAAAACCCAAAAAACCGTT -3';

PTX4 U-forward primer sequence: 5'- TGTTAATTTAAAAATTTGATTGTGT-3';

PTX4 U-reverse primer sequence: 5'- CCAATAAAACCCAAAAAACCAT-3'.

qRT-PCR

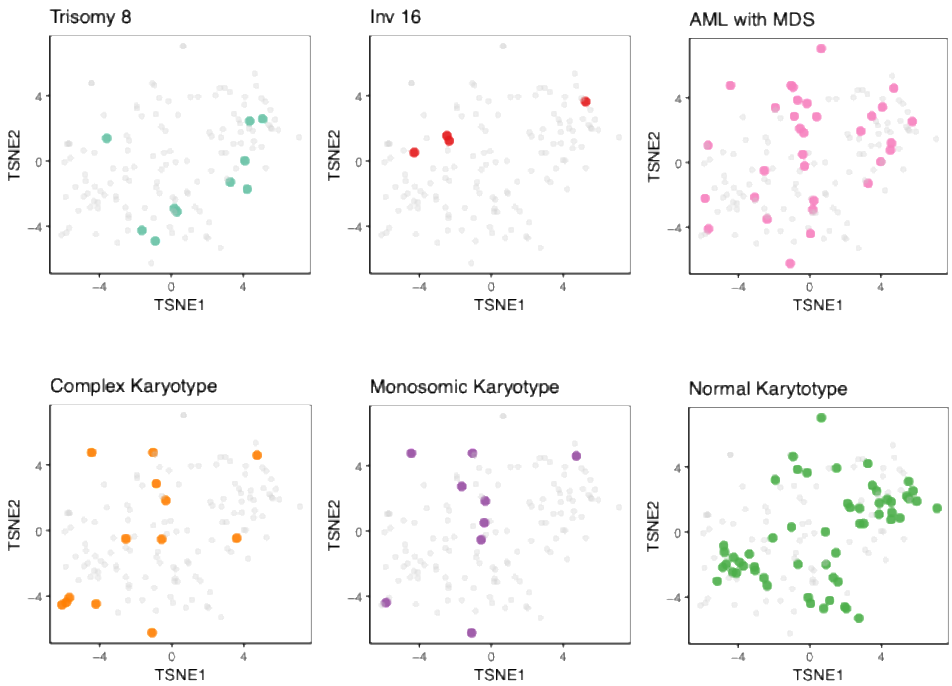
To quantify PTX4 expression, 1 μg of total RNA was used to generate cDNA by using SuperScript® III RT (Thermo Fisher Scientific, Waltham, MA, USA) with oligo-dT primer. qRT-PCR was performed using the Power SYBR® Green PCR Master Mix as recommended by the manufacturer (Applied Biosystems). GAPDH was used as the internal control. SDS 2.2.1 software (Applied Biosystems) was used to perform relative quantification of target genes using the comparative C_T ($\Delta\Delta C_T$) method. The primer sequences for PTX4 gene are PTX4-Forward: 5′-GCCGAGGCCGGAAAGTA GAC-3′ and PTX4-Reverse: 5′-CTCAGGCTGGGTTGGGGTTG-3′.

Immunoblotting analysis

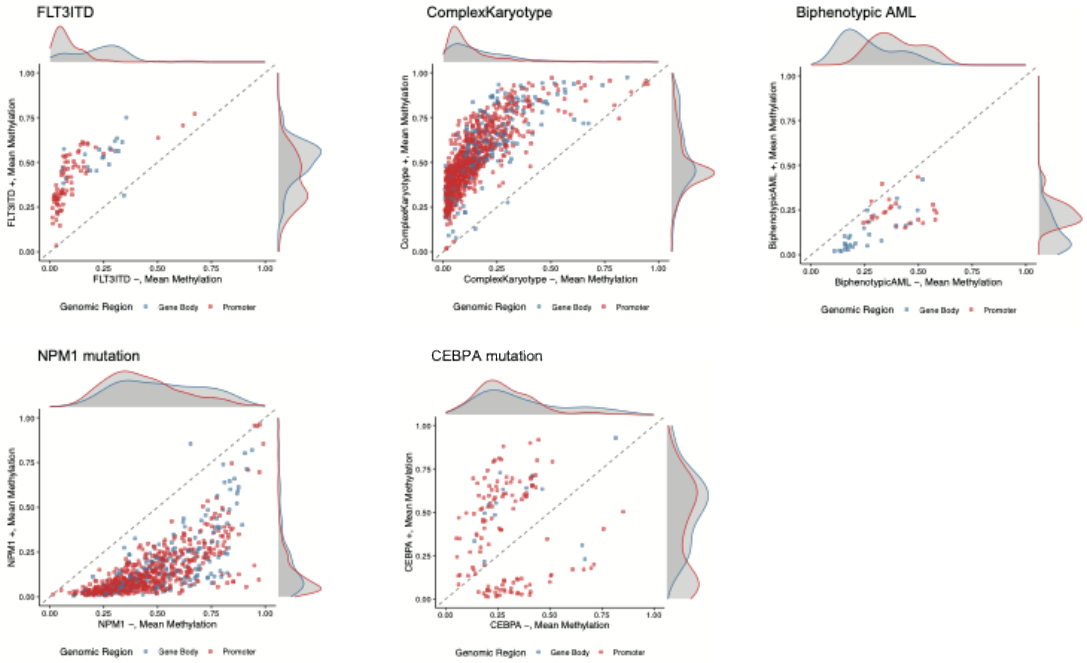
THP1 cells were incubated with 0.5 μM and 1 μM 5-Aza, or DMSO for 96 hours. Cells were harvested and lysed with radio-immunoprecipitation assay (RIPA) buffer (20 mM HEPES at pH 7.4, 1% Triton X-100,150 mM NaCl, 1 mM EDTA, 1 mM EGTA, and 1X Protease Arrest). SDS-PAGE and immunoblot analyses performed with different antibodies. Anti-PTX4 antibody was purchased from ThermoFisher (Cat # PA5-45841) and anti-□-actin from obtained from Santa Cruz Biotechnology (Santa Cruz, CA, USA).

RNA-seq and data analysis

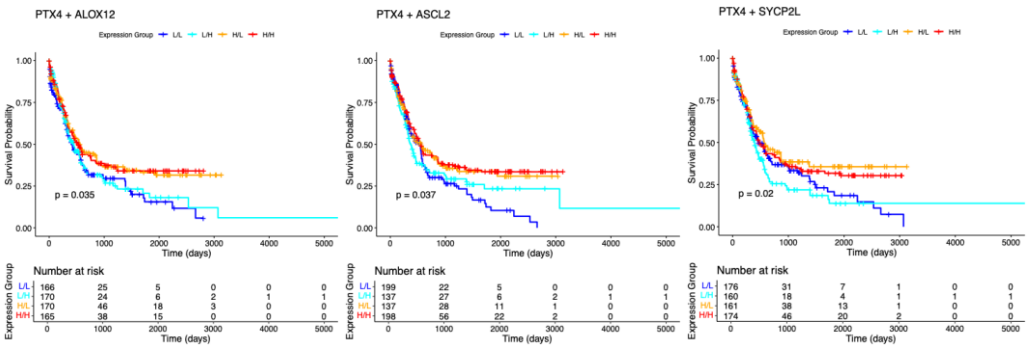
Total RNA was extracted using the RNeasy mini kit (Qiagen) from three biological replicates of THP-1-EV and THP-1-PTX4-OE cells. The RNA library construction and RNA-sequencing services were performed by service provider, Novogene Singapore. Raw RNA-seq reads were first assessed for quality using FastQC v0.11.5, followed by trimming to remove adaptor sequences, contaminants, and low-quality reads. Cleaned reads were aligned to the human reference genome (hg38) using STAR v2.4.2a. Transcript abundance was then estimated with RSEM v1.2.25 based on Gencode v24 annotations. Differential gene expression analysis was performed using EBSeq v1.20.0. Statistical comparisons were carried out using the t-test implemented in MATLAB® R2012a (Statistics Toolbox v8.0; MathWorks, Natick, MA). Functional enrichment, including pathway analyses, was conducted using the GSVA package (v1.28.0) from Bioconductor in R.



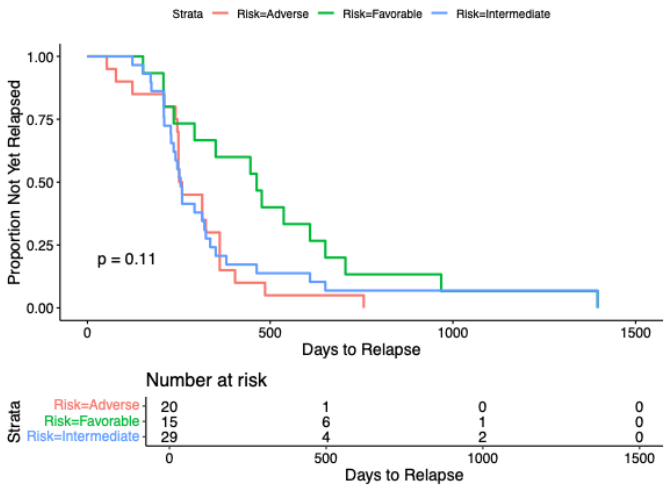
Supplementary Figure S1. Multidimensional scaling (MDS) plot of the top 100 differentially methylated positions (DMPs) across patient samples, color-coded by mutation/cytogenetic status.



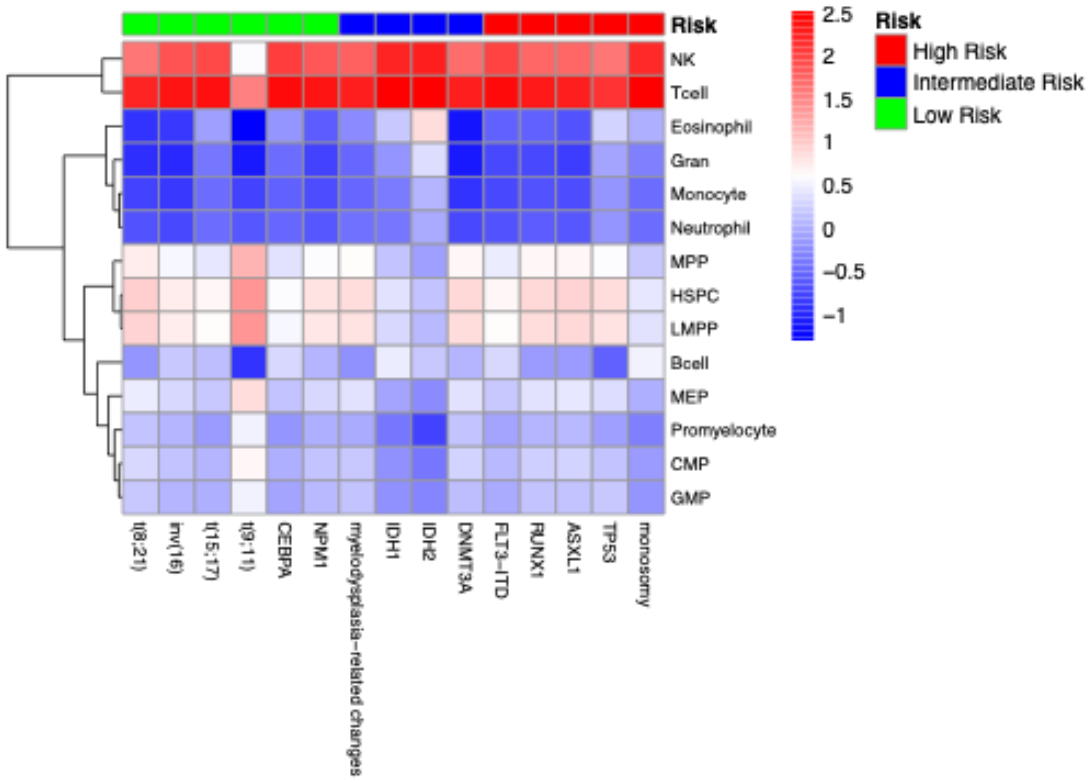
Supplementary Figure S2. Mean methylation levels of probes at DMR loci categorized by its location in gene body (blue) or promoter (red) at samples within each mutation/cytogenetic category versus those are not.



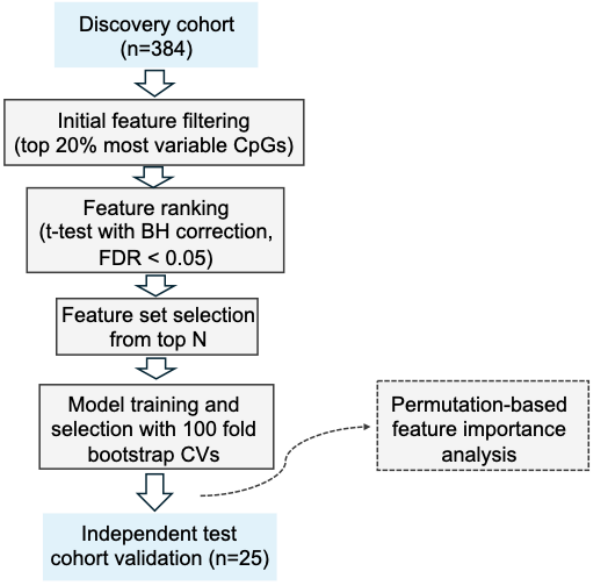
Supplementary Figure S3. Combinatorial survival analysis of PTX4 with three example genes whose functional regions exhibited differential methylation in Beat-AML cohort. From left to right: ALOX15, ASCL2, and SYCP2L. Patients were stratified by median expression levels (L: low, below median; H: high, above median).



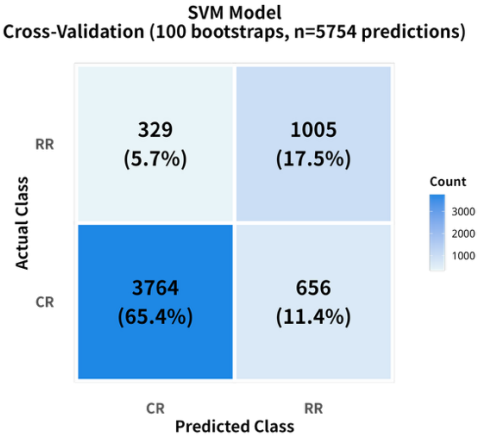
Supplementary Figure S4. Time-to-relapse of patients belonging to different ELN risk groups.



Supplementary Figure S5. Euclidean distance of AML samples with different mutation/cytogenetic changes (categorized into risk groups) to the centroid of each normal cell type of various blood cell developmental stages. Heatmap color intensity represents distance magnitude, with red indicating greater distances and blue indicating closer proximity to normal cell type centroids.



Supplementary Figure S6. Schematic workflow of machine learning-based methylome classifier development and validation.



Supplementary Figure S7. Confusion matrices of results aggregated from 100 bootstrap validations.

References

1. Tian Y, Morris TJ, Webster AP, et al. ChAMP: updated methylation analysis pipeline for Illumina BeadChips. *Bioinformatics*. 2017;33(24):3982-3984.
2. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*. 2007;8(1):118-127.
3. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013;29(1):15-21.
4. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*. 2011;12:323.
5. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. 2014;15(12):550.
6. Heinz S, Benner C, Spann N, et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell*. 2010;38(4):576-589.