Blood Science

OPEN

Charting epimutation dynamics in human hematopoietic differentiation

Xiaohuan Qin[a], Jiayi Lu[a], Peng Wu[a], Chunyong Zhang[a,b], Lei Shi[b,*], Ping Zhu[a,*]

[a]State Key Laboratory of Experimental Hematology, National Clinical Research Center for Blood Diseases, Haihe Laboratory of Cell Ecosystem, Department of Stem Cell and Regenerative Medicine, Institute of Hematology and Blood Diseases Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Tianjin, China; [b]The Province and Ministry Co-sponsored Collaborative Innovation Center for Medical Epigenetics, Key Laboratory of Immune Microenvironment and Disease (Ministry of Education), Department of Biochemistry and Molecular Biology, School of Basic Medical Sciences, Tianjin Medical University, Tianjin, China

Abstract

DNA methylation plays a critical role in hematopoietic differentiation. Epimutation is a stochastic variation in DNA methylation that induces epigenetic heterogeneity. However, the effects of epimutations on normal hematopoiesis and hematopoietic diseases remain unclear. In this study, we developed a Julia package called EpiMut that enabled rapid and accurate quantification of epimutations. EpiMut was used to evaluate and provide an epimutation landscape in steady-state hematopoietic differentiation involving 13 types of blood cells ranging from hematopoietic stem/progenitor cells to mature cells. We showed that substantial genomic regions exhibited epigenetic variations rather than significant differences in DNA methylation levels between the myeloid and lymphoid lineages. Stepwise dynamics of epimutations were observed during the differentiation of each lineage. Importantly, we found that epimutation significantly enriched signals associated with lineage differentiation. Furthermore, epimutations in hematopoietic stem cells (HSCs) derived from various sources and acute myeloid leukemia were related to the function of HSCs and malignant cell disorders. Taken together, our study comprehensively documented an epimutation map and uncovered its important roles in human hematopoiesis, thereby offering insights into hematopoietic regulation.

Key Words: DNA methylation; Epimutation; Hematopoietic differentiation; Hematopoietic stem cell

1. INTRODUCTION

The hematopoietic system is a well-known model for studying the regulatory network in cell fate decision and lineage differentiation.[1,2] In recent decades, DNA methylation has been investigated and proven to be closely involved in the maintenance of hematopoietic stem cell (HSC) function and hematopoietic differentiation.[3,4] For instance, myeloid cells undergo demethylation, whereas lymphoid cells establish a de novo methylation landscape,[5] ensuring precise stepwise differentiation through DNA methylation. Previous genome sequencing studies have revealed that enzymes, including writers and erasers of DNA methylation, have frequent mutations in blood malignancies. In particular, 5-Aza-2-deoxycytidine has been used to treat patients with myelodysplastic syndrome by inhibiting DNA methylation.[6]

Stochastic changes in DNA methylation, also known as epimutation, generate epigenetic heterogeneity. This turnover of cytosine modification is common and has been shown to occur at regulatory loci in a sequence-dependent manner.[7] Epimutation also acts as an important epigenetic molecular clock recording the lineage history of cells in the same phylogenetic tree. This characteristic of a clock is well illustrated by revealing the relationship of epigenetic information to the physiological ages of tissues.[8] It has also been used to trace the lineage relationship in blast cells of patients with chronic lymphocytic leukemia.[9] However, the epimutation landscape and its potential regulation in human hematopoiesis are unknown.

Several metrics have been proposed to evaluate epimutation, including the proportion of discordant reads (PDR),[9] methylation concurrence ratio (MCR),[10] and quantitative fraction of discordant read pairs (qFDRP).[11] PDR is used to evaluate locally disordered DNA methylation. MCR conceptualizes the degree of concurrence between active methylation and demethylation. However, MCR underestimates the real concurrence because bisulfite-seq fails to distinguish 5 hmC from 5 mC. qFDRP quantifies the disagreement between pairs of methylation states observed in the sequencing reads.

**Address correspondence: Lei Shi, The Province and Ministry Co-sponsored Collaborative Innovation Center for Medical Epigenetics, Key Laboratory of Immune Microenvironment and Disease (Ministry of Education), Department of Biochemistry and Molecular Biology, School of Basic Medical Sciences, Tianjin Medical University, Tianjin, China. E-mail address: shilei@tmu.edu.cn (L. Shi); Ping Zhu, State Key Laboratory of Experimental Hematology, National Clinical Research Center for Blood Diseases, Haihe Laboratory of Cell Ecosystem, Department of Stem Cell and Regenerative Medicine, Institute of Hematology and Blood Diseases Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Tianjin, China. E-mail address: zhuping@ihcams.ac.cn (P. Zhu).*

Conflict of interest: The authors declare that they have no conflict of interest.

X.Q., J.L., P.W., C.Z. contributed equally to this work.

This work was supported by the National Science Foundation of China (82022002, 82131430173, 81900117, 81890993), the National Key Research and Development Program of China (2021YFA1102800), the Distinguished Young Scholars of Tianjin (21JCJQJC00070), the CAMS Initiative for Innovative Medicine (2021-I2M-1-040), and the Foundation of Haihe Laboratory (HH22KYZX0018).

Blood Science (2024) 6, 1–10:e00197.

Received March 31, 2024; Accepted May 26, 2024.

http://dx.doi.org/10.1097/BS9.0000000000000197

Nonetheless, current available tools for evaluating these metrics demonstrate common limitations. One major limitation is that genomic variations, including insertion/deletion (indel) and single nucleotide polymorphism (SNP), are not considered, which may lead to errors in the estimation of epimutation. Another critical issue is that the calculation is time-consuming because the data volume of the DNA methylomes is large. Optimized utilities are required to overcome these limitations.

In this study, we developed an open-source tool called EpiMut to facilitate the rapid and accurate measurement of epimutation. Using EpiMut, we comprehensively established the epimutation landscape of human blood cells, including stem/progenitor cells and mature cells from multiple lineages. We also illustrated the unique patterns of epimutation in distinct HSCs and malignant cells from acute myeloid leukemia (AML) patients. Our findings highlight the prevalent involvement of epimutations in hematopoiesis and provide new insights into hematopoietic regulation and malignancy pathogenesis.

2. RESULTS

2.1. EpiMut was developed to assess epimutation

The available tools used to evaluate epimutations are time-consuming because bisulfite sequencing data are typically large. To address this, we first developed a Julia package named EpiMut to facilitate a more rapid and accurate estimation of epimutations (**Fig. 1A**). EpiMut estimates 3 metrics: PDR, MCR, and qFDRP (Supplemental Figure 1A and B, http://links.lww.com/BS/A104). Bisulfite treatment induces damage to DNA templates and generates genomic variations in sequencing libraries. SNPs are known to have a high frequency of C-T conversion, making the accurate evaluation of epimutation a challenge. To address this, we optimized a pipeline to retain sequencing reads containing indels rather than discarding them. This approach can rectify mismatches at CpG sites and rescue additional sequencing reads. In addition, we only estimated the methylation of CpG sites without overlapping with known SNPs, thus avoiding interference from SNPs. Therefore, EpiMut rescued more sequencing reads and CpG sites to improve the accuracy of epimutation calculations (**Fig. 1B–D** and Supplemental Figure 1C–K, http://links.lww.com/BS/A104). Furthermore, considering that epimutation estimation requires intensive computation, EpiMut offers the flexibility to calculate three distinct metrics to quantify the epimutation rate, leverage multiple cores, and utilize the Julia programming language to significantly enhance computational speed (**Fig. 1E** and Supplemental Figure 1D and H, http://links.lww.com/BS/A104). Downstream analyses, including the determination of differential epimutation, gene ontology (GO) enrichment, and genomic region enrichment, were integrated into EpiMut (**Fig. 1A**). Therefore, we anticipate that EpiMut will pave the way for the investigation of epimutation heterogeneity owing to its fast analysis of large sequencing data and accurate estimation of various epimutation metrics.

2.2. Epimutation landscape in human blood cells

Next, we examined the epimutation landscape of normal human blood cells. We used EpiMut to quantify the PDR values of 7 groups of hematopoietic stem/progenitor cells and 6 types of mature cells derived from normal human peripheral blood (PB). There was a notable increase in genome-wide epimutation from stem/progenitor cells to mature cells (**Fig. 2A**, Supplemental Figure 2A, http://links.lww.com/BS/A104). When comparing epimutation in different genomic regions, CpG

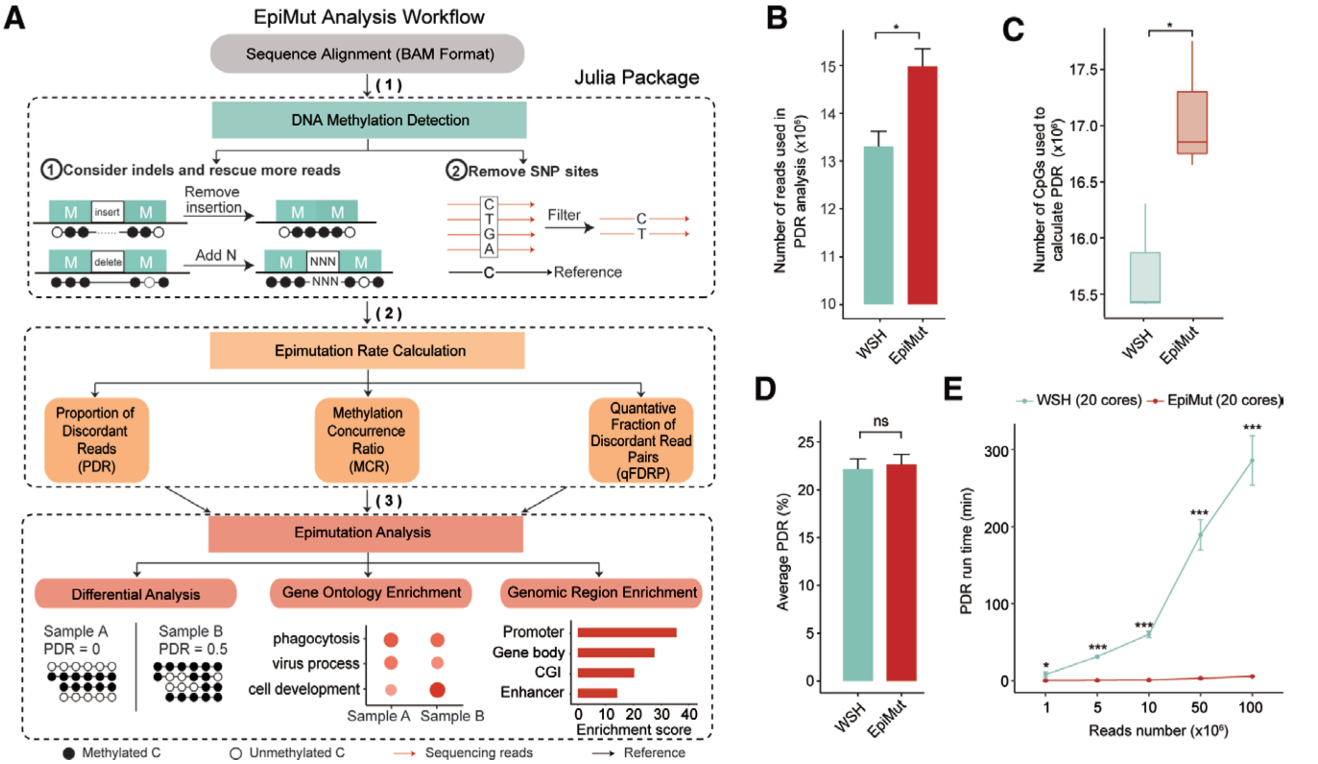

Figure 1. The development and evaluation of EpiMut. (A) The workflow of EpiMut for epimutation rate measurement. (B) Comparison of reads used in PDR analysis between EpiMut and WSH. (C) Comparison of covered CpG sites in PDR analysis between EpiMut and WSH. (D) Comparison of average PDR values calculated by EpiMut and WSH. (E) The running time of PDR calculation at different sequencing depths in EpiMut and WSH. *$P < .05$; ***$P < .001$ (unpaired 2-tailed Student *t* test). ns = not significant, PDR = proportion of discordant reads.

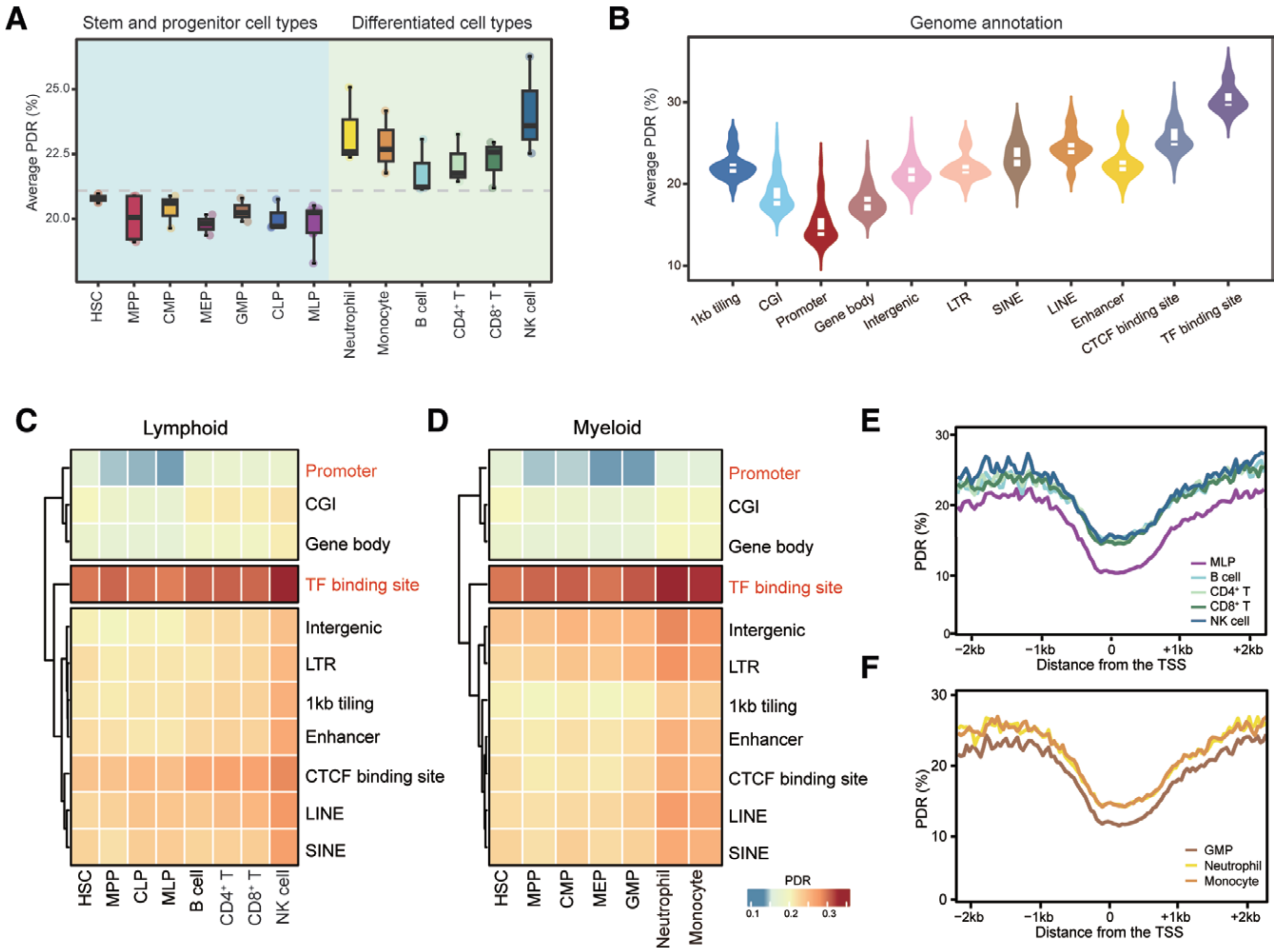

Figure 2. Overview of epimutation among human hematopoietic cell types. (A) Boxplots display the average PDR level of 13 hematopoietic cell types, targeting 1000 cells within each library. The gray dashed line represents the mean PDR level of all cell types. (B) Violin plots showing the PDR values across all cell types for different genomic regions. (C–D) Heatmap illustrating absolute PDR level across genomic regions in (C) lymphoid and (D) myeloid cells. (E–F) The average PDR levels of (E) lymphoid cells and (F) myeloid cells around the TSS (±2 kb) of all RefSeq genes. CGI = CpG islands, CLP = common lymphoid progenitors, CMP = common myeloid progenitors, CTCF = CCCTC-binding factor, GMP = granulocyte/monocyte progenitors, HSC = hematopoietic stem cell, LINE = long interspersed nuclear element, LTR = long terminal repeat, MEP = megakaryocyte/erythroid progenitor, MLP = multi-lymphoid progenitor, MPP = multipotent progenitor, NK = natural killer, PDR = proportion of discordant reads, SINE = short interspersed nuclear element, TF = transcription factor, TSS = transcription start site.

islands (CGI) and promoters exhibited the lowest PDR values, whereas regulatory regions, such as transcription factor (TF) binding sites, presented the highest PDR values (**Fig. 2B** and Supplemental Figure 2B, http://links.lww.com/BS/A104). These dynamic patterns of epimutation were also verified when only lymphoid or myeloid cells were analyzed (**Fig. 2C and D**). In particular, analysis of the distribution of PDR on the transcription start site (TSS) and their adjacent 2-kilobase (kb) regions revealed that mature cells exhibited a higher PDR than their progenitors within both lymphoid and myeloid lineages (**Fig. 2E and F**). This pattern of higher PDR in mature cells was observed despite the generally low PDR values across all cell types. This indicated that changes in epimutation correlated with hematopoietic differentiation, even within regions characterized by low epimutation levels. During hematopoietic differentiation, mature cells rapidly transform into a more disordered state. The elevated epimutation rates also accurately recorded cell differentiation history as an epigenetic molecular clock. Remarkably, low epimutation in the CGI and promoter regions suggests that unmethylated or methylated states, rather than disordered methylation states, are required to regulate gene expression. In contrast, high epimutation in TF binding sites indicates a frequent turnover of DNA methylation to dynamically regulate TF binding, representing the role of gene regulation by epimutation changes.

2.3. Epimutation distinguishes myeloid and lymphoid progenitors

One major event in hematopoietic differentiation is the lineage commitment of myeloid and lymphoid cells. To understand the relationship between epimutation and lineage commitment, differentially discordant regions (DDRs) of epimutation were identified between common myeloid progenitors (CMP) and common lymphoid progenitors (CLP). In total, 4186 DDRs were identified in the CMP and CLP groups. Specifically, 1866 DDRs exhibited lower PDR values in CMP, whereas 2320 DDRs showed lower PDR values in CLP (**Fig. 3A**). Notably, a few DDRs showed differential DNA methylation patterns between CMP and CLP (Supplemental Figure 3A, http://links.lww.com/BS/A104). This suggests that epimutation may provide more information on lineage commitment because changes in DNA methylation between lineages are subtle.

To examine the potential functional role of epimutations, we identified DDR-associated genes. GO enrichment analysis showed that regions with lower PDR in the CMP were enriched for myeloid development. In contrast, adaptive immunity, innate immunity, and immune processes were enriched in the CLP group (**Fig. 3B**). We further examined the PDR values of stemness-, myeloid-, and lymphoid-associated genes in HSC, CMP, and CLP.[12,13] We found that stemness-related genes had

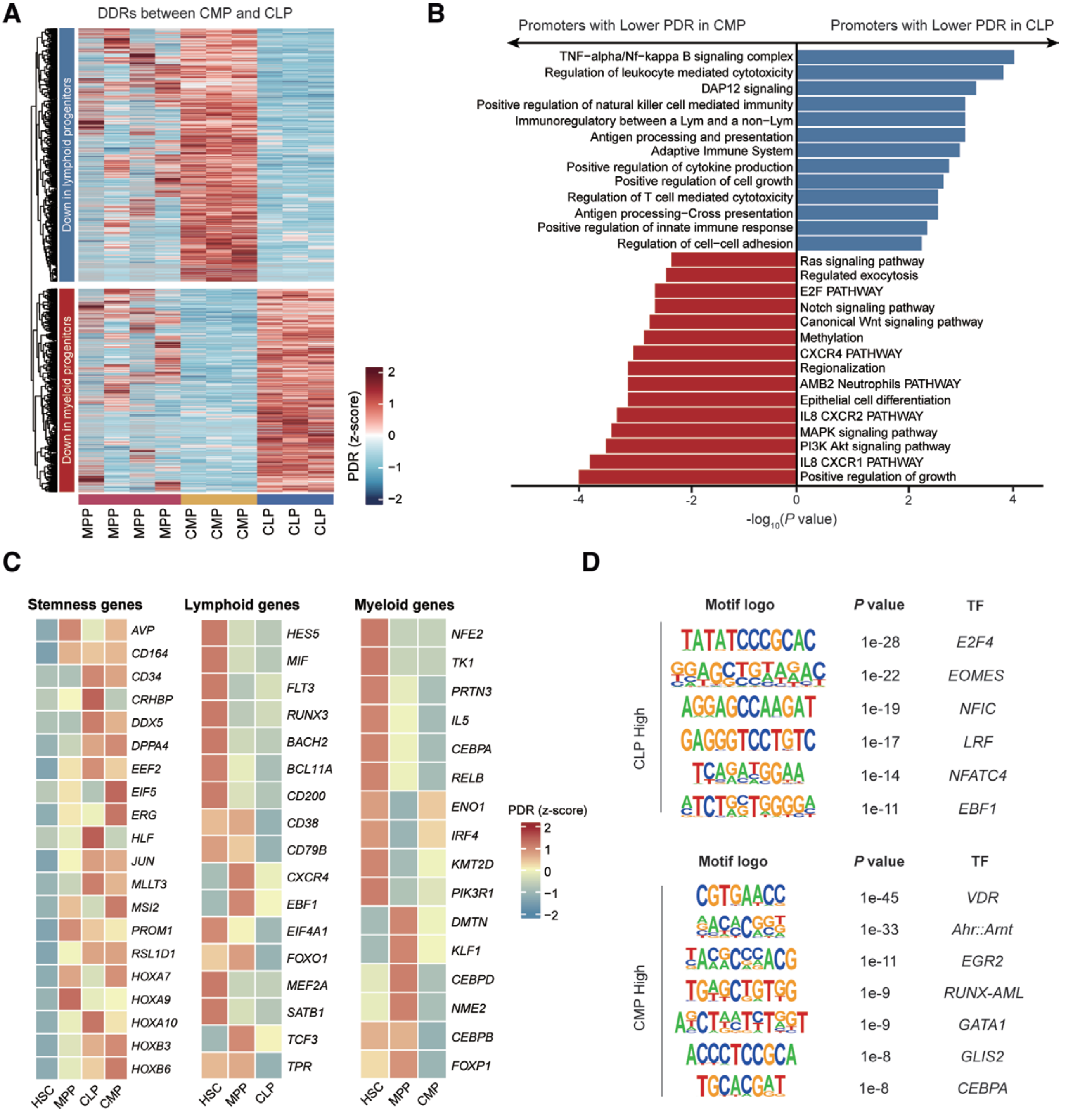

Figure 3. Epimutation distinguishes myeloid and lymphoid progenitors. (A) Heatmap showing the row-scaled PDR of DDRs defined in CMP and CLP. (B) GO enrichment analysis of DDR-related genes between CLP and CMP. The x-axis represents the negative logarithm of the *P* values. (C) Heatmaps showing the row-scaled PDR values within the promoters of stemness, lymphoid-specific, and myeloid-specific genes across hematopoietic progenitor cells. (D) TF motif enrichment of DDRs between CLP and CMP. CLP = common lymphoid progenitor, CMP = common myeloid progenitor, DDRs = differentially discordant regions, GO = gene ontology, HSC = hematopoietic stem cell, MPP = multipotent progenitor, PDR = proportion of discordant read, TF = transcription factor.

lower PDR values in HSC, along with the lowest PDR values of myeloid-related genes in CMP and the lowest PDR values of lymphoid-related genes in CLP (**Fig. 3C** and Supplemental Figure 3C and D, http://links.lww.com/BS/A104). TF motif enrichment analysis showed that the TF binding sites of well-known myeloid TFs, including EGR2, GATA1, and CEBPA, displayed significantly higher epimutation rates in CMP than in CLP.[14] Meanwhile, regions exhibiting higher PDR values in lymphoid progenitors had pronounced enrichment of TF binding sites associated with lymphoid TFs, such as EOMES and EBF1 (**Fig. 3D**).[15] These results suggest that functional elements are involved in the regulation of lineage commitment of blood cells. We also presented the PDR values of CLP-specific and CMP-specific DDRs in MPPs. However, we did not observe any epimutation patterns clearly associated with lineage differentiation within MPPs, indicating that the epimutation features of CMP and CLP gradually accumulate during the differentiation of MPPs into downstream progenitor cells.

2.4. Accumulated epimutation during myeloid and lymphoid differentiation

To explore the effect of epimutation during lineage differentiation, we compared epimutation between progenitors and

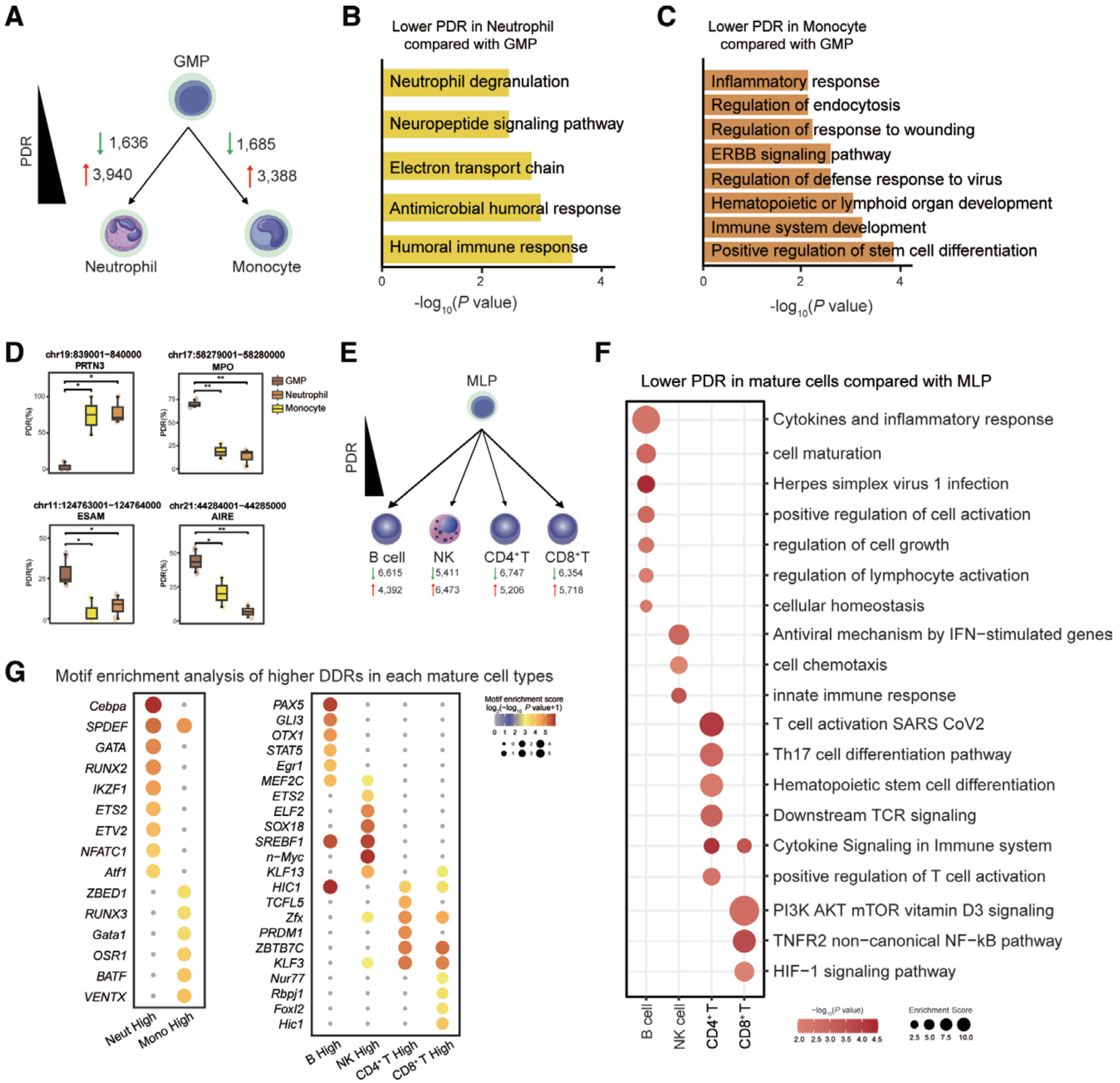

Figure 4. Different epimutation between progenitors and terminally differentiated cells. (A) Illustration depicting the numbers of DDRs in pairwise comparisons between GMP and neutrophil/monocyte. (B–C) GO enrichment analysis of genes related to DDRs with lower PDR values in (B) neutrophils and (C) monocytes compared to GMPs. The x-axis represents the negative logarithm of the *P* values. (D) Boxplots showing PDR values of cell type-specific genes across GMP, neutrophil, and monocyte. **P* < .05, ***P* < .01 (unpaired 2-tailed Student *t* test). (E) Illustration depicting the numbers of DDRs in pairwise comparisons between MLP and mature cells (NK, B, $CD4^+$, and $CD8^+$ T cells). (F) GO enrichment analysis of genes with lower PDR values in lymphoid mature cells (NK, B, $CD4^+$, and $CD8^+$ T cells) than MLP, respectively. The size of the dots represents the enrichment score, while the color indicates the negative logarithm of the *P* values. (G) TF motif enrichment of DDRs with higher PDR values in mature cells than progenitors. Both the colors and sizes of the dots indicate the motif enrichment scores. DDRs = differentially discordant regions, GMP = granulocyte/monocyte progenitor, GO = gene ontology, MLP = multi-lymphoid progenitor, NK = natural killer, PDR = proportion of discordant read, TF = transcription factor.

terminally differentiated cells within the myeloid and lymphoid lineages. From granulocyte/monocyte progenitors (GMP) to neutrophils and monocytes, PDR showed an increase in 3940 and 3388 regions, while a decrease was observed in 1636 and 1685 regions, respectively (**Fig. 4A**). There were similar PDR changes from lymphoid progenitors to natural killer (NK), B, $CD4^+$, and $CD8^+$ T cells (**Fig. 4E**). Notably, consistent with the limited overlap between DDRs and differentially methylated regions (DMRs) in lineage commitment, 90% of the DDRs identified between progenitors and mature cells did not overlap with DMRs (Supplemental Figure 4A–F, http://links.lww.com/BS/A104).

Epimutation dynamics are also functionally relevant during lineage differentiation. GO enrichment analysis showed that genomic regions exhibited a lower PDR in neutrophils than in GMP enriched in neutrophil-related biological processes (**Fig. 4B**). Additionally, regions with a lower PDR in monocytes were enriched for inflammation and endocytosis (**Fig. 4C**). We also examined the PDR values of the cell type-specific gene signatures, and the distribution supported a unique pattern of epimutation in each cell population (**Fig. 4D**). For instance, *PRTN3* exhibited the lowest PDR values with high expression in GMP.[16] During lymphoid differentiation, promoters with lower PDR values were significantly enriched for defense mechanisms against viral infections, innate immune responses, and T cell activation (**Fig. 4F**). These results illustrate that differentiation-related genes require the transition of epimutation from a disordered to an ordered state, with low epimutation rates in their promoter regions. This may be an intermediate demethylation process for the activation of cell type-specific genes. However, the more disordered states of TF binding sites in mature cells were specifically linked to the regulation of cell differentiation (**Fig. 4G** and Supplemental Figure 4G and H, http://links.lww.com/BS/A104).

2.5. Heterogeneity of epimutation in HSCs from distinct sources

To delineate the impact of epimutation on HSCs during development, we conducted a comparative analysis of epimutation rates in HSCs derived from 4 different sources: the fetal liver (FL), cord blood (CB), bone marrow (BM), and PB. Notably, HSCs derived from the adult BM had the lowest PDR, which was inconsistent with the expectation of epimutation accumulation during development (**Fig. 5A**). Although enhancers and TF binding sites showed the lowest PDR in PB, BM showed the lowest PDR in most genomic regions (**Fig. 5B and C**). Next, we identified source-specific DDRs in HSCs. In total, 1579, 389, 2239, and 8922 DDRs were identified, with the lowest PDR values in FL, CB, BM, and PB, respectively. GO enrichment analysis revealed that the regions with the lowest PDR values in the FL were related to embryonic development. In particular, several key pathways essential for HSC function were enriched in the BM, including the RHO GTPase cycle signaling, AKT signaling, WNT signaling, BMP pathway, Runx1 pathway, and Notch signaling pathway.[17–20] Regarding PB HSCs, these regions exhibited enrichment in HS-GAG degradation and cell differentiation (**Fig. 5D**).[21] Subsequently, the scores of HSC-related genes were computed using the PDR values of the gene promoters. BM had the lowest score across the quiescence, MYC targets, E2F targets, and stemness pathways (**Fig. 5E**). Representative genes (**Fig. 5F** and Supplemental Figure 5A–F, http://links.lww.com/BS/A104) and TFs (**Fig. 5G**) supported the source-specific distribution of epimutations in HSCs. These results suggest that the niche of BM HSCs plays an important role in maintaining HSC function, as represented by the distinct epimutation status.

2.6. Abnormal epimutation in AML

The enzymes involved in DNA methylation are frequently mutated in malignancies. This directly indicates the involvement of DNA methylation in abnormal hematopoiesis. Therefore, we investigated the changes in epimutation in patients with mixed-lineage rearranged AML (MLL-r AML). We observed a notable elevation in PDR values in comparison to normal HSC (**Fig. 6A**, Supplemental Figure 6A, http://links.lww.com/BS/A104). Moreover, the PDR values were higher across all genomic regions than in normal cells (**Fig. 6B**). These results suggest a higher level of epigenetic diversity in AML, indicating a more active cell cycle and proliferation of malignant cells. DDRs with lower PDR in AML were mainly enriched in the negative regulation of the immune response, cell proliferation, and activation of the Notch signaling pathway (**Fig. 6C** and Supplemental Figure 6B and C, http://links.lww.com/BS/A104). This result aligns with the immune dysfunction and blast-cell expansion in AML. In addition, the promoter regions corresponding to these known leukemia driver genes showed lower PDR levels in AML (Supplemental Figure 6D, http://links.lww.com/BS/A104), suggesting that increased expression of these genes is potentially correlated with the pathogenesis of AML.

Furthermore, we identified 2 target genes, *ARHGAP22* and *PARD6A*, annotated from DDRs with low PDR values in AML. These 2 candidates have been reported as promising therapeutic targets in cancer.[22,23] *ARHGAP22* is implicated in tumor cell motility and plays a role in the survival-mediated effects of Akt signaling.[24] In addition, *PARD6A* has been identified as an inducer of cell migration and invasion, contributing to the metastasis of ovarian cancer.[23] In our study, lower epimutation rates as well as higher expression of *ARHGAP22* and *PARD6A* were found to be correlated with shorter survival in AML (**Fig. 6D**). These results indicate that epimutations could serve as an indicator for investigating the clinical outcomes of hematological malignancies.

3. DISCUSSION

DNA methylation is essential for normal biological functions.[25] The stochastic fluctuations in DNA methylation lead to epigenetic heterogeneity. To address the limitations of current available software for epimutation analysis, we first developed EpiMut, a novel tool designed for the rapid and accurate quantification of epimutations using DNA methylation sequencing data. In this study, we documented the epimutation landscape of blood cells and emphasized the potential regulatory role of epimutation in hematopoietic differentiation and malignancies.

During lineage commitment and differentiation, we identified a substantial number of genomic regions with differential epimutations. CGI and promoter regions tended to have low epimutation rates, consistent with either high or low DNA methylation levels. We hypothesized that concordant DNA methylation in promoters enables efficient and stable transcription initiation or inhibition of hematopoietic developmental genes. Notably, TF binding sites had the highest epimutation rate. This indicates high DNA methylation heterogeneity within the binding regions of TFs, which suggests that they could facilitate the swift alteration of DNA methylation patterns with minimal methylation modifications to rapidly respond to both intracellular and extracellular signals, thus specifically regulating hematopoietic differentiation-related genes. Additionally, DNA methylation modification could have an impact on protein–DNA interactions and the binding specificity of TFs.[26] Previous studies have shown that the binding of individual TFs is directly affected by DNA methylation.[27] CpG methylation alters the binding sensitivity of TFs by modifying the local 3-dimensional of the DNA.[28] DNA methylation changes the stability of nucleosomes, which affects the local chromatin structure and accessibility of TFs to genomic DNA.[29] Therefore, the elevation of epimutation during lineage differentiation suggests intricate DNA methylation patterns are correlated to specific TF binding and specialized biological functions of the mature cells. To achieve a better mechanistic understanding of the effect of epimutation on TF binding sites, further analyses using motif databases and experiments are needed.

Furthermore, DDRs are closely associated with pathways that regulate lineage-specific differentiation. Notably, these DDRs had few overlaps with the DMRs. Because DNA methylation levels are generally high in somatic cells, especially in cells with the same differentiation system, the identification of differential DNA methylation is limited. Instead, epimutation differences may provide an alternative angle for studying mechanisms in hematopoiesis as well as other systems.

Unexpectedly, adult BM-derived HSCs had the lowest epimutation rates compared with HSCs from other developmental stages or tissues. One feature of epimutation is its ability to record the epigenetic history. Thus, the process of epimutation during the cell cycle or differentiation was enhanced. This indicates that during the migration of HSCs from the FL to the BM at the embryonic stage, HSCs undergo a transition from a disordered to an ordered state. This transition may be dependent on the microenvironment. As HSCs circulate in the blood, including CB and PB, epimutation increases.

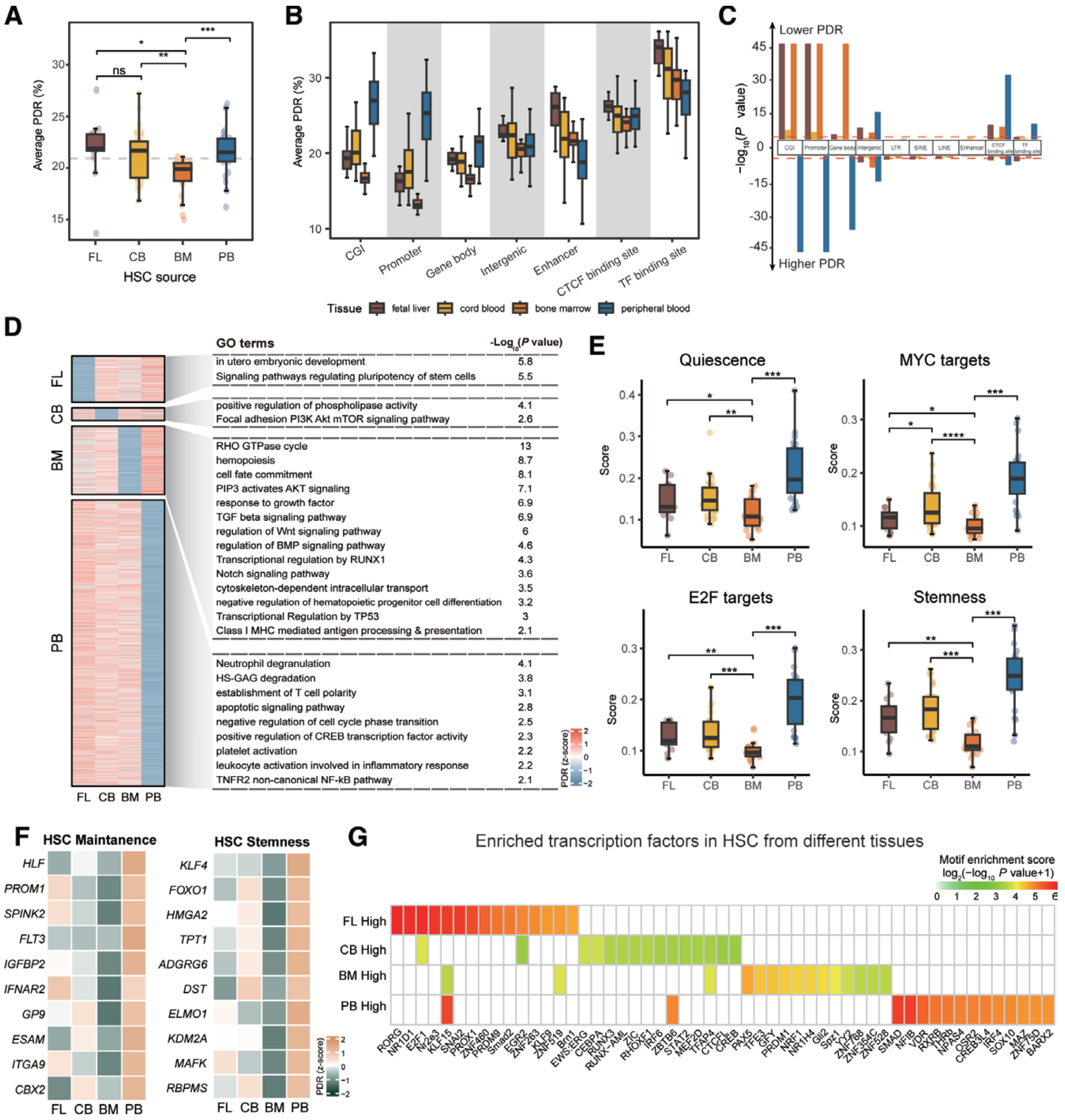

Figure 5. HSCs exhibit differential epimutation across 4 tissues. (A) Boxplots display the average PDR levels of HSCs from different tissues. The gray dashed line indicates the mean PDR level of all libraries. *P < .05; **P < .01; ***P < .001 (unpaired 2-tailed Student t test). (B) Boxplots display the PDR values across different tissues for different genomic regions. (C) Region set enrichment analysis for tissue-specific-lower DDRs (top) and tissue-specific-higher DDRs (bottom) of HSCs. The y-axis represents the negative logarithm of the P values, which is calculated by LOLA. The horizontal dashed line corresponds to a significance threshold of .05 for the P values. (D) Heatmap showing the row-scaled PDR values of tissue-specific lowest DDRs. The right column shows the GO enrichment of DDR-related genes. (E) Boxplots compare the PDR scores from each selected HALLMARK gene set across different tissues of HSCs. PDR scores are defined as the mean PDR levels of the promoters within the corresponding gene set. *P < .05, **P < .01, ***P < .001 (unpaired 2-tailed Student t test). (F) Heatmaps showing the scaled PDR values within the promoters of HSC maintenance and stemness markers across different tissues of HSCs. (G) TF motif enrichment of tissue-specific DDRs with highest PDR. The color indicates the motif enrichment scores. BM = bone marrow, CB = cord blood, CGI = CpG islands, CTCF = CCCTC-binding factor, DDRs = differentially discordant regions, FL = fetal liver, GO = gene ontology, HSC = hematopoietic stem cell, LINE = long interspersed nuclear element, LOLA = locus overlap analysis, LTR = long terminal repeat, ns = not significant, PB = peripheral blood, PDR = proportion of discordant read, SINE = short interspersed nuclear element, TF = transcription factor.

Indeed, HSCs in BM have specific epimutation features that support their functions. Moreover, malignant cells in AML show distinct and aberrant epimutations related to the dysregulation of immune signaling. Aberrant epimutations in MLL-r AML may endow blast cells with enhanced population diversity, whereas stochastic methylation alterations augment epigenetic plasticity, thereby suggesting an effective strategy for evading attacks.

Collectively, our work highlights the involvement of epimutation in hematopoietic regulation and malignancies, providing new insights into the epigenetic regulation of hematopoiesis. We anticipate that our newly developed tool, EpiMut, will pave the

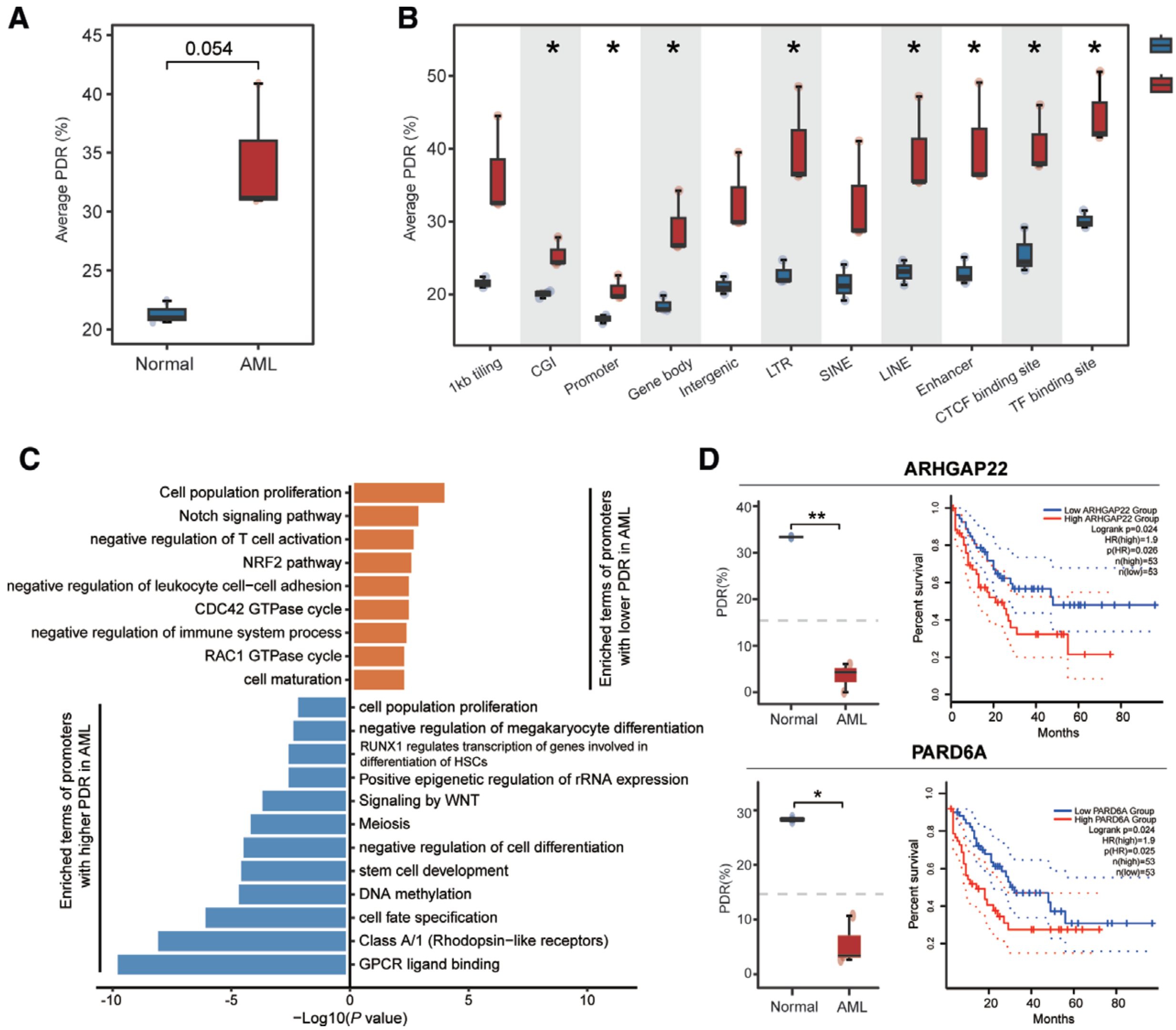

Figure 6. Epimutation associated with AML. (A) Boxplots display the average PDR levels of normal HSC (n = 3) and AML (n = 3). *P* values are shown above the box (unpaired 2-tailed Student *t* test). (B) Boxplots showing the PDR values for different genomic regions across normal HSC and AML. **P* < .05, ***P* < .01, ****P* < .001 (unpaired 2-tailed Student *t* test). (C) GO enrichment analysis of DDR-related genes between normal HSC and AML. The x-axis represents the negative logarithm of the *P* values. (D) Boxplots display the PDR levels within the promoters of *ARHGAP22* and *PARD6A* across normal HSC and AML (left). Survival analysis was performed based on the expressions of *ARHGAP22* and *PARD6A* in the TCGA-LAML patient cohort (right). **P* < .05; ***P* < .01 (unpaired 2-tailed Student *t* test). AML = acute myeloid leukemia, CGI = CpG islands, CTCF = CCCTC-binding factor, DDR = differentially discordant region, GO = gene ontology, HSC = hematopoietic stem cell, LINE = long interspersed nuclear element, LTR = long terminal repeat, PDR = proportion of discordant read, SINE = short interspersed nuclear element, TF = transcription factor.

way for the study of epimutation-based mechanisms, benefiting from its fast and accurate processing of large volumes of bisulfite sequencing data.

4. MATERIALS AND METHODS

4.1. Design of EpiMut

We used the packages named GenomicFeatures.jl, BioSequences.jl, and XAM.jl in Julia to develop EpiMut. First, the CpG methylation state was determined by comparing the sequencing reads with the reference genome. Indels (insertions or deletions) in the sequencing reads were removed or added based on the CIGAR values in the BAM. The SNP sites in the CpGs were removed to ensure accurate calculations. EpiMut can compute three key metrics of epimutation rates (PDR, MCR, and qFDRP) from BAM files aligned using tools such as BS-Seeker2. The PDR for CpGs was calculated as the number of discordant reads divided by the total number of reads. The MCR for a given CpG is equivalent to the ratio of concurrent methylation CpGs (unmethylated CpGs in partially methylated reads) to total CpGs. qFDRP was calculated using the fractions of sites that do not reflect the same methylation state in both reads of discordant read pairs divided by the total read pairs. After calculating the epimutation rates, EpiMut identified DDRs between the 2 groups using unpaired 2-tailed Student *t* tests, with *P* value adjusted using the Benjamini–Hochberg method. The visualization of PDRs was achieved by using the “cgmaptools tanghulu” command.[30] GO analysis of DDR-associated genes was conducted using an R package of clusterProfiler,[31] and the locus overlap analysis (LOLA) R package[32] was applied to the enrichment analysis of genomic regions. EpiMut software in Julia is publicly accessible at https://github.com/zhangchunyong999/EpiMut.

4.2. Evaluation of EpiMut

WSH was utilized to calculate PDR and qFDRP,[11] whereas CAMDA was employed for the computation of MCR.[10] And all these 3 metrics could be obtained by our newly developed tool, EpiMut. To compare the counts of reads used to calculate the epimutation rates between WSH[11] and EpiMut, we took 3 replicates and screened out all reads with CpGs using SAMtools[33] and XAM.jl. The read numbers calculated for epimutation rates were subsequently counted in WSH and EpiMut. Next, we applied WSH (compute_PDR.R and compute_qFDRP.R) and EpiMut (calculate PDR. jl and qFDRP. jl) to calculate the epimutation rates (PDR and qFDRP), respectively. MCR was calculated using CAMDA (CAMDA.py)[10] and EpiMut (calculateMCR.jl). To evaluate running time, we sampled reads from an alignment file in BAM format using "samtools view -s." To evaluate the running time, we took samples of reads from a BAM alignment file using "samtools view -s." Three replicates were performed for evaluating read number size of the input BAM files. The BAM files were calculated using 20 threads both in WSH (compute_PDR.R) and EpiMut (calculate PDR. jl). Each thread is assigned a CPU core. We used 1 core to calculate the MCR in CAMDA and applied it to EpiMut for comparison. Subsequently, we compared the read numbers, CpG coverage, epimutation rates, and running times of WSH, CAMDA, and EpiMut using an unpaired 2-tailed Student *t* test.

4.3. Epimutation rates for hematopoietic cells

Thirteen hematopoietic cell types from normal human PB were used for epimutation analysis based on whole-genome bisulfite sequencing datasets. Raw data were downloaded from the European Genome-phenome Archive with the accession number EGAS00001002070.[5] DNA methylation sequencing data of AML were downloaded from the NCBI GEO database with the accession number GSE135869.[34] Quality control was performed using TrimGalore (v0.4.5) (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/). The clean reads were aligned to the hg38 RefSeq reference genome (University of California, Santa Cruz Genome Browser, UCSC Genome Browser) using scBS-map (v1.0.0).[35] Subsequently, polymerase chain reaction (PCR) duplicates were removed using SAMtools (v0.1.9), and Chromosomes X, Y, and M were excluded to ensure accuracy and interpretability. Only the reads with more than four CpG sites were selected for further analysis. We used PDR values from our EpiMut software to assess the epimutation rates in hematopoietic cell analysis.

4.4. Annotation of genomic regions

The genomic features used in our analysis included whole-genome 1 kb tiling, promoters, gene bodies, intergenic regions, CpG islands, repeat elements, and regulatory regions. Promoters were defined as regions upstream 1 kb to downstream 1 kb of the TSSs. Gene bodies were defined as the regions from the TSSs to the transcription end sites. Intergenic regions were defined as stretches of DNA located between genes. The CpG islands were downloaded from the UCSC Genome Table (hg38). Repeat elements including long terminal repeats (LTRs), short interspersed nuclear elements (SINEs), and long interspersed nuclear elements (LINEs) were downloaded from the hg38 Repeat Masker. Regulatory regions, such as enhancers, CCCTC-binding factor (CTCF) binding sites, and TF binding sites, were obtained from the regulatory features in the ensemble database. The epimutation rate of a region was calculated as the average epimutation rate of all CpG sites within that region; each region was required to cover at least 5 CpG sites.

4.5. Identification of DDRs and DMRs

To estimate the epimutation value, we divided the genome into consecutive 1000 bp tiles, and only tiles with more than 5 CpG sites in the library were considered. The PDR of each tile was defined as the average PDR of all CpG sites within the tiles. We stipulated that each tile must exhibit coverage across at least 2 libraries in both groups to ensure meaningful statistical testing. Tiles with $P < .05$, unpaired 2-tailed Student *t* test, and PDR differences between the 2 groups >0.2 were defined as DDRs. For DMR analysis, the mean methylation levels of each 1000 bp tile were calculated as the average methylation values of all CpG sites within the tile. We stipulated that each tile must exhibit coverage across at least 2 libraries in both groups to ensure meaningful statistical testing. Tiles with *P* values <.05, unpaired 2-tailed Student *t* test, and differences in methylation levels between the 2 groups >0.2 were defined as DMRs.

4.6. Enrichment analysis

Genomic region enrichment was performed using LOLA software (v1.19.1)[32] with custom genomic regions described in "Annotation of genomic regions" of Methods. For genomic functional analysis of DDRs, the "bedtools intersect" command was applied to annotate the DDRs region onto the promoter to obtain the DDR-related genes, and GO enrichment analysis was performed by an online website metascape (https://metascape.org).[36] "findMotifsGenome.pl" in HOMER[37] (v4.11) was used to search for TFs binding to DDRs with the parameters "-size 2000 -S 100." Only motifs with *P* values $\leq 10^{-8}$ were retained.

4.7. Survival analysis

Overall survival data analyses were performed using Kaplan–Meier curves with the GEPIA2 webserver[38] (http://gepia2.cancer-pku.cn) based on TCGA and GTEx databases. The LAML cohort was divided into high and low expression groups for each gene using the median value (50% cutoff).

4.8. Statistics and data visualization

Student *t* test was used to test for significance between the 2 groups, and *P* values were used for significance evaluation (**P* < .05; ***P* < .01; ****P* < .001; ns, not significant).

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation of China (82022002, 82131430173, 81900117, 81890993), the National Key Research and Development Program of China (2021YFA1102800), the Distinguished Young Scholars of Tianjin (21JCJQJC00070), the CAMS Initiative for Innovative Medicine (2021-I2M-1-040), and the Foundation of Haihe Laboratory (HH22KYZX0018).

AUTHOR CONTRIBUTIONS

P.Z.., P.W., and L.S. conceived the study. J.L., P.W., C.Z., and X.Q. carried out the bioinformatics analysis and developed the package. X.Q., J.L., and P.Z. drafted the manuscript. All authors read and approved the final manuscript.

REFERENCES

[1] Till JE, McCulloch EA. Hemopoietic stem cell differentiation. *Biochim Biophys Acta* 1980;605(4):431–459.

[2] Laurenti E, Gottgens B. From haematopoietic stem cells to complex differentiation landscapes. *Nature* 2018;553(7689):418–426.

[3] Soto-Palma C, Niedernhofer LJ, Faulk CD, Dong X. Epigenetics, DNA damage, and aging. *J Clin Invest* 2022;132(16):e158446.

[4] Cabezas-Wallscheid N, Klimmeck D, Hansson J, et al. Identification of regulatory networks in HSCs and their immediate progeny via integrated proteome, transcriptome, and DNA methylome analysis. *Cell Stem Cell* 2014;15(4):507–522.

[5] Farlik M, Halbritter F, Muller F, et al. DNA methylation dynamics of human hematopoietic stem cell differentiation. *Cell Stem Cell* 2016;19(6):808–822.

[6] Laranjeira ABA, Hollingshead MG, Nguyen D, Kinders RJ, Doroshow JH, Yang SX. DNA damage, demethylation and anticancer activity of DNA methyltransferase (DNMT) inhibitors. *Sci Rep* 2023;13(1):5964.

[7] Onuchic V, Lurie E, Carrero I, et al. Allele-specific epigenome maps reveal sequence-dependent stochastic switching at regulatory loci. *Science* 2018;361(6409):eaar3146.

[8] Horvath S. DNA methylation age of human tissues and cell types. *Genome Biol* 2013;14(10):R115.

[9] Gaiti F, Chaligne R, Gu H, et al. Epigenetic evolution and lineage histories of chronic lymphocytic leukaemia. *Nature* 2019;569(7757):576–580.

[10] Shi J, Xu J, Chen YE, et al. The concurrence of DNA methylation and demethylation is associated with transcription regulation. *Nat Commun* 2021;12(1):5285.

[11] Scherer M, Nebel A, Franke A, et al. Quantitative comparison of within-sample heterogeneity scores for DNA methylation data. *Nucleic Acids Res* 2020;48(8):e46.

[12] Huo Y, Wu L, Pang A, et al. Single-cell dissection of human hematopoietic reconstitution after allogeneic hematopoietic stem cell transplantation. *Sci Immunol* 2023;8(81):eabn6429.

[13] Kaiser FMP, Janowska I, Menafra R, et al. IL-7 receptor signaling drives human B-cell progenitor differentiation and expansion. *Blood* 2023;142(13):1113–1130.

[14] Monticelli S, Natoli G. Transcriptional determination and functional specificity of myeloid cells: making sense of diversity. *Nat Rev Immunol* 2017;17(10):595–607.

[15] Mazzurana L, Czarnewski P, Jonsson V, et al. Tissue-specific transcriptional imprinting and heterogeneity in human innate lymphoid cells revealed by full-length single-cell RNA-sequencing. *Cell Res* 2021;31(5):554–568.

[16] Karamitros D, Stoilova B, Aboukhalil Z, et al. Single-cell analysis reveals the continuum of human lympho-myeloid progenitor cells. *Nat Immunol* 2018;19(1):85–97.

[17] Warsi S, Blank U, Dahl M, et al. BMP signaling is required for postnatal murine hematopoietic stem cell self-renewal. *Haematologica* 2021;106(8):2203–2214.

[18] Wu F, Chen Z, Liu J, Hou Y. The Akt-mTOR network at the interface of hematopoietic stem cell homeostasis. *Exp Hematol* 2021;103:15–23.

[19] Duncan AW, Rattis FM, DiMascio LN, et al. Integration of Notch and Wnt signaling in hematopoietic stem cell maintenance. *Nat Immunol* 2005;6(3):314–322.

[20] de Bruijn M, Dzierzak E. Runx transcription factors in the development and function of the definitive hematopoietic system. *Blood* 2017;129(15):2061–2069.

[21] Chasan S, Hesse E, Atallah P, et al. Sulfation of glycosaminoglycan hydrogels instructs cell fate and chondral versus endochondral lineage decision of skeletal stem cells in vivo. *Adv Funct Mater* 2021;32(7):2109176.

[22] El-Masry OS, Alamri AM, Alzahrani F, Alsamman K. ADAMTS14, ARHGAP22, and EPDR1 as potential novel targets in acute myeloid leukaemia. *Heliyon* 2022;8(3):e09065.

[23] Lu Z, Yuan S, Ruan L, Tu Z, Liu H. Partitioning defective 6 homolog alpha (PARD6A) promotes epithelial-mesenchymal transition via integrin beta1-ILK-SNAIL1 pathway in ovarian cancer. *Cell Death Dis* 2022;13(4):304.

[24] Mori M, Saito K, Ohta Y. ARHGAP22 localizes at endosomes and regulates actin cytoskeleton. *PLoS One* 2014;9(6):e100271.

[25] Bock C, Beerman I, Lien WH, et al. DNA methylation dynamics during in vivo differentiation of blood and skin stem cells. *Mol Cell* 2012;47(4):633–647.

[26] Dantas Machado AC, Zhou T, Rao S, et al. Evolving insights on how cytosine methylation affects protein-DNA binding. *Brief Funct Genomics* 2014;14(1):61–73.

[27] Becker PB, Ruppert S, Schütz G. Genomic footprinting reveals cell type-specific DNA binding of ubiquitous factors. *Cell* 1987;51(3):435–443.

[28] Lazarovici A, Zhou T, Shafer A, et al. Probing DNA shape and methylation state on a genomic scale with DNase I. *Proc Natl Acad Sci USA* 2013;110(16):6376–6381.

[29] Briggs JM, Portella G, Battistini F, Orozco M. Understanding the connection between epigenetic DNA methylation and nucleosome positioning from computer simulations. *PLoS Comput Biol* 2013;9(11):e1003354.

[30] Guo W, Zhu P, Pellegrini M, Zhang MQ, Wang X, Ni Z. CGmapTools improves the precision of heterozygous SNV calls and supports allele-specific methylation detection and visualization in bisulfite-sequencing data. *Bioinformatics* 2018;34(3):381–387.

[31] Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* 2012;16(5):284–287.

[32] Sheffield NC, Bock C. LOLA: enrichment analysis for genomic region sets and regulatory elements in R and Bioconductor. *Bioinformatics* 2016;32(4):587–589.

[33] Li H, Handsaker B, Wysoker A, et al. The sequence alignment/map format and SAMtools. *Bioinformatics* 2009;25(16):2078–2079.

[34] Koldobskiy MA, Abante J, Jenkinson G, et al. A dysregulated DNA methylation landscape linked to gene expression in MLL-rearranged AML. *Epigenetics* 2020;15(8):841–858.

[35] Wu P, Gao Y, Guo W, Zhu P. Using local alignment to enhance single-cell bisulfite sequencing data efficiency. *Bioinformatics* 2019;35(18):3273–3278.

[36] Zhou Y, Zhou B, Pache L, et al. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat Commun* 2019;10(1):1523.

[37] 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.

[38] Tang Z, Kang B, Li C, Chen T, Zhang Z. GEPIA2: an enhanced web server for large-scale expression profiling and interactive analysis. *Nucleic Acids Res* 2019;47(W1):W556–W560.