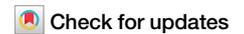


<https://doi.org/10.1038/s42003-025-08257-8>

# Discovery of candidate functional non-coding mutations in acute myeloid leukemia using single-cell chromatin accessibility sequencing



Ming Zhu<sup>1,6</sup>, Jiali Zhu<sup>1,6</sup>, Zhijuan Zhu<sup>2,6</sup>, Yiding Yang<sup>1,3</sup>, Jinxian Dai<sup>1</sup>, Hua Li<sup>2,4</sup>, Nainong Li<sup>2,5</sup>✉ & Jialiang Huang<sup>1,3</sup>✉

Mutations and gene rearrangements are crucial for the diagnosis and subtyping of acute myeloid leukemia (AML). However, the contribution of non-coding genetic variants, particularly those within cis-regulatory elements (CREs), to AML pathophysiology and heterogeneity remains poorly understood. In this study, we characterize the single-cell chromatin accessibility landscapes of 10 bone marrow samples from AML patients at diagnosis. Additionally, we develop eMut, an integrated computational pipeline for detecting, imputing, and functionally characterizing non-coding mutations in CREs at the single-cell level. Our analysis identifies 2878 potential somatic non-coding mutations, highlighting the extensive mutational heterogeneity in the non-coding genome of AML patients, with recurrent non-coding mutations displaying cell type-specific patterns. We show that mutated CREs are enriched with blood-related genetic variants, potentially linked to AML-associated genes, and harbor a higher abundance of functional CREs, suggesting their functional relevance in leukemogenesis. Importantly, we pinpoint candidate functional non-coding mutations that associate with alteration of target gene expression in AML. Collectively, our work provides a comprehensive resource of single-cell chromatin accessibility in AML and introduces an integrative approach to identify candidate functional non-coding mutations contributing to cellular heterogeneity in AML.

Acute myeloid leukemia (AML) is a heterogeneous hematologic malignancy primarily driven by gene mutations, gene rearrangements, and subsequent acquired cooperative mutations<sup>1,2</sup>. AML is characterized by over 20 frequently mutated genes, which can be broadly categorized into epigenetic regulators (e.g., *DNMT3A*, *ASXL1*, *TET2*), splicing factors (e.g., *SF3B1*, *SRSF2*), signaling transduction genes (e.g., *FLT3*, *KIT*, *KRAS*), tumor suppressors (e.g., *TP53*, *WT1*), and transcription factors (e.g., *RUNX1*, *MYC*)<sup>3,4</sup>. Identifying these mutations in AML patients is critical for guiding treatment options and prognostic stratification. However, some cases lack detectable coding mutations<sup>5</sup>, and patients within the same molecular subtype often exhibit significant clinical heterogeneity<sup>6</sup>. This suggests the involvement of

unidentified non-coding mutations in AML pathogenesis. For instance, deletions in lineage-specific enhancer clusters can disrupt the spatio-temporal expression of *MYC*, contributing to AML development<sup>7</sup>. Functional screening of enhancers with high-frequency mutations in AML has further revealed that such mutations can lead to abnormal expression of genes like *KRAS* and *PER2*<sup>8</sup>. Thus, detecting and interpreting non-coding mutations is essential for understanding AML heterogeneity.

The majority of disease-associated mutations are located in non-coding regions, particularly in cis-regulatory elements (CREs)<sup>9</sup>. Enhancers, a type of transcriptionally activated CRE, play a pivotal role in gene expression regulation and are typically found in open chromatin regions

<sup>1</sup>State Key Laboratory of Cellular Stress Biology, Xiang'an Hospital, School of Life Sciences, Faculty of Medicine and Life Sciences, Xiamen University, Xiamen, Fujian, China. <sup>2</sup>Hematopoietic Stem Cell Transplantation Center, Fujian Institute of Hematology, Fujian Provincial Key Laboratory on Hematology, Department of Hematology, Fujian Medical University Union Hospital, No.29 Xinquan Street, Gulou District, Fuzhou, China. <sup>3</sup>National Institute for Data Science in Health and Medicine, Xiamen University, Xiamen, Fujian, China. <sup>4</sup>Department of Hematology and Rheumatology, the Second Affiliated Hospital of Xiamen Medical College, No.566 Shengguang Road, Jimei District, Xiamen, China. <sup>5</sup>Translational Medicine Center on Hematology, Fujian Medical University, Fuzhou, China. <sup>6</sup>These authors contributed equally: Ming Zhu, Jiali Zhu, Zhijuan Zhu. ✉e-mail: [nainli@aliyun.com](mailto:nainli@aliyun.com); [jhuang@xmu.edu.cn](mailto:jhuang@xmu.edu.cn)

marked by specific histone modifications<sup>10</sup>. These elements contain transcription factor binding sites that recruit co-factors to initiate gene expression. Enhancers can function independently or synergistically as clusters, creating functional redundancy and complicating the study of individual enhancers<sup>11</sup>. Additionally, enhancers exhibit cell-type-specific activity, controlling the spatiotemporal expression of genes<sup>12</sup>. Mutations in the transcription factor binding sites of enhancers can disrupt this regulation, contributing to disease<sup>7</sup>. However, identifying functional non-coding mutations and elucidating their mechanisms remains a significant challenge.

Single-cell chromatin accessibility sequencing (scATAC-seq) has been employed to study epigenetic regulation in AML pathogenesis and relapse<sup>13–17</sup>, but has rarely been used to predict functional non-coding mutations. Unlike bulk sequencing, scATAC-seq enables the identification of non-coding mutations at single-cell resolution<sup>18,19</sup> and links them to cellular traits such as epigenetic regulation and cell-type abundance<sup>20–22</sup>. Although methods exist to detect non-coding mutations using single-cell chromatin accessibility data<sup>18,19,23</sup>, the functional consequences and significance of these mutations remain poorly characterized.

In this study, we performed multi-omics sequencing, including scATAC-seq, ATAC-seq, RNA-seq, and H3K27ac CUT&Tag, on bone marrow samples from 10 AML patients. We revealed cellular composition heterogeneity in AML and developed an integrated pipeline based on scATAC-seq data to detect non-coding mutations and characterize their functional consequences. We identified sample-specific and lineage-specific non-coding mutations in AML and elucidated the functional importance of the regulatory elements they reside in, as well as their potential impact on transcriptional regulation.

## Results

### Single-cell and bulk multi-omics profiling of bone marrow in AML

To better resolve the heterogeneity of AML and enable subsequent validation, we performed scATAC-seq, ATAC-seq, H3K27ac CUT&Tag, and RNA-seq on bone marrow samples from 10 AML patients (AML1 to AML10) at the time of diagnosis (Fig. 1a, Supplementary Data 1). For the scATAC-seq experiment, we incorporated modifications inspired by Paired-seq<sup>24</sup> methodology, while preserving the core ATAC sequencing component (“Methods” section). Each sample was uniquely labeled using Tn5 barcodes, allowing simultaneous processing of multiple samples with a low doublet rate (Supplementary Fig. 1a–c). We observed that the fraction of fragments in peaks (Frip) and the fraction of fragments in transcription start sites (FriTSS) were lower in the NIH3T3 cell line compared to SHARE-seq (Supplementary Fig. 1d–f), which may be caused by the lower sequencing depth. We performed scATAC-seq on 10 bone marrow samples, obtaining 69,088 high-quality cells with 1239 median unique fragments per cell (Supplementary Fig. 1g) and identifying 8 distinct clusters (Fig. 1b). Based on marker gene activity derived from scATAC-seq (Fig. 1c, Supplementary Fig. 1h) and gene expression measured by RNA-seq (Supplementary Fig. 1i), we annotated major malignant cell types, including hematopoietic stem cell-like (HSC-like, *CD34*), immature myeloid progenitor-like (IMP-like, *MPO* and *PF4*), megakaryocytic progenitor-like (Mkp-like, *PF4*), granule-monocytic progenitor-like (GMP-like, *MPO*), monocyte-like (Mono-like, *CD14*, *CEBPD*), and classical dendritic cell-like (cDC-like, *CD14*, *CLEC10A*). Each tumor sample exhibited a dominant cell cluster with the expansion of tumor cells (Fig. 1d). Furthermore, the expression of marker genes in the annotated cell types aligned with RNA-seq results (Supplementary Fig. 1i), and the promoters of these marker genes showed varying degrees of accessibility across cell types (Fig. 1e). Additionally, these scATAC clusters were enriched for lineage-determining transcription factors (TFs) (Fig. 1f), such as RUNX1 and TAL1 in hematopoietic stem/progenitor cells (HSPC) clusters, CEBPA and SPI1 (PU.1) in myeloid lineage, and GATA2 and MYC in erythroid progenitor cells<sup>25,26</sup>.

### eMut detects and interprets single-cell non-coding mutations based on scATAC-seq data

The current mutation detection pipelines using single-cell ATAC-seq (scATAC) data can be categorized into three distinct groups based on their specific applications (Supplementary Data 2): (i) Monopogen<sup>19</sup> and SComatic<sup>23</sup>, which are specifically designed for identifying putative somatic single-nucleotide variants (SNVs) in single-cell data without requiring matched normal samples; (ii) Alleloscope<sup>27</sup>, Copy-scAT<sup>28</sup>, epiAneufinder<sup>29</sup>, and AtaCNA<sup>30</sup>, which specialize in detecting copy number variants (CNVs) from scATAC data; and (iii) popular sequencing mutation detection tools, such as GATK<sup>31</sup>, SAMtools<sup>32</sup>, Strelka2<sup>33</sup>, VarScan2<sup>34</sup>, and VarCA<sup>18</sup>—an ensemble classifier integrating multiple variant callers—that have been adapted for single-cell data analysis. However, identifying functional non-coding mutations that lead to dysregulation of target genes and contribute to disease pathogenesis remains a significant challenge.

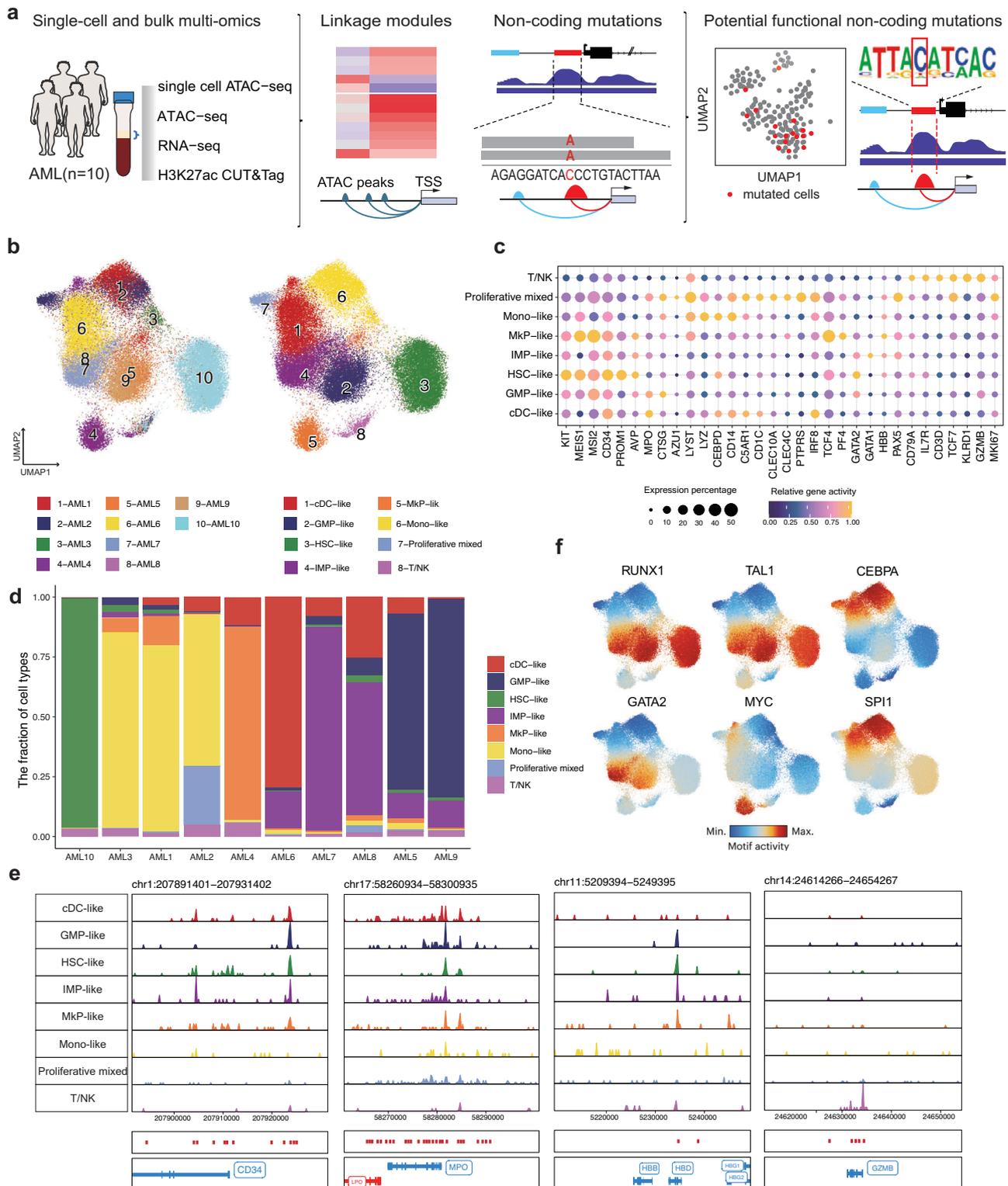
To address this, we developed eMut, an integrative pipeline designed with mutation detection and functional interpretation modules to systematically detect non-coding mutations using single-cell chromatin accessibility data and to characterize their potential functional impacts on gene regulatory programs and cellular lineages in AML (Fig. 2, see “Methods” section for details). First, leveraging reads from open chromatin regions, eMut identifies mutations in individual cells using tools such as Monopogen<sup>19</sup> or GATK Mutect2<sup>35</sup>, irrespective of whether tumor samples are paired with normal samples. Given the sparsity of scATAC-seq data, we further imputed candidate mutated cells by network propagation<sup>36</sup>, utilizing mutated cells (seed cells) within a cell-cell similarity graph (Fig. 2a). Additionally, eMut incorporates four modules for functional interpretation of detected non-coding mutations (Fig. 2b): (1) identifying cell-type- or lineage-specific mutations and assessing their potential roles in cell fate determination; (2) detecting hypermutated CREs with a significant excess of mutations to pinpoint potentially critical enhancers; (3) predicting the effects of mutations on transcription factor motifs (loss or gain) and linking mutated enhancers to target genes to identify functionally impactful non-coding mutations; (4) comparing target gene expression changes between mutated and wild-type cells (or samples). Collectively, eMut represents an integrative pipeline for detecting, imputing, and characterizing non-coding regulatory mutations with functional consequences at the single-cell resolution.

### Both coding and non-coding mutations exhibit heterogeneity among AML patients

To characterize the genomic landscape of AML, we analyzed known coding genetic variants in 10 AML patients using diagnostic genetic testing, including cytogenetic and molecular genetic analyses (Fig. 3a, Supplementary Data 1). Consistent with previous studies<sup>37</sup>, coding mutations were detected in FLT3-ITD rearrangement ( $n=3$ ), and other AML-associated genes, such as *NPM1*, *DNMT3A*, *CEBPA*, *IDH2*, *NRAS*, and *WT1*. These mutations displayed distinct combinatorial patterns, unique to each patient (Fig. 3a). For non-coding mutations, we employed eMut to identify mutations in accessible chromatin regions using scATAC-seq data from these patients (without paired normal samples) (Fig. 3b). In total, we identified 2878 potential somatic non-coding mutations (median 317 per patient), most of which were patient-specific, underscoring the heterogeneity of non-coding mutations in AML (Fig. 3b, Supplementary Data 2). Among these, 1930 (67%) were supported by bulk ATAC-seq data, and 1569 (54.5%) were validated by H3K27ac CUT&Tag data (Supplementary Fig. 2). In addition, we found a positive correlation between the fraction of cells with somatic non-coding mutations and the percentage of AML blast cells (Pearson  $R=0.57$ ,  $p=0.053$ ; Fig. 3c and Supplementary Data 1). These non-coding mutations were mainly concentrated in the 5' flanking regions, introns, and intergenic regions (Fig. 3d). To explore mutagenic mechanisms, we analyzed mutational spectra and signatures<sup>38</sup>. The non-coding mutation spectrum derived from scATAC-seq data showed high concordance with whole genome

sequencing (WGS) from public AML datasets<sup>39</sup> (cosine similarity 0.82), comparable to bulk ATAC-seq predictions (Fig. 3e, “Methods” section). By fitting COSMIC signatures to the observed mutational spectra, we identified three COSMIC single base substitution (SBS)

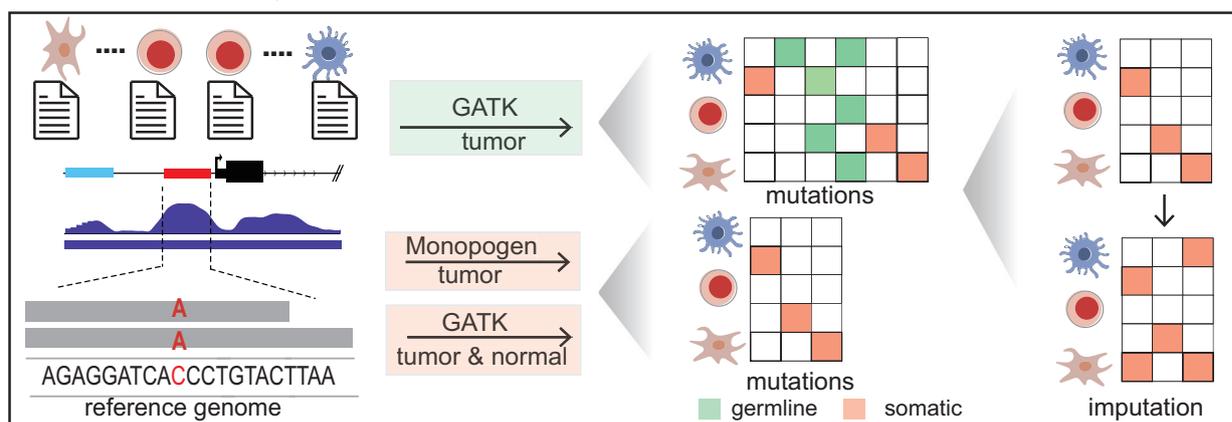
signatures: a myeloid signature, SBS5 (a clock-like signature), and SBS6 (DNA mismatch repair) (Fig. 3f). These findings suggest both non-coding and coding mutations exhibit heterogeneity among AML patients but share underlying mutational processes.



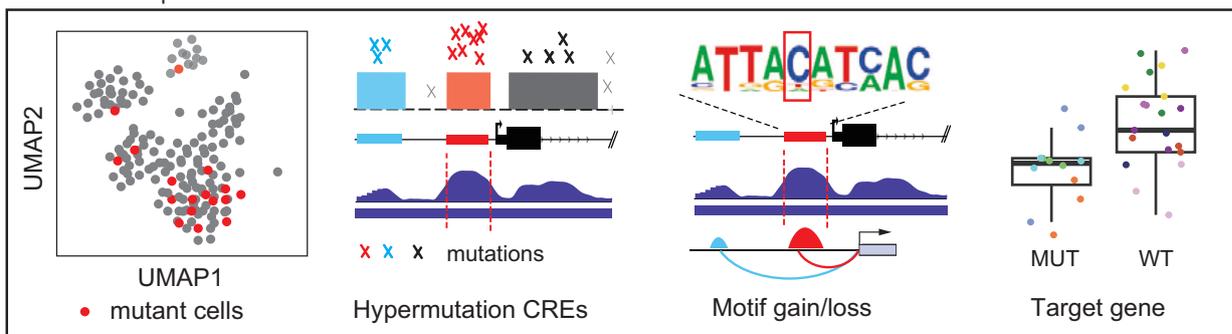
**Fig. 1 | Single-cell and bulk multi-omics profiling of bone marrow samples from AML patients at diagnosis.** **a** Overview of study design, sample collection, single-cell profiling, and analyses. **b** UMAP visualization for scATAC-seq across samples (left) and major hematopoietic cell types (right). **c** scATAC-seq gene activity scores for the representative marker genes across scATAC-seq clusters. Color indicates

relative gene activity across all clusters, and dot size indicates the percentage of cells showing gene activity in each cluster. **d** Fraction of cells in each scATAC-seq cluster across samples. **e** Genome tracks of the accessibility near the marker genes. **f** UMAP for the representative motif activity score identified by ChromVAR across major hematopoietic cell types.

**a** Mutation detection and imputation



**b** Functional interpretation



**Fig. 2 | Overview of eMut workflow to detect and interpret non-coding mutations at the single-cell level. a** Mutation detection and imputation. eMut splits the alignment file (BAM-formatted) according to cellular barcode and detects mutations in open regions of the genome by matching sequence information to the reference

genome. For each mutation, the mutated cells were imputed by network propagation using the SCAVENGE approach. **b** Functional interpretation of mutations at different levels, including for cell-type enrichment, hypermutated CREs identification, impact on transcription factor binding event, and target gene expression.

**Recurrent non-coding mutations across AML patients are cell-type-specific**

AML is a disease characterized by abnormal hematopoietic differentiation and proliferation<sup>40</sup>. To investigate the potential role of non-coding mutations, particularly recurrent ones, in cell fate determination, we identified 52 recurrent non-coding mutations present in at least 3 patients (Fig. 4a). Cell-type enrichment analysis revealed that these recurrent non-coding mutations are cell-type-specific across AML patients (Fig. 4b, “Methods” section). For example, the mutation chr17:8156823:C:A, located on the enhancer of *PER1*, was identified in 6 AML samples, and showing significant enrichment in proliferating cells across various cell types (Fig. 4c). Due to the sparsity of single-cell data, we performed imputation on mutated cells to infer additional potential mutated cells (“Methods” section). *Per1*, which is involved in circadian rhythm transcription-translation feedback loops, has been implicated in leukemia stem cell function<sup>41</sup>. Another mutation, chr8:38030499:C:G, located on the promoter of *EIF4EBP1*, a gene essential for sustaining AML cells proliferation and survival<sup>42</sup>, was enriched in the HSC-like cluster (Fig. 4d). Additionally, we identified a mutation chr16:74666845:G:C, enriched in the Mono-like lineage and located in the promoter of *RFWD3* (Fig. 4e), as well as a mutation chr20:62832771:G:A, enriched in the GMP-like lineage and situated in the enhancer of *DPH3P1* (Fig. 4f). Collectively, these findings suggest that cell-type-specific recurrent non-coding mutations may play a role in determining cell fate in AML.

**Cis-regulatory elements with mutations implicated in leukemogenesis**

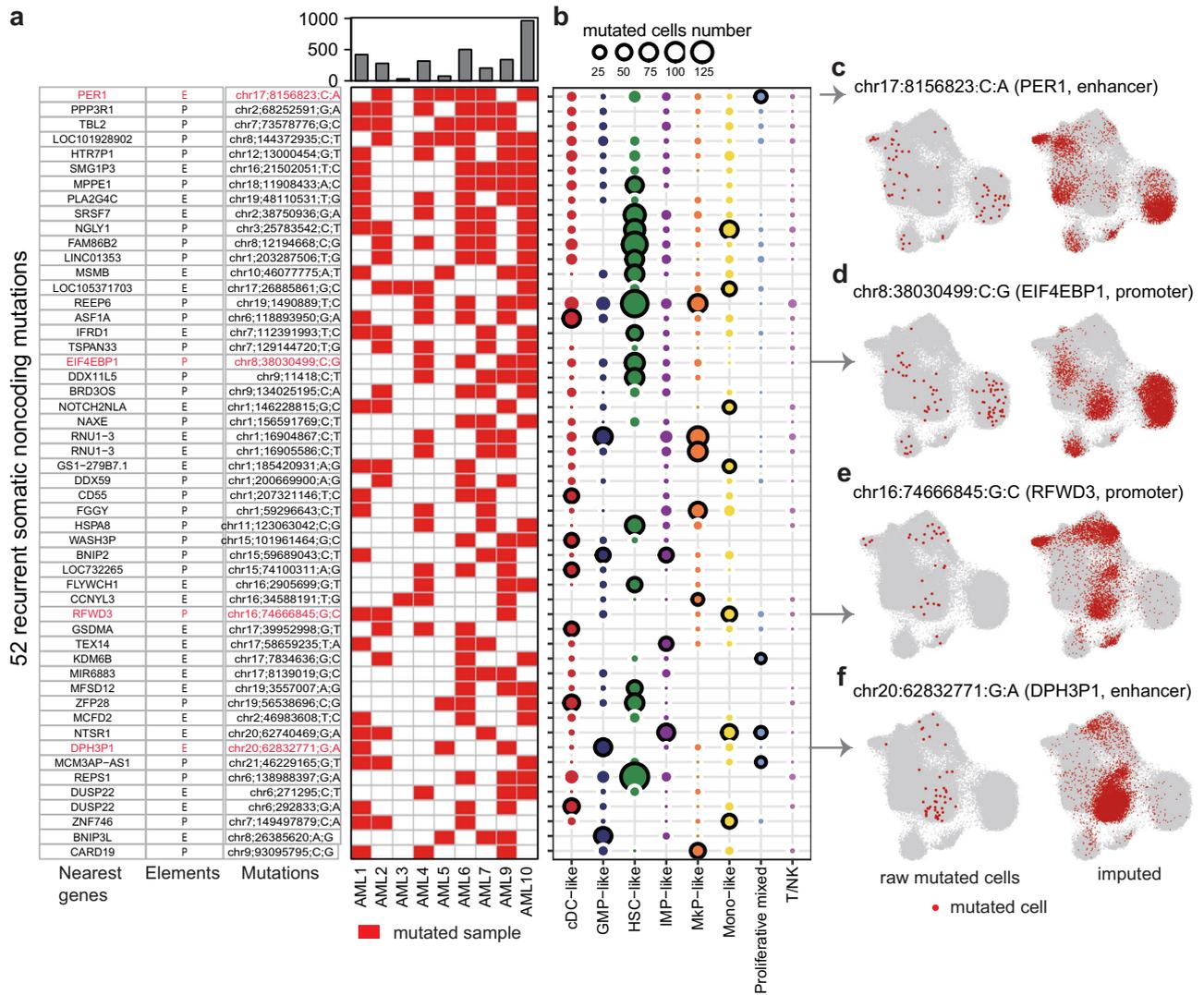
To investigate the hypothesis that genetic variants may drive enhancer dysregulation and contribute to AML, we mapped the identified putative somatic mutations to accessible cis-regulatory elements (CREs), identifying

2542 mutated CREs (Supplementary Data 3). Target genes linked to these mutated CREs showed significantly higher enrichment of AML-related genes and cancer-driver genes compared to CREs without somatic mutations (median enrichment of 5.7-fold, Fig. 5a). Functional enrichment analysis<sup>43</sup> revealed that these mutated CREs were associated with apoptosis and hematopoietic cell fate determination, including the regulation of lymphocyte differentiation (Fig. 5b).

In addition to somatic mutations, germline mutations have also been implicated in leukemogenesis<sup>44,45</sup>. Therefore, we analyzed germline mutations in AML (Supplementary Fig. 3). Given the substantial number of germline mutations, we adapted our approach to identify hypermutated CREs (defined by an observed excess expected somatic/germline mutations, see “Methods” section for details) rather than simply mutated CREs (with or without somatic mutations). Similarly, target genes of hypermutated CREs were enriched for AML-related genes and cancer-driver genes (Fig. 5c). We further assessed the prevalence of GWAS SNPs<sup>46</sup> and cis-eQTL<sup>47</sup> within these hypermutated CREs, observing a 1.6-fold enrichment of SNPs associated with blood-related traits and a 1.4-fold enrichment of cis-eQTL (Fig. 5d, e).

To validate the mutations detected by eMut could help screen for functional CREs in leukemogenesis, we collected experimentally discovered functional CREs in human leukemia<sup>8</sup>. For comparison, we also applied another mutation detection method, SComatic, to identify putative somatic mutations (Supplementary Fig. 4 and Data 2). Compared to SComatic, eMut demonstrated a superior balance between the number and proportion of functional CREs enriched among mutated CREs (Fig. 5f–h). For the mutation detection pipelines Monopogen and GATK utilized in eMut, the additional functional interpretation modules—such as the perturbed TF binding sites module for Monopogen and the hypermutated CREs module





**Fig. 4 | Recurrent cell-type-specific non-coding mutations in AML.** **a** Heatmap shows the 52 recurrent non-coding mutation profiles across patients (ranked by the number of patients with individual mutation). The nearest neighbor genes, annotation (E, enhancer; P, promoter), and chromosome coordinate for each mutation are shown on the left of the heatmap. The cumulative numbers of mutations for each sample are shown at the top of the heatmap. **b** The enrichment of mutations across

various cell types. The dot size represents the number of mutated cells, the outer circle represents the enrichment significance using the hypergeometric test. **c–f** The distribution of raw mutated cells (left) and imputed mutated cells (right) with mutations highlighted in **(a)**, chr17:8156823:C:A (*PER1*, enhancer) **(c)**, chr8:38030499:C:G (*EIF4EBP1*, promoter) **(d)**, chr16:74666845:G:C (*RFWD3*, promoter) **(e)**, and chr20:62832771:G:A (*DPH3P1*, enhancer) **(f)**.

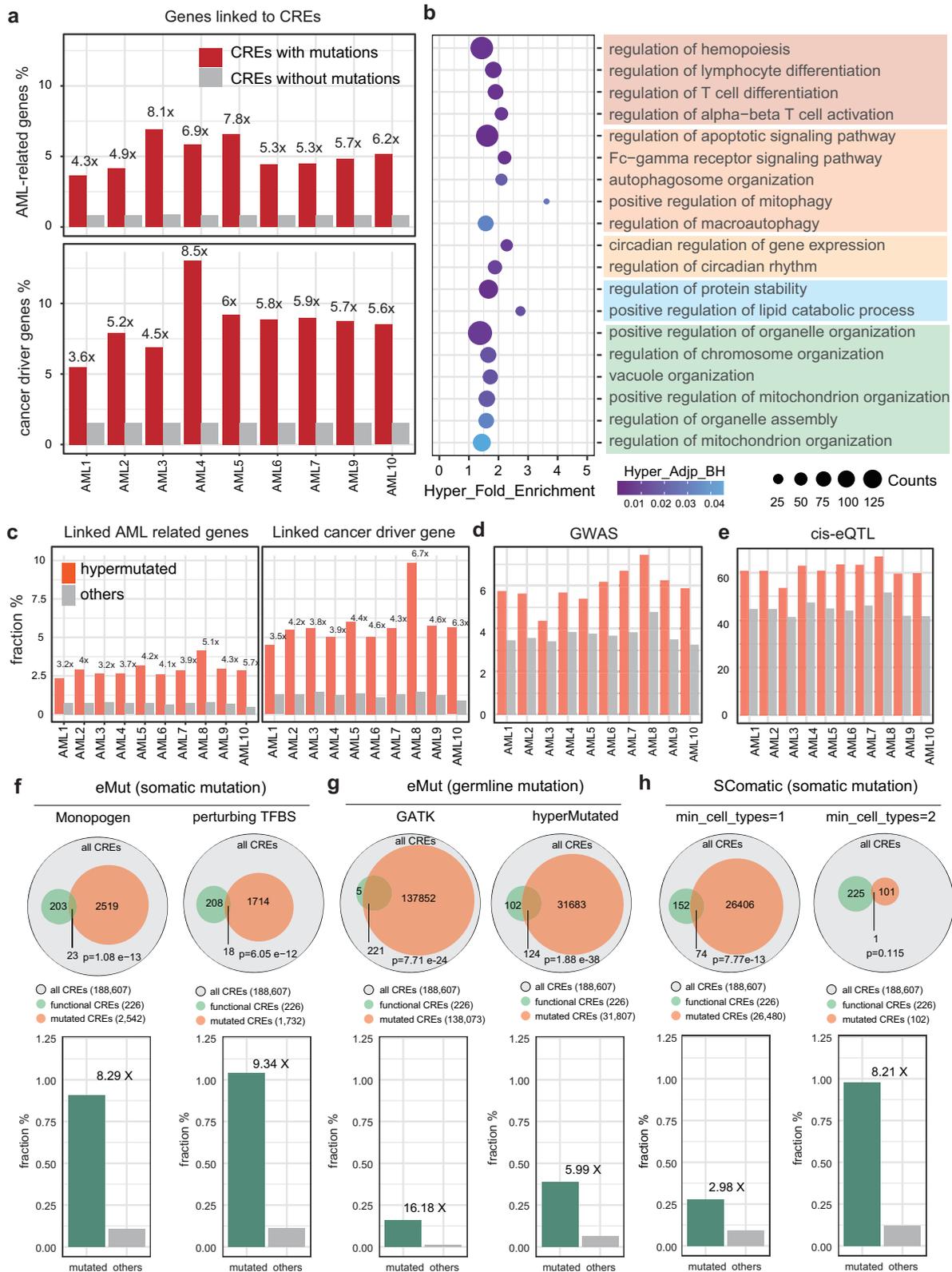
for GATK—enhanced the screening efficiency of functional CREs (Fig. 5f, g). Together, these findings suggest that CREs harboring genetic variants detected by eMut may play a critical role in regulating leukemia-related genes and contributing to leukemogenesis.

**Candidate functional non-coding mutations in AML**

Enhancers serve as platforms for binding lineage-regulating TFs, which control gene expression<sup>13,21</sup>. To further investigate the impact of non-coding mutations on transcriptional regulation, we assessed alterations of transcription factor binding motifs caused by these mutations. We found that the majority (83%) of somatic mutations significantly altered TF binding motifs, resulting in either loss or gain of binding sites for prominent factors such as EGR1 and lineage-determining TFs like SPI1 (Fig. 6a). Given the sparsity of single-cell data, we examined the impact of non-coding mutations on target gene expression by comparing mutated samples with wild-type samples for each mutation. Notably, 334 non-coding mutations were found to affect downstream target genes, with 208 mutations associated with upregulated expression and 126 mutations with downregulated expression, respectively (Fig. 6b). These findings suggested that non-coding mutations

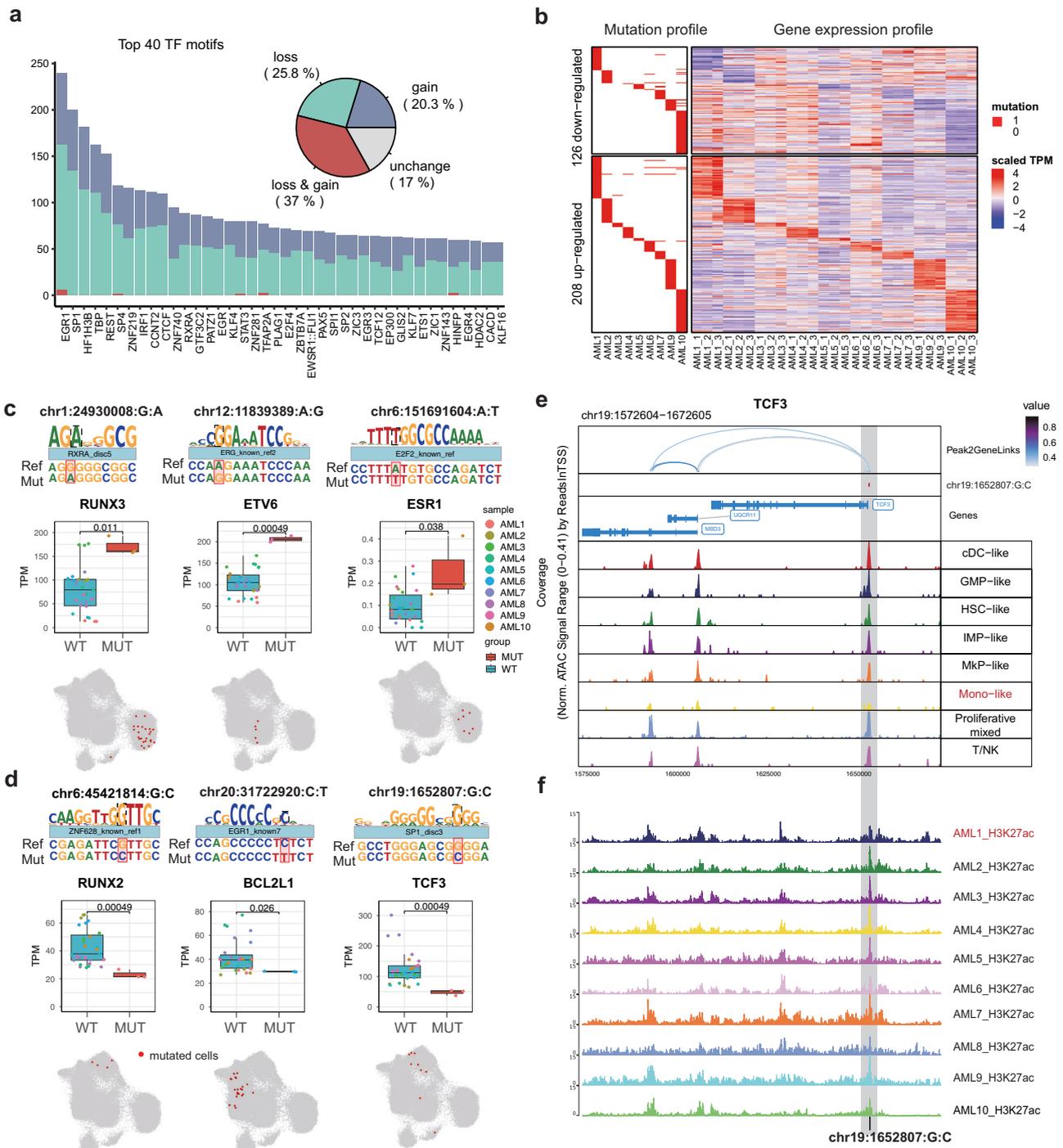
can induce transcriptional dysregulation to a significant extent. Furthermore, these transcriptionally dysregulated genes were significantly enriched for AML survival-related genes<sup>18</sup> (Supplementary Fig. 5a), offering a potential basis for developing diagnostic biomarkers for risk stratification.

Therefore, from the perspective of transcriptional dysregulation, we identified potential functional mutations that could directly alter TF binding and induce gene expression dysregulation (Supplementary Data 4). For instance, certain mutations were predicted to create a potential TF motif for transcriptional activators, which might associate with alterations in gene expression. For example, in HSC-like cells, gains in RXRA and E2F2 motifs were associated with increased expression of *RUNX3* and *ESR1*, respectively, whereas ERG motif gain correlated with elevated *ETV6* levels in GMP-like cells (Fig. 6c, Supplementary Fig. 6). In contrast, the disruption of ZNF628 motifs coincided with reduced *RUNX2* expression in mono-like cells, and the loss of EGR1 motifs was accompanied by downregulation of *BCL2L1* in cDC-like cells (Fig. 6d, Supplementary Fig. 6). Similarly, the loss of SPI1 motifs was primarily associated with reduced *TCF3* expression in mono-like cells, alongside decreased chromatin accessibility and diminished H3K27ac signals (Fig. 6d–f, Supplementary Fig. 6). Notably, these functional somatic



**Fig. 5 | Characterization and validation of mutated or hypermutated CREs.** **a** The observed fraction of AML-related genes and cancer-driver genes (COSMIC) within the genes linked to CREs with (red) or without (gray) somatic mutations is shown. **b** The selected top functional enrichment analysis results (via rGREAT) of CREs with somatic mutations. **c** The observed fraction of AML-related genes and cancer-driver genes within the genes linked to hypermutated CREs is shown in orange, with gray representing the remaining CREs. **d, e** Percentage of mutations annotated with blood-related GWAS SNPs (**d**) and cis-eQTL (**e**). **f-h** Functional validation of

mutated CREs identified through the somatic mutation mode (**f**), germline mode (**g**), using eMut and by SComatic (**h**). Venn diagram representing the overlap of all CREs, mutated CREs, and functional CREs (overlapped with CREs reported by previous study) (top panel), along with a bar plot illustrating the percentage of mutated CREs confirmed by experimental validation compared to the rest of the CREs (bottom panel). The *p*-value was computed using the hypergeometric distribution (lower.tail = F).



**Fig. 6 | Candidate functional non-coding mutations in AML.** **a** The top 40 transcription factors ranked based on the number of mutations that could potentially affect their binding motifs. Pie chart showing the proportion of mutations affecting transcription factor binding (gain or loss). **b** The left heatmap displays the mutation profiles across patients, while the right heatmap shows the gene expression of the linked target genes for each mutation across samples by bulk RNA-seq (mutated patients vs wild-type, with three replicates per patient). **c, d** Genomic visualization of six mutations. In each example, the second row represents the reference sequence, the bottom row represents the mutated sequence (with the alt-allele base), and the

top row represents the gain (c) or loss (d) motif by the mutation (top panel). The mutated position is highlighted by the red box. The boxplot shows target gene expression between mutated and wild-type samples (middle panel). The UMAP plots show the distribution of raw mutated cells (bottom panel). **e** The genomic track around the *TCF3* locus. The position of the mutation is shown on the scATAC-seq tracks with the mutated cluster labeled in red. The CRE with mutation is highlighted in a gray background, and significant peak-to-gene linkages are indicated by loops. **f** The genomic track of H3K27ac CUT&Tag signal around the *TCF3* with the mutated sample labeled in red.

mutations affecting TF binding and AML-related gene expression were specifically depleted in T/NK cells, whereas passenger mutations showed no such depletion in these cells (Supplementary Fig. 5b–j, see “Methods” section for details).

Additionally, our observations revealed that in certain cases, the expression alterations of downstream target genes may result from the cumulative effects of multiple mutations through regulatory cascades, rather than individual mutation events. For example, the

mutation chr11:32421211:C:T resides in an active enhancer exhibited preferential accessibility in the HSPC-like clusters and was linked to the promoter of *WT1* (Wilms tumor 1) (Supplementary Fig. 7a), a critical regulator of normal and malignant hematopoiesis<sup>49</sup>. This mutation was predicted to gain binding motifs of MYC and ATF4, but a downregulation trend in its target gene was observed in AML10 (Supplementary Fig. 7b–d). Interestingly, recurrent non-coding somatic and germline mutations within this enhancer suppressed enhancer activity and *WT1* expression by disrupting MYB binding in acute promyelocytic leukemia (APL)<sup>45</sup>. Consequently, we focused our investigation on MYB, the primary transcription factor that natively occupies this enhancer region. Our analysis revealed that a specific mutation chr6:135184142:A:C, within the enhancer region was predicted to disrupt the RARA binding motif, resulting in decreased MYB expression in the same patient (Supplementary Fig. 7e–h). This regulatory alteration may potentially contribute to the observed downregulation of *WT1* expression.

In summary, our findings demonstrated that functional non-coding mutations can alter TF motifs, thereby modulating the expression of downstream target genes, suggesting their potential role as regulatory drivers in AML pathogenesis. While further experimental validation is needed to fully establish the causal relationship between non-coding mutations and gene expression changes, the current evidence provides a foundation for understanding their contribution to disease mechanisms.

## Discussion

Identifying mutations and gene rearrangements is crucial for the diagnosis and subtyping of AML. However, the role of non-coding genetic variants in AML pathophysiology and heterogeneity remains largely unclear. Here, we performed single-cell and bulk multi-omics profiling of bone marrow in AML and explored candidate functional non-coding mutations. Large-scale perturbation screening and functional analyses have demonstrated that non-coding mutations contribute to enhancer dysregulation in leukemogenesis<sup>8</sup>. Nevertheless, many of these mutations lack functional interpretation at the resolution of individual cell types. In this study, we developed eMut to identify and analyze potential functional non-coding mutations that contribute to cellular heterogeneity in AML. Similar to existing single-cell mutation detection methods, eMut addresses the challenges of low-quality and sparse single-cell data<sup>18,19,23</sup> by performing imputation to enhance mutation profiles. Importantly, eMut employed a strategy with several key differences compared to the previous mutation detection pipelines, focusing on the functional consequences of non-coding mutations at four levels: cell-type-specific enrichment, the functional importance of the mutated enhancers, the potential disruption of transcription factor binding motifs, and the effects on downstream target genes.

Our results indicate that the majority of non-coding mutations are patient-specific, while recurrent mutations are cell-type-specific, highlighting the diversity of non-coding mutations in AML. Mutated CREs are enriched with blood-related genetic variants, potentially linked to AML-associated genes, and harbor a higher abundance of functional CREs, supporting the development of therapies targeting functionally mutated CREs, such as those targeting *RUNX1*<sup>44</sup>. Furthermore, these functional non-coding mutations, which alter TF binding and impact proto-oncogenes (via gain-of-function) and tumor suppressors (via loss-of-function), help prioritize candidate mutations for experimental validation and may also serve as potential therapeutic targets through genetic, such as base editing of *BCL11A* enhancer for  $\beta$ -hemoglobinopathies<sup>50</sup>. Our results provide insights into the complex heterogeneity and pathogenesis of AML and offer a valuable reference to guide the development of diagnostic biomarkers for risk stratification and to advance the identification of therapeutic targets beyond the coding region.

This study has several limitations. First, unlike solid tumors, obtaining paired normal blood samples for somatic mutation detection in AML is challenging. Although using other normal tissues as controls is feasible, it is

important to acknowledge that chromatin accessibility varies across cell types and patients<sup>12,13</sup>. This variability may limit the utility of scATAC-seq data in detecting somatic mutations, particularly those specific to hematopoietic cell types. To address this, we employed the Monopogen method<sup>19</sup> to predict potential non-coding mutations within the open chromatin regions. However, it is essential to note that this method currently only predicts non-coding mutations on autosomes. Second, due to the lack of matched single-cell transcriptome data for most samples and the sparsity of such data, we assessed the impact of non-coding mutations on downstream target genes using bulk RNA-seq data. Consequently, it was not possible to entirely exclude the influence of co-occurring mutations on these genes. Third, while these analyses provide a valuable framework for linking genetic variation to disease phenotypes, experimental validation in an appropriate cellular context is essential to confirm the regulatory interactions between individual mutations and genes. In the future, it will be important to collect additional AML samples and conduct further experiments to validate the functional impact of selected candidate non-coding mutations, as demonstrated by the validation of *WT1* non-coding mutations<sup>45</sup>.

## Methods

### Sample acquisition and patient consent

This study includes ten samples of acute myeloid leukemia (AML). Samples were collected from the Fujian Institute of Hematology, the Fujian Provincial Key Laboratory on Hematology, and Fujian Medical University Union Hospital, and all patients provided written informed consent. Clinical information (e.g., age and gender) along with diagnostic genetic testing (cytogenetic and molecular genetic) were obtained for each patient (Supplementary Data 1). The studies involving human participants were reviewed and approved by the ethics committee of Union Hospital, Fujian Medical University. All ethical regulations relevant to human research participants were followed.

### Cell lines

The HEK293T cell line was a gift from Huafeng Xie's laboratory at South China University of Technology, and the NIH3T3 cell line was a gift from Dawang Zhou's laboratory at Xiamen University. Both HEK293T and NIH3T3 cells were cultured in DMEM with 10% FBS medium at 37 °C with 5% CO<sub>2</sub> in a humidified incubator.

### Bone marrow sample preparation

Bone marrow samples from patients were 1:1 mixed with cold D-PBS, then loaded to Ficoll-Paque PLUS (cytiva, 17144002) and centrifuged at 400 × g for 30 min. The bone marrow mononuclear cells (BMMCs) in the middle layer were transferred out and washed twice with cold D-PBS.

### RNA-seq

One million BMMCs were collected, and total RNA was extracted by the MiniBEST Universal RNA Extraction Kit (Takara, 9767). Libraries of mRNA were constructed by Novogene (Beijing, China). Briefly, mRNA was purified using poly-T oligo-attached magnetic beads. Fragmentation was performed by divalent cations before first-strand cDNA synthesis by random hexamer and M-MuLV reverse transcriptase. The second strands were synthesized using DNA polymerase I and RNase H. After adapter ligation, 370–420 bp cDNA fragments were purified and amplified for PE150 sequencing by Illumina NovaSeq 6000.

### Bulk ATAC-seq

One million BMMCs were collected and centrifuged at 300 × g for 5 min, resuspended in 200  $\mu$ L lysis buffer (10 mM Tris buffer pH 8.0, 10 mM NaCl, 3 mM MgCl<sub>2</sub>, 0.1% CA630, 0.1% Tween-20, 0.01% Digitonin) on ice for 3 min, after adding 1 ml wash buffer, cells were centrifuged at 500 × g for 5 min, and resuspend with wash buffer. 50,000 cells from each AML sample were taken for ATAC reaction. ATAC reactions were performed following the manuscript of the TruePrep DNA Library Prep Kit (Vazyme, TD501).

## CUT&Tag

CUT&Tag experiments were performed according to the manuscript of the Hyperactive Universal CUT&Tag Assay Kit (Vazyme, TD903). Briefly, 50,000 BMMCs were collected and incubated with ConA Beads, anti-H3K27ac antibody (Abcam, ab177178), goat-anti-rabbit secondary antibody (Vazyme), pA/G-Tnp, after tagmentation, accessible chromatin regions were purified and amplified for PE150 sequencing by Illumina NovaSeq 6000.

## Diagnostic genetic testing

Diagnostic genetic testing was conducted by Tianjin Sino-US Diagnostics Laboratory, with G-band analysis performed on bone marrow samples obtained from patients to determine karyotypes. Molecular mutation screening for 40 genes (Supplementary Data 1) was performed using the Illumina high-throughput sequencing platform, based on target region probe capture technology. The mutation data were filtered and analyzed using public databases such as dbSNP (V156), 1000 Genomes<sup>51</sup>, gnomAD<sup>52</sup>, COSMIC<sup>53</sup>, and ClinVar<sup>54</sup>, alongside functional prediction software like PolyPhen-2<sup>55</sup> and SIFT<sup>56</sup>, as well as the internal database of Tianjin Sino-US Diagnostics Laboratory, to obtain the final analysis results.

## Single-cell ATAC-seq library preparation

Cell preparation was similar to bulk ATAC-seq. Single-cell ATAC library construction steps were mainly referenced from Paired-seq<sup>24</sup> and SHARE-seq<sup>57</sup>. Briefly, ATAC reactions were performed in 50ul ATAC buffer (20 mM Tris-HCl pH 8.0, 66 mM KCl, 3 mM MgCl<sub>2</sub>, 16% DMF) with barcoded-Tn5 at 37 °C for 30 min, and stopped by adding 10 μL stop buffer (50 mM Tris-HCl pH 8.0, 50 mM EDTA). Transposed cells were then pooled and split into 96-well plates containing distinct barcodes (Supplementary Data 5), after three rounds of barcode ligation, cells were pooled together and split into sub-libraries. Sub-libraries were reverse-crosslinked, and purified using PCR & DNA Cleanup Kit (NEB, T1030S), and amplified for 15 cycles. PCR products were then purified by 0.8× DNA Clean Beads (Vazyme, N411-02) for PE150 sequencing by Illumina NovaSeq X plus or BGI DNBSEQ-T7.

## Processing and analysis of scATAC-seq data

**Data preprocessing.** The scATAC-seq sequencing data from 10 individuals were processed similarly to SHARE-seq<sup>57</sup>. First, cell barcodes were extracted from Read2 and added to the SeqID of Read1. Merged Read1 was mapped to the genome by BWA<sup>58</sup>. PCR duplicates were removed after mapping using Picard (<http://broadinstitute.github.io/picard/>). Fragment files were generated from BAM files using a custom script.

**Identification of cell types.** Then, preprocessing and downstream analysis of data used fragment files as input to ArchR (Version 1.0.2)<sup>59</sup>. Cells with a FriTSS below 2 or nFragments below 300 were all filtered from the following analysis. After quality control, we used the ‘addIterativeLSI’ function with 25,000 variable features and 30 dimensions for dimensionality reduction and ‘addClusters’ with resolution=0.8 to identify clusters. The two-dimensional representation of the data was generated using the ‘addUMAP’ function. The gene activity scores were similarly smoothed using the MAGIC algorithm<sup>60</sup>, and motif activity scores were computed with ChromVAR<sup>61</sup>. Both sets of scores were then visualized on Uniform Manifold Approximation and Projection (UMAP).

**Linkage of cis-regulatory elements to target genes.** Candidate peak-gene links were identified by the ‘addPeak2GeneLinks’ function in ArchR<sup>59</sup>. This procedure firstly links peaks located within a genomic distance of 250 kb to the TSS of each gene. Subsequently, the Pearson correlation coefficient of log<sub>2</sub>-normalized accessibility and gene activity scores was computed. Eventually, 33,483 high-confidence peak-to-gene

linkages were determined by selecting those pairs with a Pearson correlation coefficient exceeding 0.2.

## eMut overview

**Mutation detection at the single-cell level.** The eMut began with individual BAM files of single-cell sequencing data. It removed reads that have high alignment mismatches (default of four mismatches) and low mapping quality. For tumor samples with paired normal samples, somatic single-nucleotide variants (SNVs) and small insertions and deletions (INDELs) were called using GATK Mutect2<sup>35</sup> with its default settings and retained based on “PASS” labeling. For tumor samples without paired normal samples, eMut applied GATK Mutect2 to identify germline and somatic mutations and utilized Monopogen<sup>19</sup> with its default setting to predict potential somatic mutations for each cell. All mutations were annotated by VEP (V102) and the dbSNP database (V156). The vcf2maf method was used to convert the VCF format files of annotated mutations into the MAF format, which is subsequently better handled.

**Imputation of mutation profile.** Considering dropout events due to the sparsity of single-cell technical data, we applied SCAVENGE<sup>36</sup> with its default parameters to infer potential mutated cells. Briefly, an M-kNN graph was constructed based on scATAC-seq data to represent cell-cell similarity. For a given mutation, the mutated cells (as seed cells) were projected onto the M-kNN graph. Through network propagation of these seed nodes, relevant cells were identified as potential mutated cells.

**Identification of cell-type-specific mutations.** For each cell type, we determined whether the number of mutated cells for a given mutation significantly exceeds what would be expected by chance using a hypergeometric test. We defined mutations as over-occurring in a specific cell type if they have a false discovery rate (FDR) of less than 0.05.

**Prediction of the disruptiveness of mutation on TF binding sites.** To explore the impact of mutations in their located enhancer, motifbreakR<sup>62</sup> was applied to predict TF motif disruptions (loss or gain) for numerous single-nucleotide variants using various sources of TF motifs (e.g., JASPAR<sup>63</sup>, ENCODE<sup>64</sup>). In the predicted results, “strong” effect motif change will be considered as the potential impact of mutations.

**Identification of hypermutated CREs.** Among all accessible regions, we adapted the ActiveDriverWGS method<sup>65</sup> with modifications to identify hypermutated CREs based on scATAC-seq data. Specifically, we changed adjacent flanking genomic regions to flanking accessible regions (±500 kbps) for training the model of expected mutations, we identified hypermutated CREs (observed excess expected mutations) in each sample.

## Mutation analysis in AML

**Mutation detection.** Considering AML tumor samples without paired normal samples, we applied the Monopogen module in eMut to identify their potential somatic mutations, and the GATK module to identify germline mutations. Additionally, SComatic was employed to call somatic mutations using parameters min\_cell\_type = 1 and min\_cell\_type = 2.

**Hypermutated CREs annotation.** We identified the hypermutated CREs of each AML sample based on germline and somatic mutations identified by the GATK module in eMut. The AML-related genes (MalaCards), hematopoietic-related GWAS<sup>46</sup>, and cis-eQTL<sup>47</sup> (<https://www.eqtngen.org/phase1.html>) datasets were downloaded to annotate these hypermutated CREs.

**Mutational spectra and signatures.** We conducted mutational spectra analyses for mutations in AML identified by scATAC-seq (detected by Monopogen or SComatic), ATAC-seq (detected by GATK), WES-seq

(The Cancer Genome Atlas, <https://portal.gdc.cancer.gov>), and WGS-seq (ICGC<sup>39</sup>, [http://dcc.icgc.org/releases/PCAWG/consensus\\_snv\\_indel/](http://dcc.icgc.org/releases/PCAWG/consensus_snv_indel/)) using the R package MutationalPatterns<sup>38</sup>. Based on the COSMIC Mutational Signatures catalog v.3<sup>66</sup>, we estimated the similarity of selected mutational processes to the observed mutational spectrum using the function `get_known_signatures` with default parameter values.

**The depletion of functional somatic mutations in T/NK.** Statistical analysis of the mutated cell frequency (for somatic mutations) and mutation frequency (for germline mutations) revealed a significant reduction in T/NK cells (Supplementary Fig. 5b–e). To account for confounding effects of open chromatin regions and sequencing depth, we normalized somatic mutation frequency using germline mutation frequency as a background for each cell (Supplementary Fig. 5f). Additionally, to determine whether functional somatic mutations were depleted in T/NK cells while passenger mutations were not, we calculated the somatic-to-germline mutation rate for four somatic mutation categories: all somatic mutations, somatic mutations altering TF motif, somatic mutations altering TF motif near AML-related genes, and somatic mutations altering TF motif that modulate AML-related gene expression. Cells were ranked based on these rates, and the top cells at different thresholds were selected to assess whether T/NK cells exhibited significant depletion compared to random.

### Statistics and reproducibility

Gene expression differences and mutation frequencies/ratios were compared using the Wilcoxon rank-sum test (two-sided). Enrichment analyses, including cell type, functional CRE, survival gene, and mutated cell enrichments, were conducted using the hypergeometric test. Pearson correlation analysis evaluated the correlation between the mutated cell and blast cell proportions. Survival differences between groups were assessed using the log-rank test. All p-values were adjusted for multiple testing using the Benjamini–Hochberg false discovery rate (FDR) correction. Statistical significance was defined as:  $P \geq 0.05$  (not significant, ns);  $P < 0.05$  (\*);  $P < 0.01$  (\*\*); and  $P < 0.001$  (\*\*\*)). To ensure reproducibility, bulk ATAC-seq served as biological replicates (10 AML samples) to validate scATAC-seq findings, while RNA sequencing included three technical replicates for data robustness.

### Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

### Data availability

The datasets utilized in this study were listed in Supplementary Table 1. The scATAC-seq data produced in this study have been deposited in the Genome Sequence Archive (GSA-human, <https://ngdc.cnbc.ac.cn/gsa/>)<sup>67</sup> under accession code HRA007220. The matched bulk ATAC-seq, CUT&Tag, and RNA-seq data are accessible via the GSA-human accession code HRA007242. Access to patients' raw sequencing data is restricted to protect privacy, and access can be requested on the GSA-human website.

### Code availability

The code of the eMut pipeline for this study is available on GitHub (<https://github.com/xmuhuanglab/eMut>). Our analyses also utilize published software tools, with detailed descriptions of their usage and parameter settings provided in the “Methods” section.

Received: 23 December 2024; Accepted: 20 May 2025;

Published online: 26 May 2025

### References

- Arber, D. A. et al. The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia. *Blood* **127**, 2391–2405 (2016).
- Dohner, H. et al. Diagnosis and management of AML in adults: 2017 ELN recommendations from an international expert panel. *Blood* **129**, 424–447 (2017).
- Alaggio, R. et al. The 5th edition of the World Health Organization classification of haematolymphoid tumours: lymphoid neoplasms. *Leukemia* **36**, 1720–1748 (2022).
- 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).
- Desai, P. et al. Somatic mutations precede acute myeloid leukemia years before diagnosis. *Nat. Med.* **24**, 1015–1023 (2018).
- Mer, A. S. et al. Biological and therapeutic implications of a unique subtype of NPM1 mutated AML. *Nat. Commun.* **12**, 1054 (2021).
- Bahr, C. et al. A Myc enhancer cluster regulates normal and leukaemic haematopoietic stem cell hierarchies. *Nature* **553**, 515–520 (2018).
- Li, K. et al. Noncoding variants connect enhancer dysregulation with nuclear receptor signaling in hematopoietic malignancies. *Cancer Discov.* **10**, 724–745 (2020).
- Maurano, M. T. et al. Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190–1195 (2012).
- Shlyueva, D., Stampfel, G. & Stark, A. Transcriptional enhancers: from properties to genome-wide predictions. *Nat. Rev. Genet.* **15**, 272–286 (2014).
- Kvon, E. Z., Waymack, R., Gad, M. & Wunderlich, Z. Enhancer redundancy in development and disease. *Nat. Rev. Genet.* **22**, 324–336 (2021).
- Corces, M. R. et al. Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nat. Genet.* **48**, 1193–1203 (2016).
- Granja, J. M. et al. Single-cell multiomic analysis identifies regulatory programs in mixed-phenotype acute leukemia. *Nat. Biotechnol.* **37**, 1458–1465 (2019).
- Chen, C. et al. Single-cell multiomics reveals increased plasticity, resistant populations, and stem-cell-like blasts in KMT2A-rearranged leukemia. *Blood* **139**, 2198–2211 (2022).
- Fan, H. et al. Single-cell chromatin accessibility profiling of acute myeloid leukemia reveals heterogeneous lineage composition upon therapy-resistance. *Commun. Biol.* **6**, 765 (2023).
- Lambo, S. et al. A longitudinal single-cell atlas of treatment response in pediatric AML. *Cancer Cell* **41**, 2117–2135.e12 (2023).
- Nuno, K. et al. Convergent epigenetic evolution drives relapse in acute myeloid leukemia. *eLife* **13**, e93019 (2024).
- Massarat, A. R. et al. Discovering single nucleotide variants and indels from bulk and single-cell ATAC-seq. *Nucleic Acids Res.* **49**, 7986–7994 (2021).
- Dou, J. et al. Single-nucleotide variant calling in single-cell sequencing data with Monopogen. *Nat. Biotechnol.* (2023).
- Ameen, M. et al. Integrative single-cell analysis of cardiogenesis identifies developmental trajectories and non-coding mutations in congenital heart disease. *Cell* **185**, 4937–4953.e23 (2022).
- Ober-Reynolds, B. et al. Integrated single-cell chromatin and transcriptomic analyses of human scalp identify gene-regulatory programs and critical cell types for hair and skin diseases. *Nat. Genet.* **55**, 1288–1300 (2023).
- Xiong, K. et al. Multimodal integration of single cell ATAC-seq data enables highly accurate delineation of clinically relevant tumor cell subpopulations. Preprint at *bioRxiv* <https://doi.org/10.1101/2024.10.11.617736> (2024).
- Muyas, F. et al. De novo detection of somatic mutations in high-throughput single-cell profiling data sets. *Nat. Biotechnol.* (2023).
- Zhu, C. et al. An ultra high-throughput method for single-cell joint analysis of open chromatin and transcriptome. *Nat. Struct. Mol. Biol.* **26**, 1063–1070 (2019).

25. Rosenbauer, F. & Tenen, D. G. Transcription factors in myeloid development: balancing differentiation with transformation. *Nat. Rev. Immunol.* **7**, 105–117 (2007).
26. Orkin, S. H. & Zon, L. I. Hematopoiesis: an evolving paradigm for stem cell biology. *Cell* **132**, 631–644 (2008).
27. Wu, C. Y. et al. Integrative single-cell analysis of allele-specific copy number alterations and chromatin accessibility in cancer. *Nat. Biotechnol.* **39**, 1259–1269 (2021).
28. Nikolic, A. et al. Copy-scAT: deconvoluting single-cell chromatin accessibility of genetic subclones in cancer. *Sci. Adv.* **7**, eabg6045 (2021).
29. Ramakrishnan, A. et al. epiAneufinder identifies copy number alterations from single-cell ATAC-seq data. *Nat. Commun.* **14**, 5846 (2023).
30. Wang, X., Jin, Z., Shi, Y. & Xi, R. Detecting copy-number alterations from single-cell chromatin sequencing data by AtacCNA. *Cell Rep. Methods* **5**, 100939 (2025).
31. DePristo, M. A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
32. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
33. Kim, S. et al. Strelka2: fast and accurate calling of germline and somatic variants. *Nat. Methods* **15**, 591–594 (2018).
34. Reble, E., Castellani, C. A., Melka, M. G., O'Reilly, R. & Singh, S. M. VarScan2 analysis of de novo variants in monozygotic twins discordant for schizophrenia. *Psychiatr. Genet.* **27**, 62–70 (2017).
35. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
36. Yu, F. et al. Variant to function mapping at single-cell resolution through network propagation. *Nat. Biotechnol.* **40**, 1644–1653 (2022).
37. Papaemmanuil, E. et al. Genomic classification and prognosis in acute myeloid leukemia. *N. Engl. J. Med.* **374**, 2209–2221 (2016).
38. Blokzijl, F., Janssen, R., van Boxtel, R. & Cuppen, E. MutationalPatterns: comprehensive genome-wide analysis of mutational processes. *Genome Med.* **10**, 33 (2018).
39. Consortium, I.T.P.-C.A.o.W.G. Pan-cancer analysis of whole genomes. *Nature* **578**, 82–93 (2020).
40. Khwaja, A. et al. Acute myeloid leukaemia. *Nat. Rev. Dis. Prim.* **2**, 16010 (2016).
41. Puram, R. V. et al. Core Circadian clock genes regulate leukemia stem cells in AML. *Cell* **165**, 303–316 (2016).
42. Jiang, Y. et al. CEBPG promotes acute myeloid leukemia progression by enhancing EIF4EBP1. *Cancer Cell Int.* **21**, 598 (2021).
43. Gu, Z. & Hubschmann, D. rGREAT: an R/bioconductor package for functional enrichment on genomic regions. *Bioinformatics* **39**, btac745 (2023).
44. Mill, C. P. et al. RUNX1-targeted therapy for AML expressing somatic or germline mutation in RUNX1. *Blood* **134**, 59–73 (2019).
45. Song, H. et al. Recurrent noncoding somatic and germline WT1 variants converge to disrupt MYB binding in acute promyelocytic leukemia. *Blood* **140**, 1132–1144 (2022).
46. Vuckovic, D. et al. The polygenic and monogenic basis of blood traits and diseases. *Cell* **182**, 1214–1231.e11 (2020).
47. Vosa, U. et al. Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nat. Genet.* **53**, 1300–1310 (2021).
48. Colaprico, A. et al. TCGAAbilinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res.* **44**, e71 (2016).
49. Pronier, E. et al. Genetic and epigenetic evolution as a contributor to WT1-mutant leukemogenesis. *Blood* **132**, 1265–1278 (2018).
50. Zeng, J. et al. Therapeutic base editing of human hematopoietic stem cells. *Nat. Med.* **26**, 535–541 (2020).
51. Genomes Project, C. et al. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
52. Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
53. Tate, J. G. et al. COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res.* **47**, D941–D947 (2019).
54. Landrum, M. J. et al. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* **46**, D1062–D1067 (2018).
55. Adzhubei, I. A. et al. A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248–249 (2010).
56. Ng, P. C. & Henikoff, S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res.* **31**, 3812–3814 (2003).
57. Ma, S. et al. Chromatin potential identified by shared single-cell profiling of RNA and Chromatin. *Cell* **183**, 1103–1116.e20 (2020).
58. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at <https://arxiv.org/abs/1303.3997> (2013).
59. Granja, J. M. et al. ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nat. Genet.* **53**, 403–411 (2021).
60. van Dijk, D. et al. Recovering gene interactions from single-cell data using data diffusion. *Cell* **174**, 716–729.e27 (2018).
61. Schep, A. N., Wu, B., Buenrostro, J. D. & Greenleaf, W. J. chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nat. Methods* **14**, 975–978 (2017).
62. Coetzee, S. G., Coetzee, G. A. & Hazelett, D. J. motifbreakR: an R/Bioconductor package for predicting variant effects at transcription factor binding sites. *Bioinformatics* **31**, 3847–3849 (2015).
63. Castro-Mondragon, J. A. et al. JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **50**, D165–D173 (2022).
64. Kheradpour, P. & Kellis, M. Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. *Nucleic Acids Res.* **42**, 2976–2987 (2014).
65. Zhu, H. et al. Candidate cancer driver mutations in distal regulatory elements and long-range chromatin interaction networks. *Mol. Cell* **77**, 1307–1321.e10 (2020).
66. Alexandrov, L. B. et al. The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).
67. Chen, T. et al. The genome sequence archive family: toward explosive data growth and diverse data types. *Genom. Proteom. Bioinform.* **19**, 578–583 (2021).

## Acknowledgements

This work was sponsored by the National Natural Science Foundation of China (92474104, 82470211, 32370586). This work was also supported by Fujian provincial health technology project (2021ZD01005), Joint Funds for the innovation of science and Technology, Fujian province (2021Y9050) and Fujian Province Science and Technology Major Special Project (2022Y2034016), the Fundamental Research Funds for the Central Universities (20720230068) and the Wang Deyao Project Fund.

## Author contributions

M.Z., J.Z. and J.H. conceived and designed the experiments. J.H. and N.L. supervised the study. Z.Z. and H.L. collected the patient samples and M.Z. and J.D. processed the samples. M.Z., Y.Y. and J.D. performed experiments. J.Z. and M.Z. performed the analysis. M.Z., J.Z. and J.H. wrote the manuscript with the input from other authors. All authors read and approved the final version of the manuscript.

## 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/s42003-025-08257-8>.

**Correspondence** and requests for materials should be addressed to Nainong Li or Jialiang Huang.

**Peer review information** *Communications Biology* thanks the anonymous reviewers for their contribution to the peer review of this work. Primary Handling Editors: Kuangyu Yen and Aylin Bircan. 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-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, 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 you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. 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-nc-nd/4.0/>.

© The Author(s) 2025