SCIENCE CHINA
Life Sciences

•RESEARCH PAPER•

May 2020 Vol.63 No.5: 675–687
https://doi.org/10.1007/s11427-019-1574-x

Characterization of functional transposable element enhancers in acute myeloid leukemia

Yingying Zeng[1†], Yaqiang Cao[1†], Rivka Sukenik Halevy[2,3,4†], Picard Nguyen[2,3], Denghui Liu[1], Xiaoli Zhang[1], Nadav Ahituv[2,3*] & Jing-Dong J. Han[1,5*]

[1]*CAS Key Laboratory of Computational Biology, CAS-MPG Partner Institute for Computational Biology, Shanghai Institute of Nutrition and Health, Chinese Academy of Sciences Center for Excellence in Molecular Cell Science, Collaborative Innovation Center for Genetics and Developmental Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China;*
[2]*Department of Bioengineering and Therapeutic Sciences, University of California San Francisco, San Francisco 94158, USA;*
[3]*Institute for Human Genetics, University of California San Francisco, San Francisco 94143, USA;*
[4]*Sackler School of Medicine, Tel-Aviv University, Tel Aviv 6997801, Israel;*
[5]*Peking-Tsinghua Center for Life Sciences, Academy for Advanced Interdisciplinary Studies, Center for Quantitative Biology, Peking University, Beijing 100871, China*

Received September 10, 2019; accepted October 24, 2019; published online March 6, 2020

Transposable elements (TEs) have been shown to have important gene regulatory functions and their alteration could lead to disease phenotypes. Acute myeloid leukemia (AML) develops as a consequence of a series of genetic changes in hematopoietic precursor cells, including mutations in epigenetic factors. Here, we set out to study the gene regulatory role of TEs in AML. We first explored the epigenetic landscape of TEs in AML patients using ATAC-seq data. We show that a large number of TEs in general, and more specifically mammalian-wide interspersed repeats (MIRs), are more enriched in AML cells than in normal blood cells. We obtained a similar finding when analyzing histone modification data in AML patients. Gene Ontology enrichment analysis showed that genes near MIRs in open chromatin regions are involved in leukemogenesis. To functionally validate their regulatory role, we selected 19 MIR regions in AML cells, and tested them for enhancer activity in an AML cell line (Kasumi-1) and a chronic myeloid leukemia (CML) cell line (K562); the results revealed several MIRs to be functional enhancers. Taken together, our results suggest that TEs are potentially involved in myeloid leukemogenesis and highlight these sequences as potential candidates harboring AML-associated variation.

transposable element, enhancer, promoter, MIR, acute myeloid leukemia

Citation: Zeng, Y., Cao, Y., Halevy, R.S., Nguyen, P., Liu, D., Zhang, X., Ahituv, N., and Han, J.D.J. (2020). Characterization of functional transposable element enhancers in acute myeloid leukemia. Sci China Life Sci 63, 675–687. https://doi.org/10.1007/s11427-019-1574-x

INTRODUCTION

Transposable elements (TEs), including DNA transposons and retrotransposons, comprise nearly half of the human genome (Rebollo et al., 2012). Retrotransposons including long interspersed repeats (LINE), short interspersed repeats (SINE), and long terminal repeat elements constitute major components of the human genome (Burns and Boeke, 2012). TEs have been shown to be involved in human diseases by causing insertional mutations in specific genes (Belancio et al., 2009) as well as by creating recombined *cis*-acting signals that alter gene expression (Rebollo et al., 2012). Only a small proportion of retrotransposons can transpose their DNA, which suggests that they could function in other ways

†Contributed equally to this work
*Corresponding authors (Nadav Ahituv, email: Nadav.Ahituv@ucsf.edu; Jing-Dong J. Han, email: jackie.han@pku.edu.cn)

(Göke and Ng, 2016). For example, TE-derived lncRNAs can modulate pluripotency and impact human development (Durruthy-Durruthy et al., 2016). Hypomethylation of TEs was also detected in malignancies of the lung (Daskalos et al., 2009) and bladder (Wolff et al., 2010). A large proportion of retrotransposons are marked by H3K4me1 and H3K27ac, bound by transcription factors, and thus might act as regulators to influence the expression of nearby genes (Göke and Ng, 2016; Su et al., 2014; Trizzino et al., 2017) and mark cell identities (Cao et al., 2019). For example, CTCF binding repeat DXPas34 in the Xist locus regulates imprinted X inactivation (Cohen et al., 2007); moreover, mammalian-wide interspersed repeats (MIRs) are under strong selection (Kamal et al., 2006), and some MIRs can function as enhancer boosters (Cao et al., 2019; Smith et al., 2008). L1 insertions also frequently occur in cancers at hypomethylated genomic regions, which could disrupt the expression of target genes (Göke and Ng, 2016).

MIRs belong to the SINE elements and are about 260 bp in length, including a tRNA-related 5′ cap and a 70 bp conserved central domain (Carnevali et al., 2016). MIRs account for 2.93% of the human genome, and about 500,000 of them have been annotated, which include four subfamilies: MIR, MIRb, MIRc, and MIR3 (Carnevali et al., 2016; Rodić and Burns, 2013). MIRs have been shown to participate in the regulation of gene expression, site cleavage, and polyadenylation (Hughes, 2000), and interact with other TEs such as L2 in the 3D genome (Cao et al., 2019). The integration of MIRs into genes is important for gene control and evolution. For example, the zinc finger gene zfOC1 contains MIRs that are thought to play a role in human heart and vision impairment (Hughes, 2000). MIRs have also been found to play regulatory roles in the human genome, serving as chromosomal barriers or elements that insulate enhancers (Carnevali et al., 2016). The binding of MIRs by the transcription factors POLR2A, TBP, MAZ, MAX, and YY1 highly expressed in some cell lines suggests that the expression of MIRs may be regulated by these transcription factors (Carnevali et al., 2016). MIRs are also the only type of transposon element positively associated with tissue-specific gene expression in their vicinity (Jjingo et al., 2011). A series of studies have indicated that MIRs can contain binding sites for transcription factors (Cao et al., 2019; Polavarapu et al., 2008; Thornburg et al., 2006). MIRs can also contain the constituent sequences of enhancers and act as enhancers (Cao et al., 2019; Huda et al., 2011; Jjingo et al., 2014). Moreover, some MIRs have sites overlapping with microRNAs (Piriyapongsa et al., 2007) or with antisense transcripts (Conley et al., 2008; Jjingo et al., 2014).

Acute myeloid leukemia (AML) is a relatively common leukemic type associated with a low survival rate (Oran and Weisdorf, 2012). AML could be caused by constituent mutations in hematopoietic precursor cells, leading to the production of abnormal cells in the bone marrow and blood (Kumar, 2011). AML cells have high genetic stability (Ptasinska et al., 2014) and molecular heterogeneity (Solh et al., 2014; Valk et al., 2004). The known genetic alterations in AML include mutations in oncogenes and tumor suppressor genes as well as cytogenetic abnormalities. Recurrent chromosomal structural variations such as t(8;21) have been well studied in AML cells (Buenrostro et al., 2013; Rowley, 2008). With the introduction of large-cohort whole-genome sequencing studies, additional genetic hallmarks of AML are being revealed. These include multiple somatic mutations, such as genetic rearrangements, mutations in genes encoding DNA methyltransferases, chromatin modifiers, and transcription factors (Ley et al., 2013; Papaemmanuil et al., 2016). Like in other human diseases, the role of TEs in AML is unclear; we therefore set out to investigate whether TEs can act as enhancers or promoters during the development of AML.

The analysis of epigenetic modifications could enable the detection of TEs that are associated with human disease (Lamprecht et al., 2010). Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq) (Buenrostro et al., 2013) is now a widely used method to assess genome-wide chromatin accessibility. Here, we focused on understanding the role of TEs and their relevance to AML. We first integrated ATAC-seq data and histone modification ChIP-seq data in AML cell lines to assess the relationship between TEs and AML. We found a cluster of TEs that show high ATAC-seq signals specifically enriched in AML cells. These regions also show enrichment for H3K4me1 and H3K27ac histone marks, suggesting their role as enhancers. Since MIRs were enriched in the group of TEs with high ATAC-seq signals in AML cells, we focused on MIRs to test the role of TEs as active regulatory elements in AML. We show that some MIRs, which were specifically enriched in open chromatin regions in AML cells, are positive enhancers in Kasumi-1 and chronic myeloid leukemia (CML) cell lines (K562). Our results suggest a possible important role for TE and specifically MIRs in myeloid leukemogenesis.

RESULTS

Identification of AML-specific TEs in open chromatin regions

To explore whether TEs are in an open or closed chromatin state in AML, bulk ATAC-seq data (GSE74912; Corces et al., 2016) of AML patient blast samples and normal blood progenitor cells were analyzed, which included 23 samples from AML blast cells, 7 samples from hematopoietic stem cells (HSCs), 6 samples from multipotent progenitor cells (MPPs), 3 samples from lymphocyte-initiated pluripotent progenitor cells (LMPP), 8 samples from common myeloid

progenitor cells (CMP), 5 samples from common lymphoid progenitor cells (CLP), 7 samples from particle granulocyte macrophage progenitor cells (GMP), 7 samples from megakaryocyte erythroid progenitor cells (MEPs), 5 samples from $CD4^+$ T cells, 5 samples from $CD8^+$ T cells, 4 samples from $CD19^+CD20^+$ B cells, 6 samples from natural killer cells, 6 samples from monocytes, and 8 samples from erythroblasts. Repeat annotations for all species used in this study were downloaded from UCSC (Karolchik et al., 2004). ATAC-seq signals shown in log2-transformed FPKM (Fragments Per Kilobase per Million mapped) values of TEs were computed by iteres (Xie et al., 2013) and then transformed into a 0/1 binary matrix. Specifically, as the nonzero ATAC-seq signal values of TEs containing 0 ATAC-seq signal value in more than half of all samples follow a normal distribution, we used this distribution as noise background and set the log2-transformed FPKM signal to 1 if it was >mean+3std in this noise distribution, and to 0 otherwise. TEs having all zeros in the matrix were removed, and the remaining TEs were grouped into four clusters (Methods). Principal component analysis (PCA)-based *k*-means clustering (Ding and He, 2004; Pedregosa et al., 2011) was carried out to explore the open/closed (active/inactive) patterns of these TEs (Methods, Figure 1A). A total of 42,443 TEs in open chromatin regions were obtained and grouped into four clusters. Our analyses found a cluster of TEs in open chromatin regions specifically enriched in AML samples, the C3 cluster, totaling 20,869 TEs (Table S1 in Supporting Information), including 27.29% MIRs, 14.51% LINE-2 (L2) elements, and other TE families. Although ATAC-seq signals of C3 TEs have some difference among the 23 AML blast samples, which might reflect the heterogeneity across different patients, for example, different gene mutations, there are high correlations among 12 of the 23 samples and high correlations among the other 11 samples, suggesting two subtypes of the AML samples (Figure S1A in Supporting Information). The information on AML patients derived from Corces et al. (2016) was added in Table S2 in Supporting Information. TE family enrichment analysis was performed in all four clusters (Methods); interestingly, MIRs were highly enriched in all four clusters, which may suggest that MIRs were located at open chromatin regions in blood-related cells. In particular, MIRs and L2s were both highly enriched in the C3 cluster (FDR<1×10^{-10}) (Figure 1B). Therefore, we wonder whether MIRs in C3 play a role in the process of AML, perhaps as enhancers.

Transcription factors (TFs) play important roles in the progression of AML. For example, the transcription factor CEBPA is known to be critical for myeloid differentiation (Gonzalez et al., 2017), FOSB is essential for growth in human AML cells (Bergerson et al., 2012), and JUN is a key regulator of unfolded protein response in AML (Zhou et al., 2017). AP-1 proteins are composed of one member of each of the two different families of related bZIP proteins, the Fos family (c-Fos, Fra-1, Fra-2, and FosB) and the Jun family (c-Jun, JunB, and JunD), providing a multiplicity of regulatory control (Karin et al., 1997). AP-1 factors have been shown to play an important role in AML (Czibere et al., 2008). To identify the role of TFs in AML TEs specifically enriched in open chromatin regions, we conducted motif analysis in C3 cluster TEs. Compared to a random genomic background and all TEs in open chromatin regions, the motifs of AP-1 factors and CEBP families were significantly enriched in C3 TEs (Figure 2A). We found that motifs of FOSB, FOS, and JUND, which are related to the MAPK pathway (Vinciguerra et al., 2004), were also highly enriched in the C3 cluster. Furthermore, about 8% of the MIR-containing regions in C3 had these TF motifs. We also computed the expression value of these TFs in the AML blast samples with normal blood progenitor cells using the RNA-seq data (GSE74246) (Figure S1B in Supporting Information). We found that these TFs had higher expression in AML blast samples than in most normal blood cell lines. We collected the published CEBPA ChIP-seq data in the Kasumi-1 cell line from GSM1501162 (Ptasinska et al., 2014), and found that CEBPA significantly binds to C3 TEs in this cell line (Chi-squared test *P*-value<2.2×10^{-16}). Although other ChIP-seq data for other TFs containing these enriched motifs are still needed to verify these TFs' function in AML, these results suggested that these TFs might be associated with TEs enriched in the open chromatin region and have some function in AML.

In addition, functional enrichment analysis using GREAT (McLean et al., 2010) revealed that genes near C3 TEs, C3 MIRs, and C3 L2s are highly related to immune response (Figure 2B and C; Figure S1D in Supporting Information). In particular, the most significant disease annotation term among genes near C3 MIRs is AML (Figure 2D). The genes near TEs were defined by GREAT (Figure 2B and C) or HOMER (Heinz et al., 2010) (Figure 2D). HOMER analysis of disease terms from DisGeNET (Piñero et al., 2017) for the C3 nearest genes revealed enrichment for AML regulatory genes near MIRs, and the most significant annotated disease term for C3 MIRs was AML. For the disease terms annotated for the genes near C3 L2, the most significant term was related to encephalomyelitis, while the third annotated disease term was leukemia (Figure S1C in Supporting Information). Taken together, our results suggest that C3 TEs, in particular MIRs, might be functional elements in AML.

TEs could potentially function as enhancers or promoters in AML

We next set out to analyze the potential regulatory function of C3 TEs in AML. To identify whether they have potential promoter or enhancer activity in AML cells, we analyzed the following datasets: (i) AML H3K27ac, H3K4me1, and

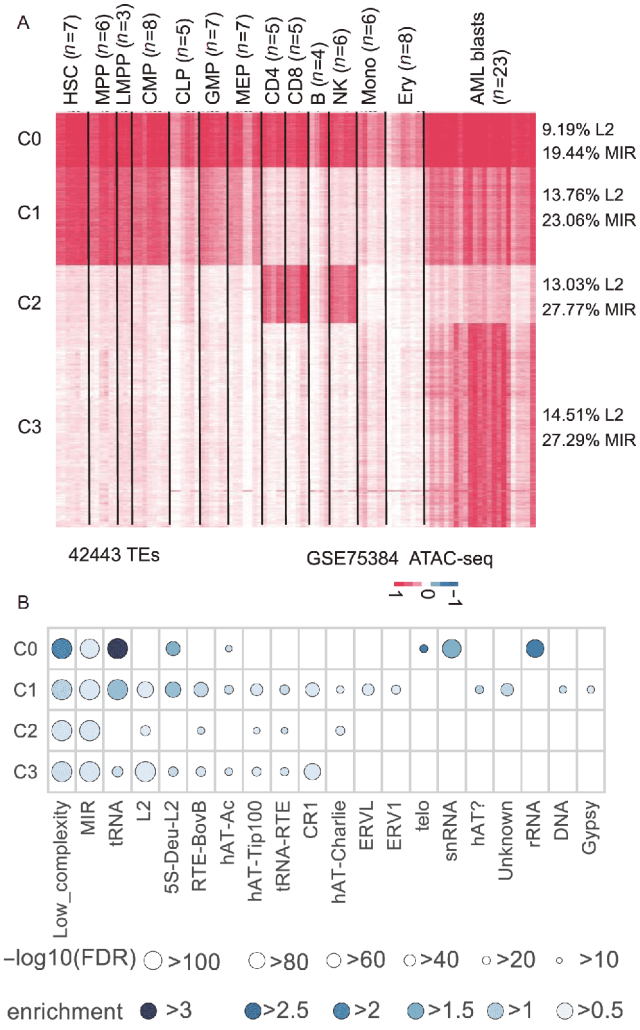

Figure 1 TEs specifically enriched in open chromatin regions of AML patients' blast cells. A, Heatmap of PCA-based *k*-means clustering for the TE binary matrix (enriched in open chromatin region or not), totaling 42,443 TEs originating from ATAC-seq (GSE75384). Red stands for enriched in open chromatin regions and white for not enriched in open chromatin regions. Different clusters are labeled with a prefix of C. The C3 cluster comprises acute myeloid leukemia (AML)-specific enriched in open chromatin region TEs. B, TE family enrichment in different clusters of TEs defined in (A). Bubble size indicates corrected enrichment *P*-value and color marks enrichment score. The enrichment test was performed with a combination of the binomial test and hypergeometric test.

H3K4me3 ChIP-seq peaks from the BLUEPRINT Database 2016 release (http://dcc.blueprint-epigenome.eu/#/home) (Adams et al., 2012); (ii) H3K27ac and H3K4me1 profiles in the MOLM-14 cell line (GSE65138; Pelish et al., 2015), which is an established MLL-AF9 rearranged AML cell line (Matsuo et al., 1997); and (iii) p300 ChIP-seq data from the Kasumi-1 cell line (GSE76464; Mandoli et al., 2016), which is an established t(8,21) translocation (RUNX1-ETO fusion gene) AML cell line (Ptasinska et al., 2014).

Using the AML H3K27ac, H3K4me1, and H3K4me3 peak datasets from the BLUEPRINT Database (Adams et al., 2012), we obtained the enhancer-like repeats (ELRs) and promoter-like repeats (PLRs). ELRs are TEs overlapping with both H3K4me1 and H3K27ac peaks, while PLRs are TEs overlapping with both H3K4me3 and H3K27ac peaks. From the results of Gene Set Enrichment Analysis (GSEA), we found that TEs in C3 are significantly enriched for the BLUEPRINT AML TE enhancers, namely, ELRs (Figure 2E). The C3 TE activity profiles were further corroborated by the H3K27ac and H3K4me1 profiles in the MOLM-14 cell line (GSE65138; Pelish et al., 2015) (Figure 3A). In addition, the TEs overlapping between C3 TEs and ELRs or PLRs both show high activity for H3K27ac and H3K4me1. We also collected H3K27ac ChIP-seq data in the THP-1 cell line, which was established on the MLL-AF9 translocation cell line (GSM2108046; Prange et al., 2017). C3 TEs were significantly enriched in TEs overlapping with H3K27ac peaks (Chi-squared test, *P*-value $<2.2\times10^{-16}$). This suggests that C3 TEs are significantly enriched in these AML-related cell lines and might play roles during this process. The MIRs in C3 also have higher H3K27ac, H3K4me1, and H3K4me3 activity than other repressive histone modifications or input in 10 BLUEPRINT AML patients. The histone modification profiles for 10 BLUEPRINT AML patients were variable; for

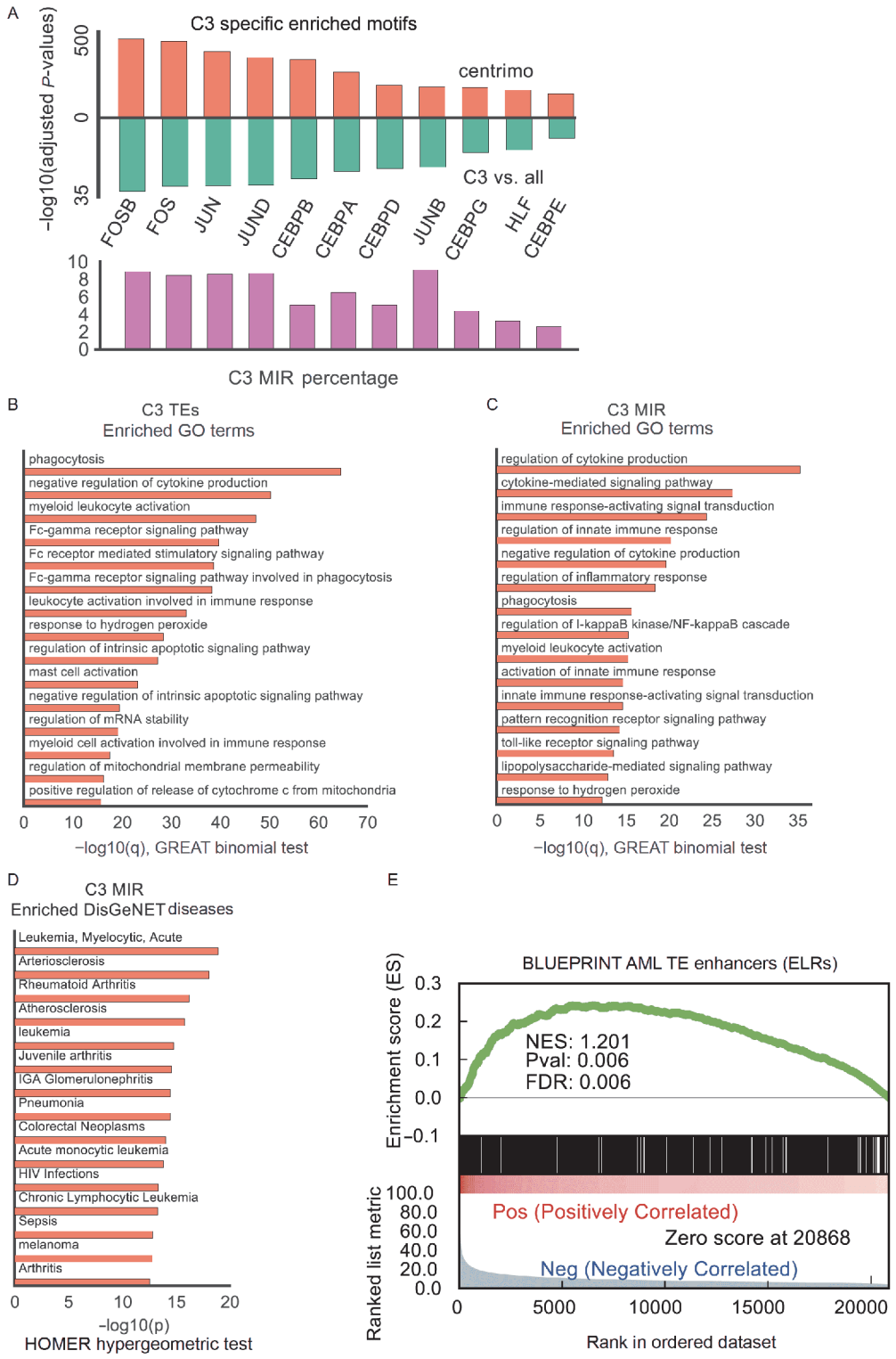

Figure 2 Functional annotation for AML-specific active TEs. A, Top enriched motifs for C3 TEs. The significance test was carried out by CentriMo. Orange color indicates the significance called with random background; (cyan color indicates the significance called by comparing with all enriched in open chromatin region TEs) cyan color indicates the significance called with all detected open TEs treated as background. B, Bar plot of the top 15 enriched GO terms for C3 TEs, annotated by GREAT. C, Bar plot of the top 15 enriched GO terms for the MIRs in the C3 group annotated by GREAT. D, Bar plot of the top 15 enriched DisGeNET diseases for C3 MIRs, annotated by HOMER. E, Gene Set Enrichment Analysis (GSEA) result of C3 TEs to BLUEPRINT AML TE enhancers (2016 release). C3 TEs are not significantly enriched for BLUEPRINT AML TE promoters. ES represents enrichment score.

example, eight patients have higher H3K4me3 than H3K4me1 signal, while two patients have the opposite pattern. However, this might reflect different age, sex, or other different confounding factors. Nonetheless, the common

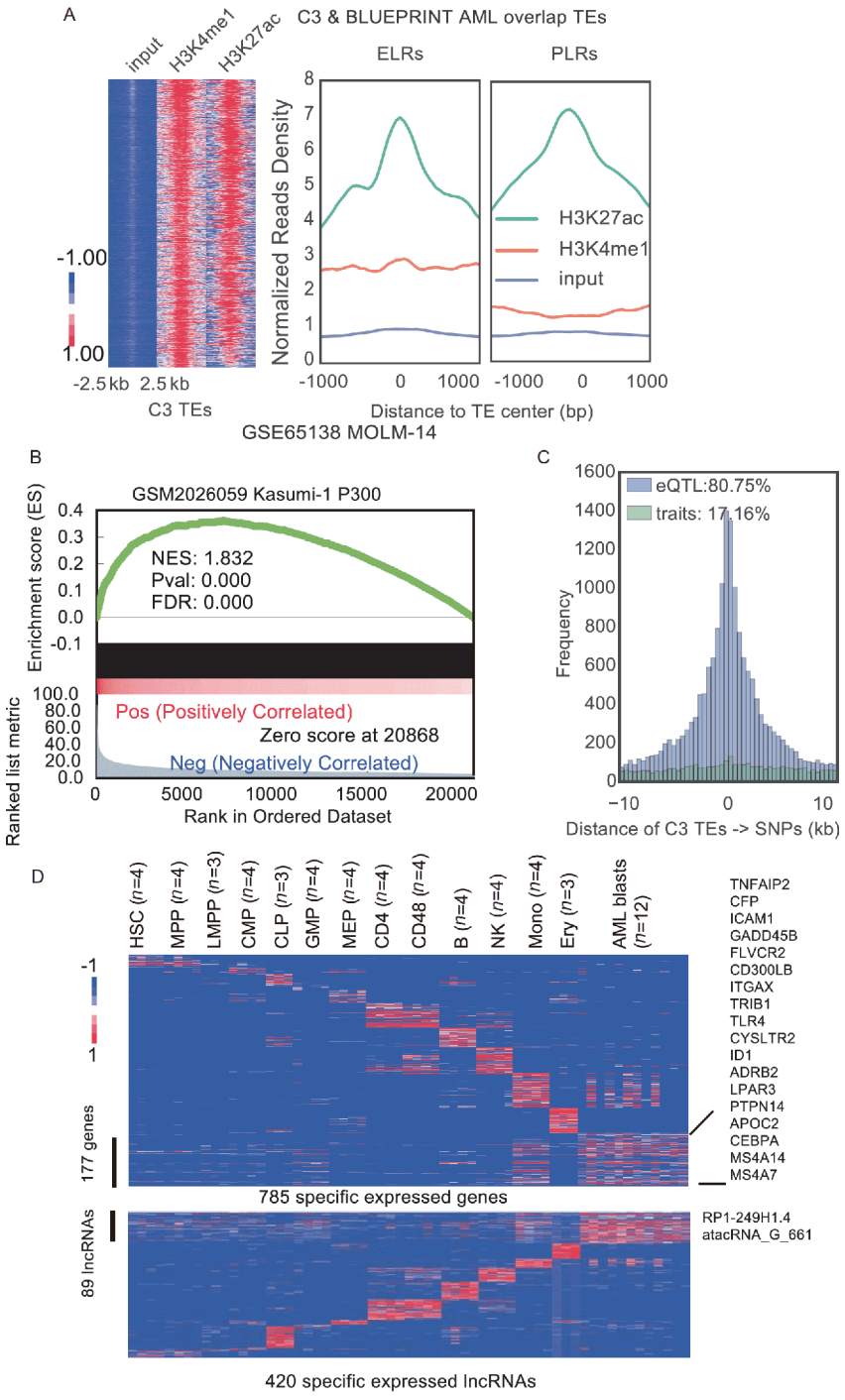

Figure 3 Epigenetic, genetic, and gene expression evidence of AML-specific active TEs. A, H3K4me1 and H3K27ac signals of C3 TEs and average H3K4me1 and H3K27ac profile of enhancer-like repeats (ELRs) and promoter-like repeats (PLRs) in the MOLM-14 cell line (GSE65138). ELRs are those C3 TEs that overlap with both H3K4me1 and H3K27ac peaks from BLUEPRINT Database AML samples; other remaining C3 TEs are defined as PLRs. B, Gene Set Enrichment Analysis (GSEA) result of C3 TEs to TEs bound by p300 in the AML Kasumi-1 cell line (GSE76464). C, Histogram for distance of C3 TEs to nearest expression quantitative trait loci (eQTLs) from Genome-Wide Repository of Associations Between SNPs and Phenotypes (GRASP) and Genome-wide association study (GWAS) Catalog SNPs. D, Heatmap of selected specifically expressed protein-coding genes (upper panel) and lncRNAs (lower panel).

pattern is that H3K4me1, H3K27ac, and H3K4me3 signals of C3 MIRs are higher than those of other repressive histone markers and input control. This suggests that C3 MIRs might act as enhancers or promoters in AML (Figure S2A in Supporting Information). From the distribution of the distance of C3 TEs to their nearest genes, 73.9% of C3 TEs are more than 10 kb away from their nearest genes (Figure S3A in Supporting Information). In addition, 70.9% of C3 MIRs

are more than 10 kb away from their nearest genes (Figure S3B in Supporting Information). This indicates that most C3 TEs might act as enhancers, but not promoters. From the GSEA results of the EP300 ChIP-seq peaks from the Kasumi-1 cell line (GSE76464) (Mandoli et al., 2016), C3 TEs are significantly enriched for the TEs overlapping with EP300 peaks (Figure 3B, FDR<0.001). A total of 2,490 TEs of 20,869 C3 TEs overlap with EP300 peaks, and 690 MIRs in C3 have EP300 binding. MOLM-14 is an MLL-AF9 rearranged AML cell line, and Kasumi-1 is a RUNX1-ETO fusion AML cell line. C3 TEs were significantly enriched in the Kasumi-1 cell EP300 peaks. Taken together, these findings suggest that most C3 TEs might function as enhancers, while some might function as promoters in AML-related cell lines.

Genome-wide association studies (GWAS) (Manolio, 2010) have demonstrated that the majority of genetic variants are found in noncoding regions of the genome and are therefore likely to be involved in regulating gene expression. An expression quantitative locus (eQTL) is a locus that has genetic variants and can explain the expression change of genes (Nica and Dermitzakis, 2013). We wondered whether the disease-related C3 may be enriched with disease-related SNPs, which would suggest regulatory roles in disease pathogenesis. Using human cell line/tissue eQTLs, we analyzed C3 TEs for the enrichment of eQTL. For eQTL analysis, we collected the SNP data from Genome-Wide Repository of Associations Between SNPs and Phenotypes (GRASP) (Eicher et al., 2015) and GWAS Catalog (MacArthur et al., 2017). The distance between C3 TEs and SNPs is shown in Table S3 in Supporting Information. The distribution of the distance of C3 TEs to the nearest eQTLs from GRASP and GWAS Catalog SNPs (Figure 3C**)** showed that 80.75% of C3 TEs have eQTLs and 17.16% have an SNP within 10 kb upstream and downstream, indicating that these TEs are associated with disease traits. To study the evolutionary conservation of the C3 TEs, we compared them to the sets of TEs enriched in open chromatin regions in mouse AML cells on day 8 and day 27 (Sen et al., 2016) using ATAC-seq data (GSE87646). C3 TEs are significantly enriched in mouse AML TEs enriched in open chromatin regions (P=0.038 by GSEA), according to sequence similarities (Figure S3D in Supporting Information, Methods**)**. Although 72% of C3 TEs cannot be mapped to the mouse genome, MIR is the most conserved TE in both human and mouse AML cells (Figure S3E in Supporting Information).

Potential regulation of AML specifically expressed genes by TEs

We next set out to examine whether genes that play important roles in AML reside near TEs that are predicted to function as gene regulatory elements in our study. RNA-seq data of 13 normal hematopoietic cell types and AML blasts from GSE74246 (Corces et al., 2016) were analyzed. Overall, 782 AML specifically differentially expressed genes and 420 specifically differentially expressed lncRNAs were obtained by the entropy-based Jensen-Shannon divergence (JSD) method (Cabili et al., 2011) (Figure 3D, Methods). Examples of MIRs enriched in open chromatin regions that potentially function as enhancers or promoters to regulate highly expressed genes in AML blasts are shown in Figure 4A and Figure S3C in Supporting Information. Data used in these two figures were from the following published datasets: ATAC-seq data of 13 normal hematopoietic cell types and 3 AML cell types (pHSC, LSC, blasts) (GSE75384). Among them, pHSC stands for pre-leukemic hematopoietic stem cell, LSC for leukemia stem cell, where replicates belonging to one cell type were aggregated together; then, MACS2 (Zhang et al., 2008) was used to generate the RPM (Reads per Million mapped) normalized visualization tracks. H3K27ac, H3K4me1 ChIP-seq, and input data in the MOLM-14 cell line (GSE65138) and leukemia blast RNA-seq data (GSE75384) (Corces et al., 2016) were also used. In Figure 4A, the first example of an MIR belongs to ELRs and its nearby gene is TNFAIP2, which is associated with tumor progression and highly expressed in AML blasts, suggesting that this MIR might function as an enhancer. The second example of an MIR, which is in the promoter region of nearby gene CFP, indicates that this MIR can act as a promoter to activate the nearest gene expression.

MIRs are functional enhancers in Kasumi-1 and K562 cell lines

The following criteria were used to select candidate C3 cluster MIRs for experimental validation of enhancer activities: (i) the intensity of the ATAC-seq signal; (ii) the presence of a EP300 ChIP-seq peak; (iii) the presence of AML-associated transcription factor binding motifs; and (iv) the vicinity to known AML-related genes. Nineteen regions containing MIRs enriched in AML cells (data in Table S4 in Supporting Information**)** were selected and tested for enhancer activity using dual luciferase reporter assay in AML cells (Kasumi-1). Our *in vitro* enhancer assays were initially designed for an AML cell line (Kasumi-1). However, the transfection efficiency was very low for this cell line in all methods used, including X-tremeGENE, lipofectamine, polyethylenimine, Max and electroporation. Despite having low transfection efficiency in these cells, four regions showed strong enhancer activity (Figure 4B). As an additional validation, we also tested all 19 regions in the CML cell line (K562) as their infection efficiency is much higher. All four sequences that were positive for enhancer activity in Kasumi-1 cells were also positive in these cell lines, along with seven additional regions (Figure 4B, data in Table S5 in Supporting Information). Furthermore, we also examined the

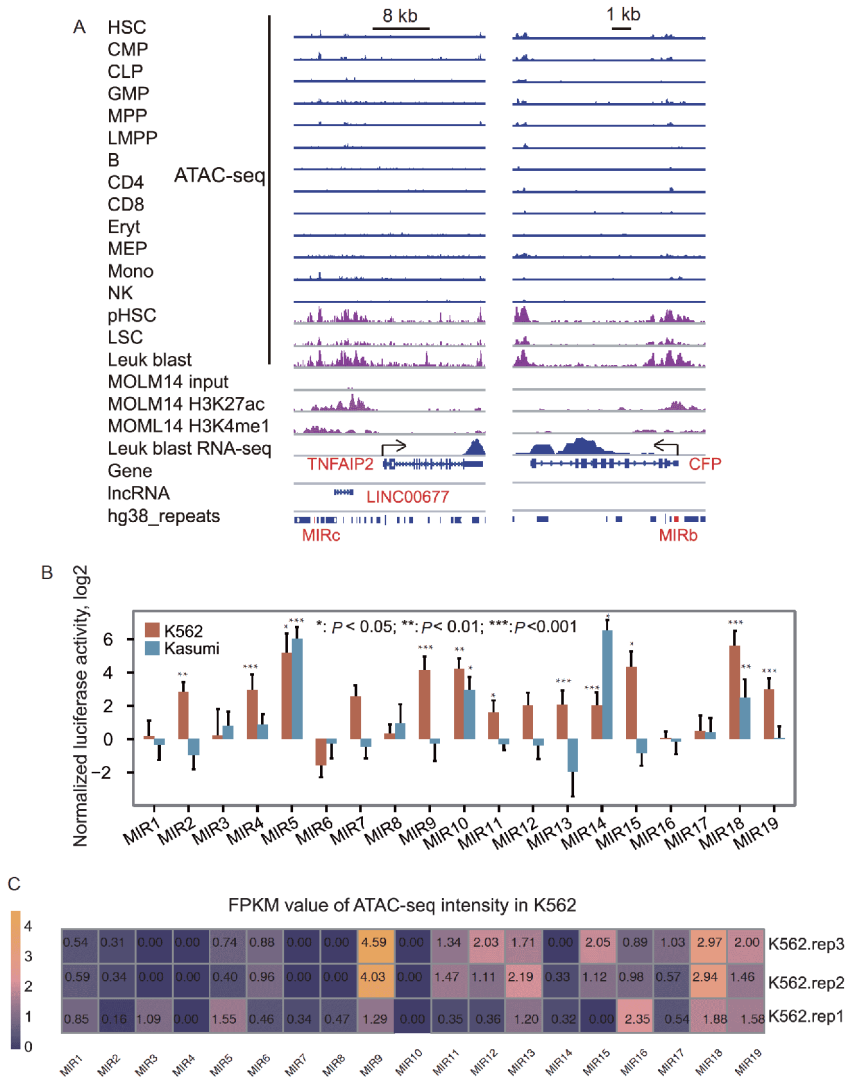

Figure 4 Validation of enhancer activities of TEs that are specifically associated with AML open chromatic regions. A, Two examples of suggested enhancer activity of MIR (for *TNFAIP2*) and promoter activity (for *CFP*). pHSC: pre-leukemic hematopoietic stem cell; LSC: leukemia stem cell. FPKM value was used to visualize the signal; the same scale was used for the same data type when visualizing in the genome browser. B, Enhancer assay results for the 19 selected MIRs in Kasumi-1 and K562 cell lines. C, ATAC-seq signal of MIR in the K562 cell line.

ATAC-seq signal of MIRs in the K562 cell line using the data from GSE99173 (Liu et al., 2017), and found that MIR9 and MIR18, which have high enhancer luciferase activity, also had high ATAC-seq signal in K562 (Figure 4C). Taken together, our results show that MIRs could be functional enhancers in these two cancer cell types.

DISCUSSION

TEs are important elements for genome plasticity and evolution. The insertion of TEs in the human genome can cause genetic dysfunction and alteration of gene expression, contributing to cancer and other human diseases (Chenais, 2015). For example, the deletion of TEs like Alus could account for mutations in several genes, which could cause von Hippel-Lindau disease (Casarin et al., 2006). Altered expression of TEs and other active elements in different phases of cancer cell progression could be responsible for driving tumorigenesis (Burns, 2017).

Here, we studied the potential important gene regulatory roles of TEs, and MIRs in particular, in AML. Through the analysis of data in AML cell lines, we found a cluster of TEs (C3) that showed special enrichment and a gain in activity in AML compared with that in normal blood cells. Interestingly, some transcription factor motifs, such as those for FOSB, FOS, and JUND, which are related to the MAPK pathway, are enriched in C3 TEs. Inappropriate constitutive activation of the MAPK pathway is known to play a role in the leukemic transformation of myeloid cells (Milella et al., 2001). Thus, TEs enriched in open chromatin regions may serve as response elements in leukemogenesis downstream

of the MAPK pathway. Furthermore, the motif for the transcription factor CEBPA is also enriched in C3. This transcription factor is known to coordinate proliferation arrest and the differentiation of myeloid progenitors, and its mutation has also been detected in AML cells, and used as a prognostic marker (Grossmann et al., 2012; Matsuo et al., 2014). This suggests a specific role for these TEs in AML. In addition, most of the C3 TEs had high H3K27ac and H3K4me1 activity, suggesting enhancer activity of these regions. In addition, the genes near MIRs in the C3 cluster are particularly associated with AML leukemogenesis and immune response. TNFAIP2, which was originally identified as a TNFα-inducible gene in endothelial cells, was shown to be involved in the pathogenesis of AML (Rusiniak et al., 2000). We found that it is in the vicinity of MIR and has a high expression level in AML. Intercellular adhesion molecule-1 (ICAM1) can interact with lymphocyte function-associated antigen-1 (LFA-1), which functions in the adhesion and migration of AML cells (Zhang et al., 2006). We found an MIR in the vicinity of ICAM1 that is highly expressed in AML samples. Moreover, blocking Adrb2 in leukemic mice can significantly reduce their survival (Hanoun et al., 2014). In our analysis, ADRB2 was shown to be highly expressed and near the L3. These examples of genes suggest that the TEs we identified might serve as potential enhancers to modulate key factors involved in AML initiation or progression. Enhancer assays performed in Kasumi-1 and K562 cells revealed several MIRs that function as enhancers *in vitro*.

MIRs have been reported to have widespread enhancer activities, some of which even function as enhancer boosters (Cao et al., 2019; Smith et al., 2008). This study further shows that MIRs can be epigenetically active in disease models and further substantiates the hypothesis that all TEs might be epigenetically active under certain normal or disease biological contexts, providing regulatory sites/sources for TFs. This study also provides a new perspective for AML researchers and clinicians to enlarge their scope for studying disease etiology and outcome. MIRs are potentially involved in myeloid leukemogenesis, but the specific mechanisms for MIR involvement in AML are still unknown. We speculate that MIRs function as independent enhancers or through interaction with other important upstream regulators to alter gene expression in cancer cells. Future studies are needed to explore relationships between MIRs and the most enriched TFs. Our study suggests a potential role for TEs, especially MIRs, as enhancers in AML and CML.

METHODS

Activated human TEs identified from ATAC-seq

Bulk ATAC-seq data from GSE75384 were mapped to the human genome build hg38 by bowtie2 (v2.2.3) (Langmead and Salzberg, 2012) using the adapter trimmed reads and the same parameters as in a previous study (Corces et al., 2016). For mapping, bowtie2 (Langmead and Salzberg, 2012) was used with specific parameters set to "-q –end-to-end –very-sensitive –no-discordant –no-mixed –maxins 2000," and the reads/pairs that could be aligned to multiple locations were kept and mapped to the best matched position. Next, duplicates were removed and unmapped reads were discarded, while multi-mapped reads were retained for the next step of analysis. According to Derrien et al. (2012), if the read length is ≥100 bp, even using a unique mapping strategy, the mappability is still high in most repeat regions. The ATAC-seq data used here are paired-end, and each end's read length is 75 bp (together is 150 bp; single-ended mapped reads were removed); hence, these data mostly overcame the mappability problems, even using the unique mapping strategy. Unique mapping was used for ATAC-seq data as the read length (75 bp paired-end reads) was long enough to overcome the issue of mappability in TE regions (Derrien et al., 2012). Transposable element annotations are from RepeatMasker identified by the Repbase (Jurka et al., 2005; Kohany et al., 2006) classification system. The quantitative FPKM of each individual TE was calculated by iteres (Xie et al., 2013) and then log2-transformed. We used a simple method to model the ATAC-seq signal to noise for TEs as follows: For a TE, if more than half of the samples in the same group have 0 signal, then nonzero signals were collected to draw a normal distribution, and the mean+3std was set as the noise cut-off. The values higher than the cut-off were set as 1 (open chromatin region) and those lower were set as 0 (closed chromatin region) to obtain the binary matrix.

Active mouse TEs identified from ATAC-seq

ATAC-seq data of mouse $CD8^+$ cells in the set conditions (8 days after acute LCMV infection, 27 days after acute LCMV infection) (GSE87646; Sen et al., 2016) were mapped to the mouse genome build mm10 by bowtie (v2.2.3) (Langmead and Salzberg, 2012) using the adapter trimmed reads and the same parameters as in a previous study (Corces et al., 2016). The quantitative FPKM of each individual TE was calculated by iteres (Xie et al., 2013) and then log2-transformed. The values were converted to binary values using the same method as TEs enriched in open chromatin regions identified from human ATAC-seq. Then, we defined the TEs enriched in open chromatin regions in mouse AML cells as those TEs are enriched in open chromatin regions in mouse on day 8 and day 27 (Sen et al., 2016). bnMapper (Denas et al., 2015) was used to map TEs between mouse AML TEs and human C3 TEs with the following command: "bnMapper.py -f BED4 –gap 20 –threshold 0.1."

Method for analyzing TEs' functional enrichment

Gene Set Enrichment Analysis (GSEA) (Subramanian et al., 2005) was used to test whether a predefined set of TEs has significantly high or low ranking in a ranked list of TEs sorted by the ATAC-seq signal. The python package gseapy (Kuleshov et al., 2016; Subramanian et al., 2005) was used for this with the following command: "gseapy prerank -f pdf –max-size 100000 –r a.rnk -g b.gmt –o result."

PCA-based *k*-means clustering

PCA-based *k*-means clustering implemented in scikit-learn (Pedregosa et al., 2011) was used to cluster TEs defined from bulk ATAC-seq data. By analyzing the variance ratio explained by the first 30 components in PCA, we used k=5 as the number of clusters for *k*-means with the initial centers of the first five PCs, as the first five PCs explained the majority of variance and the curve gradually saturated. After the *k*-means clustering, we abandoned one cluster showing a pattern like noise.

Transposable element family enrichment analysis

P-values obtained by hypergeometric test and binomial test were combined by Stouffer's *Z*-score method (Darlington and Hayes, 2000; Stouffer, 1949). FDR correlations were performed using these combined *P*-values. The hypergeometric test, binomial test, and FDR functions were implemented in Orange Bioinformatics Toolbox (Demsar et al., 2013) and Stouffer's *Z*-score method was implemented in Scipy (Jones et al., 2015). For the test of the significance of TE families, FDR $>1\times10^{-100}$ was assigned as 1×10^{-100} and FDR=1×10^{-10} was defined as the significance cut-off. Enrichment score was calculated as $ES = \frac{k}{m} / \frac{n}{N}$, where k is the number of specific TE families in the input list, m is the number of TEs in the input list, n is the number of TE families in the genome, and N is the number of all TEs.

Motif analysis

CentriMo (v4.10.0) (Bailey and Machanick, 2012) from the MEME package (Bailey et al., 2006) was used to identify the motifs showing significant preference at the center of a set of TE sequences. The motif datasets curated by MEME called HOCOMOCOv10_HUMAN_mono were used, and the motif set was collected by HOCOMOCO (Kulakovskiy et al., 2013). Fimo from MEME was used to identify the locations of motifs within each individual TE with the same motif database.

RNA-seq quantification and *de novo* lncRNA assembly

Sequence reads were aligned using STAR (v2.4.0d) (Dobin et al., 2013). Cufflinks (v2.2.1) (Trapnell et al., 2010) was used to assemble the transcript and compute the FPKM value for every gene. We used the parameter "-g" of cufflinks to guide lncRNA *de novo* assembly.

AML specifically expressed genes

Cell-specifically expressed genes/lncRNAs were obtained by an entropy-based method. Briefly, Jensen-Shannon divergence (JSD) was used to measure the specificity of gene expression patterns and predefined tissue-specific patterns, by a method similar to that reported previously (Cabili et al., 2011). The predefined cell-specific pattern was a binary vector as samples in target tissue are marked 1 and others as 0. Finally, a JSD cut-off >0.7 was used to define the cell-specific genes.

Enhancer assay

The selected MIRs were amplified from human DNA (Roche) and cloned into the pGL4.23 enhancer vector (Promega; Pasquali et al., 2014) using InFusion cloning (Clontech). This vector can be used for gateway cloning of candidate enhancer elements. Cloned regions were verified for proper inserts using Sanger sequencing.

Kasumi-1 cells were grown in suspension in RPMI1640 medium with 20% fetal bovine serum. Transfection efficiency was tested with a number of methods, including a range of electroporation protocols, X-tremeGENE (Version 08), Lipofectamine LTX with Plus reagent (ThermoFisher Scientific), and Polyethylenimine "Max" (Mw 40,000)—High Potency Linear PEI from Polysciences Inc. Lipofectamine LTX with Plus reagent resulted in the highest transfection efficiency (~5%) and was thus used for subsequent experiments. K562 cells were grown in suspension in Iscove's Modified Dulbecco's Medium with 10% fetal bovine serum. Cells were transfected with X-tremeGENE in accordance with the manufacturer's protocol. Transfections were performed in 24-well plates. On the day of transfection, approximately 400,000 Kasumi-1 cells or K562 cells were plated in each well and 1 μg of the PGL4.23 insert contacting plasmid was transfected along with 50 ng of a Renilla plasmid (p-RL-TK; Promega) to correct for transfection efficiency. Transfections were performed in triplicate. We also used an empty pGL4.23 vector as a negative control and the pGL4.13 vector (Promega) as a positive control. Two days after transfection, cells were lysed and luciferase activity was measured using the Dual-Luciferase Reporter Assay Kit using a Glomex 96-microplate luminometer (Promega). Each experiment was repeated using three biological replicates on two different days (a total of six biological replicates). The readings from each tested MIR (luciferase/Renilla) were compared to the reading of the pGL4.23 empty

vector using a *t*-test (one-sided, type 2).

Data availability

Data that support the findings of this study were all published by others; a summary of the used data is presented in Table S6 in Supporting Information.

Compliance and ethics *The author(s) declare that they have no conflict of interest.*

Acknowledgements *This work was supported by grants from the National Natural Science Foundation of China (91749205, 91329302, and 31210103916), Ministry of Science and Technology of China (2015CB964803 and 2016YFE0108700), and a Max Planck fellowship to J. D.J.H. This work was also supported by the National Human Genome Research Institute (NHGRI) and the National Cancer Institute (1R01CA197139), NHGRI (1UM1HG009408), and the National Health Lung and Blood Institute (1R01HL138424) to N.A. R.S.H is a Recipient of a Fellowship Scholarship from the American Healthcare Professionals and Friends for Medicine in Israel. This study made use of data generated by the Blueprint Consortium.*

References

Adams, D., Altucci, L., Antonarakis, S.E., Ballesteros, J., Beck, S., Bird, A., Bock, C., Boehm, B., Campo, E., Caricasole, A., et al. (2012). BLUEPRINT to decode the epigenetic signature written in blood. Nat Biotechnol 30, 224–226.

Bailey, T.L., and Machanick, P. (2012). Inferring direct DNA binding from ChIP-seq. Nucleic Acids Res 40, e128.

Bailey, T.L., Williams, N., Misleh, C., and Li, W.W. (2006). MEME: discovering and analyzing DNA and protein sequence motifs. Nucleic Acids Res 34, W369–W373.

Belancio, V.P., Deininger, P.L., and Roy-Engel, A.M. (2009). LINE dancing in the human genome: transposable elements and disease. Genome Med 1, 97.

Bergerson, R.J., Collier, L.S., Sarver, A.L., Been, R.A., Lugthart, S., Diers, M.D., Zuber, J., Rappaport, A.R., Nixon, M.J., Silverstein, K.A.T., et al. (2012). An insertional mutagenesis screen identifies genes that cooperate with Mll-AF9 in a murine leukemogenesis model. Blood 119, 4512–4523.

Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y., and Greenleaf, W.J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat Methods 10, 1213–1218.

Burns, K.H. (2017). Transposable elements in cancer. Nat Rev Cancer 17, 415–424.

Burns, K.H., and Boeke, J.D. (2012). Human transposon tectonics. Cell 149, 740–752.

Cabili, M.N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., and Rinn, J.L. (2011). Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev 25, 1915–1927.

Cao, Y., Chen, G., Wu, G., Zhang, X., McDermott, J., Chen, X., Xu, C., Jiang, Q., Chen, Z., Zeng, Y., et al. (2019). Widespread roles of enhancer-like transposable elements in cell identity and long-range genomic interactions. Genome Res 40–52.

Carnevali, D., Conti, A., Pellegrini, M., and Dieci, G. (2016). Whole-genome expression analysis of mammalian-wide interspersed repeat elements in human cell lines. DNA Res dsw048.

Casarin, A., Martella, M., Polli, R., Leonardi, E., Anesi, L., and Murgia, A. (2006). Molecular characterization of large deletions in the von Hippel-Lindau (VHL) gene by quantitative real-time PCR. Mol Diag Ther 10, 243–249.

Chenais, B. (2015). Transposable elements in cancer and other human diseases. Cur Cancer Drug Targets 15, 227–242.

Cohen, D.E., Davidow, L.S., Erwin, J.A., Xu, N., Warshawsky, D., and Lee, J.T. (2007). The DXPas34 repeat regulates random and imprinted X inactivation. Dev Cell 12, 57–71.

Conley, A.B., Miller, W.J., and Jordan, I.K. (2008). Human cis natural antisense transcripts initiated by transposable elements. Trends Genet 24, 53–56.

Corces, M.R., Buenrostro, J.D., Wu, B., Greenside, P.G., Chan, S.M., Koenig, J.L., Snyder, M.P., Pritchard, J.K., Kundaje, A., Greenleaf, W. J., et al. (2016). Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. Nat Genet 48, 1193–1203.

Czibere, A., Singh, R., Bruns, I., Zerbini, L.F., and Haas, R. (2008). Ap-1 family members C-Jun, JunB and Fra-2 mediate apoptosis and differentiation in AML through activation of GADD45 alpha and ADFP following non-steroidal anti-inflammatory drug treatment. Blood 112, 436–436.

Darlington, R.B., and Hayes, A.F. (2000). Combining independent *p* values: Extensions of the Stouffer and binomial methods. Psychol Methods 5, 496–515.

Daskalos, A., Nikolaidis, G., Xinarianos, G., Savvari, P., Cassidy, A., Zakopoulou, R., Kotsinas, A., Gorgoulis, V., Field, J.K., and Liloglou, T. (2009). Hypomethylation of retrotransposable elements correlates with genomic instability in non-small cell lung cancer. Int J Cancer 124, 81–87.

Demsar, J., Curk, T., Erjavec, A., Gorup, C., Hocevar, T., Milutinovic, M., Mozina, M., Polajnar, M., Toplak, M., Staric, A., et al. (2013). Orange: Data Mining Toolbox in Python. J Mach Learn Res 14, 2349–2353.

Denas, O., Sandstrom, R., Cheng, Y., Beal, K., Herrero, J., Hardison, R.C., and Taylor, J. (2015). Genome-wide comparative analysis reveals human-mouse regulatory landscape and evolution. BMC Genomics 16, 87.

Derrien, T., Estellé, J., Marco Sola, S., Knowles, D.G., Raineri, E., Guigó, R., and Ribeca, P. (2012). Fast computation and applications of genome mappability. PLoS ONE 7, e30377.

Ding, C., and He, X.F. (2004). Cluster structure of K-means clustering via principal component analysis. Lect Notes Artif Int 3056, 414–418.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15–21.

Durruthy-Durruthy, J., Sebastiano, V., Wossidlo, M., Cepeda, D., Cui, J., Grow, E.J., Davila, J., Mall, M., Wong, W.H., Wysocka, J., et al. (2016). The primate-specific noncoding RNA HPAT5 regulates pluripotency during human preimplantation development and nuclear reprogramming. Nat Genet 48, 44–52.

Eicher, J.D., Landowski, C., Stackhouse, B., Sloan, A., Chen, W., Jensen, N., Lien, J.P., Leslie, R., and Johnson, A.D. (2015). GRASP v2.0: an update on the Genome-Wide Repository of Associations between SNPs and phenotypes. Nucleic Acids Res 43, D799–D804.

Göke, J., and Ng, H.H. (2016). CTRL+INSERT: retrotransposons and their contribution to regulation and innovation of the transcriptome. EMBO Rep 17, 1131–1144.

Gonzalez, D., Luyten, A., Bartholdy, B., Zhou, Q., Kardosova, M., Ebralidze, A., Swanson, K.D., Radomska, H.S., Zhang, P., Kobayashi, S.S., et al. (2017). ZNF143 protein is an important regulator of the myeloid transcription factor C/EBPα. J Biol Chem 18924–18936.

Grossmann, V., Bacher, U., Kohlmann, A., Butschalowski, K., Roller, A., Jeromin, S., Dicker, F., Kern, W., Schnittger, S., Haferlach, T., et al. (2012). Expression of CEBPA is reduced in RUNX1-mutated acute myeloid leukemia. Blood Cancer J e86.

Hanoun, M., Zhang, D., Mizoguchi, T., Pinho, S., Pierce, H., Kunisaki, Y., Lacombe, J., Armstrong, S.A., Dührsen, U., and Frenette, P.S. (2014). Acute myelogenous leukemia-induced sympathetic neuropathy promotes malignancy in an altered hematopoietic stem cell niche.

Cell Stem Cell 15, 365–375.
Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell 38, 576–589.
Huda, A., Bowen, N.J., Conley, A.B., and Jordan, I.K. (2011). Epigenetic regulation of transposable element derived human gene promoters. Gene 475, 39–48.
Hughes, D.C. (2000). MIRs as agents of mammalian gene evolution. Trends Genet 16, 60–62.
Jjingo, D., Conley, A.B., Wang, J., Mariño-Ramírez, L., Lunyak, V.V., and Jordan, I.K. (2014). Mammalian-wide interspersed repeat (MIR)-derived enhancers and the regulation of human gene expression. Mobile DNA 5, 14.
Jjingo, D., Huda, A., Gundapuneni, M., Mariño-Ramírez, L., and Jordan, I. K. (2011). Effect of the transposable element environment of human genes on gene length and expression. Genome Biol Evol 3, 259–271.
Jones, E., Oliphant, T., and Peterson, P. (2015). SciPy: Open source scientific tools for Python, 2001. http://www scipy org 73, 86.
Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O., and Walichiewicz, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res 110, 462–467.
Kamal, M., Xie, X., and Lander, E.S. (2006). A large family of ancient repeat elements in the human genome is under strong selection. Proc Natl Acad Sci USA 103, 2740–2745.
Karin, M., Liu, Z., and Zandi, E. (1997). AP-1 function and regulation. Curr Opin Cell Biol 9, 240–246.
Karolchik, D., Hinrichs, A.S., Furey, T.S., Roskin, K.M., Sugnet, C.W., Haussler, D., and Kent, W.J. (2004). The UCSC Table Browser data retrieval tool. Nucleic Acids Res 493D–496.
Kohany, O., Gentles, A.J., Hankus, L., and Jurka, J. (2006). Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. BMC BioInf 7, 474.
Kulakovskiy, I.V., Medvedeva, Y.A., Schaefer, U., Kasianov, A.S., Vorontsov, I.E., Bajic, V.B., and Makeev, V.J. (2013). HOCOMOCO: a comprehensive collection of human transcription factor binding sites models. Nucleic Acids Res 41, D195–D202.
Kuleshov, M.V., Jones, M.R., Rouillard, A.D., Fernandez, N.F., Duan, Q., Wang, Z., Koplev, S., Jenkins, S.L., Jagodnik, K.M., Lachmann, A., et al. (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res 44, W90–W97.
Kumar, C.C. (2011). Genetic abnormalities and challenges in the treatment of acute myeloid leukemia. Genes Cancer 2, 95–107.
Lamprecht, B., Walter, K., Kreher, S., Kumar, R., Hummel, M., Lenze, D., Köchert, K., Bouhlel, M.A., Richter, J., Soler, E., et al. (2010). Derepression of an endogenous long terminal repeat activates the CSF1R proto-oncogene in human lymphoma. Nat Med 16, 571–579.
Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat Methods 9, 357–359.
Ley, T.J., Miller, C., Ding, L., Raphael, B.J., Mungall, A.J., Robertson, A. G., Hoadley, K., Triche, T.J., Laird, P.W., et al. (2013). Genomic and epigenomic landscapes of adult *de novo* acute myeloid leukemia. N Engl J Med 368, 2059–2074.
Liu, X., Zhang, Y., Chen, Y., Li, M., Zhou, F., Li, K., Cao, H., Ni, M., Liu, Y., Gu, Z., et al. (2017). *In situ* capture of chromatin interactions by biotinylated dCas9. Cell 170, 1028–1043.e19.
MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., Junkins, H., McMahon, A., Milano, A., Morales, J., et al. (2017). The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res 45, D896–D901.
Mandoli, A., Singh, A.A., Prange, K.H.M., Tijchon, E., Oerlemans, M., Dirks, R., Ter Huurne, M., Wierenga, A.T.J., Janssen-Megens, E.M., Berentsen, K., et al. (2016). The hematopoietic transcription factors RUNX1 and ERG prevent AML1-ETO oncogene overexpression and onset of the apoptosis program in t(8;21) AMLs. Cell Rep 17, 2087–2100.
Manolio, T.A. (2010). Genomewide association studies and assessment of the risk of disease. N Engl J Med 363, 166–176.
Matsuo, H., Kajihara, M., Tomizawa, D., Watanabe, T., Saito, A.M., Fujimoto, J., Horibe, K., Kodama, K., Tokumasu, M., Itoh, H., et al. (2014). Prognostic implications of CEBPA mutations in pediatric acute myeloid leukemia: a report from the Japanese Pediatric Leukemia/Lymphoma Study Group. Blood Cancer J 4, e226.
Matsuo, Y., MacLeod, R.A., Uphoff, C.C., Drexler, H.G., Nishizaki, C., Katayama, Y., Kimura, G., Fujii, N., Omoto, E., Harada, M., et al. (1997). Two acute monocytic leukemia (AML-M5a) cell lines (MOLM-13 and MOLM-14) with interclonal phenotypic heterogeneity showing MLL-AF9 fusion resulting from an occult chromosome insertion, ins (11;9)(q23;p22p23). Leukemia 11, 1469–1477.
McLean, C.Y., Bristor, D., Hiller, M., Clarke, S.L., Schaar, B.T., Lowe, C. B., Wenger, A.M., and Bejerano, G. (2010). GREAT improves functional interpretation of cis-regulatory regions. Nat Biotechnol 28, 495–501.
Milella, M., Kornblau, S.M., Estrov, Z., Carter, B.Z., Lapillonne, H., Harris, D., Konopleva, M., Zhao, S., Estey, E., and Andreeff, M. (2001). Therapeutic targeting of the MEK/MAPK signal transduction module in acute myeloid leukemia. J Clin Invest 108, 851–859.
Nica, A.C., and Dermitzakis, E.T. (2013). Expression quantitative trait loci: present and future. Phil Trans R Soc B 368, 20120362.
Oran, B., and Weisdorf, D.J. (2012). Survival for older patients with acute myeloid leukemia: a population-based study. Haematologica 97, 1916–1924.
Papaemmanuil, E., Gerstung, M., Bullinger, L., Gaidzik, V.I., Paschka, P., Roberts, N.D., Potter, N.E., Heuser, M., Thol, F., Bolli, N., et al. (2016). Genomic classification and prognosis in acute myeloid leukemia. N Engl J Med 374, 2209–2221.
Pasquali, L., Gaulton, K.J., Rodríguez-Seguí, S.A., Mularoni, L., Miguel-Escalada, I., Akerman, İ., Tena, J.J., Morán, I., Gómez-Marín, C., van de Bunt, M., et al. (2014). Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants. Nat Genet 46, 136–143.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., and Dubourg, V. (2011). Scikit-learn: Machine learning in Python. J Mach Learn Res 12, 2825–2830.
Pelish, H.E., Liau, B.B., Nitulescu, I.I., Tangpeerachaikul, A., Poss, Z.C., Da Silva, D.H., Caruso, B.T., Arefolov, A., Fadeyi, O., Christie, A.L., et al. (2015). Mediator kinase inhibition further activates super-enhancer-associated genes in AML. Nature 526, 273–276.
Piñero, J., Bravo, À., Queralt-Rosinach, N., Gutiérrez-Sacristán, A., Deu-Pons, J., Centeno, E., García-García, J., Sanz, F., and Furlong, L.I. (2017). DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. Nucleic Acids Res D833–D839.
Piriyapongsa, J., Mariño-Ramírez, L., and Jordan, I.K. (2007). Origin and evolution of human microRNAs from transposable elements. Genetics 176, 1323–1337.
Polavarapu, N., Mariño-Ramírez, L., Landsman, D., McDonald, J.F., and Jordan, I.K. (2008). Evolutionary rates and patterns for human transcription factor binding sites derived from repetitive DNA. BMC Genomics 9, 226.
Prange, K.H.M., Mandoli, A., Kuznetsova, T., Wang, S.Y., Sotoca, A.M., Marneth, A.E., van der Reijden, B.A., Stunnenberg, H.G., and Martens, J.H.A. (2017). MLL-AF9 and MLL-AF4 oncofusion proteins bind a distinct enhancer repertoire and target the RUNX1 program in 11q23 acute myeloid leukemia. Oncogene 36, 3346–3356.
Ptasinska, A., Assi, S.A., Martinez-Soria, N., Imperato, M.R., Piper, J., Cauchy, P., Pickin, A., James, S.R., Hoogenkamp, M., Williamson, D., et al. (2014). Identification of a dynamic core transcriptional network in t(8;21) AML that regulates differentiation block and self-renewal. Cell Rep 8, 1974–1988.
Rebollo, R., Romanish, M.T., and Mager, D.L. (2012). Transposable elements: an abundant and natural source of regulatory sequences for host genes. Annu Rev Genet 46, 21–42.

Rodić, N., and Burns, K.H. (2013). Long Interspersed Element–1 (LINE-1): Passenger or Driver in Human Neoplasms? PLoS Genet e1003402.
Rowley, J.D. (2008). Chromosomal translocations: revisited yet again. Blood 112, 2183–2189.
Rusiniak, M.E., Yu, M., Ross, D.T., Tolhurst, E.C., and Slack, J.L. (2000). Identification of B94 (TNFAIP2) as a potential retinoic acid target gene in acute promyelocytic leukemia. Cancer Res 60, 1824–1829.
Sen, D.R., Kaminski, J., Barnitz, R.A., Kurachi, M., Gerdemann, U., Yates, K.B., Tsao, H.W., Godec, J., LaFleur, M.W., Brown, F.D., et al. (2016). The epigenetic landscape of T cell exhaustion. Science 1165–1169.
Smith, A.M., Sanchez, M.J., Follows, G.A., Kinston, S., Donaldson, I.J., Green, A.R., and Göttgens, B. (2008). A novel mode of enhancer evolution: the Tal1 stem cell enhancer recruited a MIR element to specifically boost its activity. Genome Res 18, 1422–1432.
Solh, M., Yohe, S., Weisdorf, D., and Ustun, C. (2014). Core-binding factor acute myeloid leukemia: Heterogeneity, monitoring, and therapy. Am J Hematol 89, 1121–1131.
Stouffer, S.A. (1949). Studies in Social Psychology in World War II: The American Soldier: Adjustment During Army Life (Princeton University Press).
Su, M., Han, D., Boyd-Kirkup, J., Yu, X., and Han, J.D.J. (2014). Evolution of Alu elements toward enhancers. Cell Rep 7, 376–385.
Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA 102, 15545–15550.
Thornburg, B.G., Gotea, V., and Makałowski, W. (2006). Transposable elements as a significant source of transcription regulating signals. Gene 365, 104–110.
Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol 28, 511–515.
Trizzino, M., Park, Y.S., Holsbach-Beltrame, M., Aracena, K., Mika, K., Caliskan, M., Perry, G.H., Lynch, V.J., and Brown, C.D. (2017). Transposable elements are the primary source of novelty in primate gene regulation. Genome Res 27, 1623–1633.
Valk, P.J.M., Verhaak, R.G.W., Beijen, M.A., Erpelinck, C.A.J., Barjesteh van Waalwijk van Doorn-Khosrovani, S., Boer, J.M., Beverloo, H.B., Moorhouse, M.J., van der Spek, P.J., Löwenberg, B., et al. (2004). Prognostically useful gene-expression profiles in acute myeloid leukemia. N Engl J Med 350, 1617–1628.
Vinciguerra, M., Vivacqua, A., Fasanella, G., Gallo, A., Cuozzo, C., Morano, A., Maggiolini, M., and Musti, A.M. (2004). Differential phosphorylation of c-Jun and JunD in response to the epidermal growth factor is determined by the structure of MAPK targeting sequences. J Biol Chem 279, 9634–9641.
Wolff, E.M., Byun, H.M., Han, H.F., Sharma, S., Nichols, P.W., Siegmund, K.D., Yang, A.S., Jones, P.A., and Liang, G. (2010). Hypomethylation of a LINE-1 promoter activates an alternate transcript of the MET oncogene in bladders with cancer. PLoS Genet 6, e1000917.
Xie, M., Hong, C., Zhang, B., Lowdon, R.F., Xing, X., Li, D., Zhou, X., Lee, H.J., Maire, C.L., Ligon, K.L., et al. (2013). DNA hypomethylation within specific transposable element families associates with tissue-specific enhancer landscape. Nat Genet 45, 836–841.
Zhang, W., Zhang, X., Fan, X., Li, D., and Qiao, Z. (2006). Effect of ICAM-1 and LFA-1 in hyperleukocytic acute myeloid leukaemia. Clin Lab Haematol 28, 177–182.
Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B. E., Nussbaum, C., Myers, R.M., Brown, M., Li, W., et al. (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biol 9, R137.
Zhou, C., Martinez, E., Di Marcantonio, D., Solanki-Patel, N., Aghayev, T., Peri, S., Ferraro, F., Skorski, T., Scholl, C., Fröhling, S., et al. (2017). JUN is a key transcriptional regulator of the unfolded protein response in acute myeloid leukemia. Leukemia 31, 1196–1205.

SUPPORTING INFORMATION

Figure S1 Function annotation for AML specific active TEs.

Figure S2 Histone modification profile of C3 MIRs in BLUEPRINT AML patients' bone marrow.

Figure S3 Conserved active C3 TEs between human and mouse.

Table S1 The location of C3 TEs

Table S2 AML patients infomation

Table S3 Distance between C3 TEs and SNPs

Table S4 Designed primers for 19 testified MIRs

Table S5 Luciferase reporter assay result of validated MIRs

Table S6 Used public datasets

The supporting information is available online at http://life.scichina.com and https://link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.