Cell Reports, Volume *26*

Supplemental Information

Chromatin-Based Classification of Genetically Heterogeneous AMLs into Two Distinct Subtypes with Diverse Stemness Phenotypes

Guoqiang Yi, Albertus T.J. Wierenga, Francesca Petraglia, Pankaj Narang, Eva M. Janssen-Megens, Amit Mandoli, Angelika Merkel, Kim Berentsen, Bowon Kim, Filomena Matarese, Abhishek A. Singh, Ehsan Habibi, Koen H.M. Prange, André B. Mulder, Joop H. Jansen, Laura Clarke, Simon Heath, Bert A. van der Reijden, Paul Flicek, Marie-Laure Yaspo, Ivo Gut, Christoph Bock, Jan Jacob Schuringa, Lucia Altucci, Edo Vellenga, Hendrik G. Stunnenberg, and Joost H.A. Martens

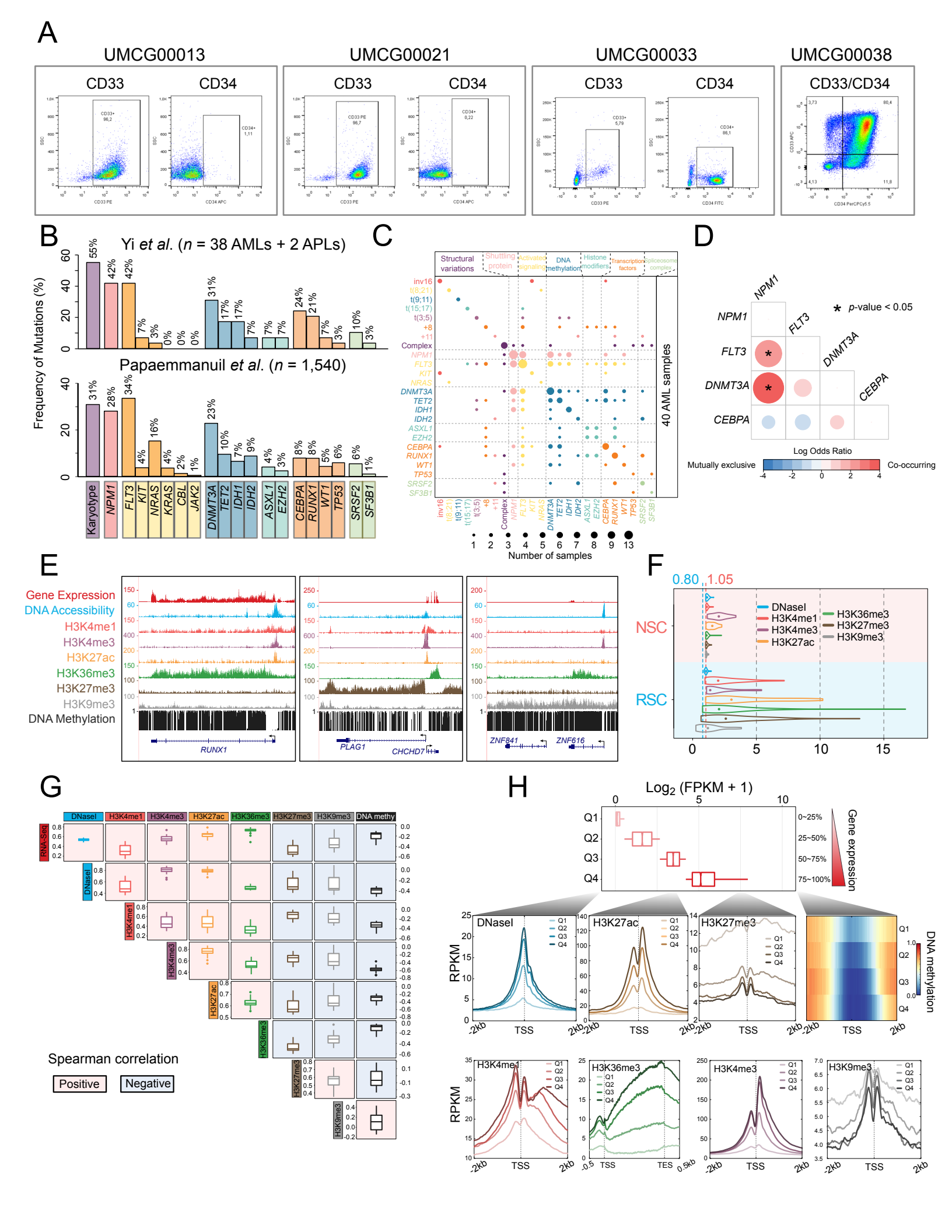

A
UMCG00013
UMCG00021
UMCG00033
UMCG00038
CD33
CD34
CD33/CD34
B
Yi et al. (n = 38 AMLs + 2 APLs)
Papaemmanuil et al. (n = 1,540)
Frequency of Mutations (%)
Karyotype
NPM1
FLT3
KIT
NRAS
KRAS
CBL
JAK2
DNMT3A
TET2
IDH1
IDH2
ASXL1
EZH2
CEBPA
RUNX1
WT1
TP53
SRSF2
SF3B1
C
Structural variations
Shuttling protein
Activated signaling
DNA methylation
Histone modifiers
Transcription factors
Spliceosome complex
40 AML samples
Number of samples
D
NPM1
FLT3
DNMT3A
CEBPA
* p-value < 0.05
Log Odds Ratio
Mutually exclusive
Co-occurring
E
Gene Expression
DNA Accessibility
H3K4me1
H3K4me3
H3K27ac
H3K36me3
H3K27me3
H3K9me3
DNA Methylation
RUNX1
PLAG1
CHCHD7
ZNF841
ZNF616
F
0.80
1.05
NSC
RSC
DNaseI
H3K4me1
H3K4me3
H3K27ac
H3K36me3
H3K27me3
H3K9me3
G
RNA-Seq
DNA methy
Spearman correlation
Positive
Negative
H
Log2 (FPKM + 1)
Q1
Q2
Q3
Q4
0~25%
25~50%
50~75%
75~100%
Gene expression
RPKM
DNA methylation
TSS
TES
-2kb
2kb
0.5kb

Figure S1, related to Figure 1. Mutational Landscape and Comprehensive Multi-omics Profiling across 40 AML samples.

(A) Sorting strategy for leukemic blasts. Four representative examples of sorting strategies and post sort analysis for the $CD33^{+}/CD34^{+}$ populations isolated in the present study. See also Table S1.

(B) Distribution of 20 driver mutations in our study (Yi *et al*.) and a large AML cohort (Papaemmanuil *et al*., 2016). Two APL patients were included in mutational analysis but excluded in the subsequent subtype classification. The frequency of gene mutations is calculated based on those samples genotyped by sequencing-based assay. Mutational profiling revealed high similarity in the frequency of each genetic disorder between the two studies, suggesting that AML patient composition in our study well represents the molecular heterogeneity of AML.

(C) Landscape of molecular aberrations in AMLs. Each dot represents the number of patients carrying this mutation, and those dots in upper or lower triangles mean co-occurrence of two mutations in the same sample. All genomic defects are grouped based on their molecular functions.

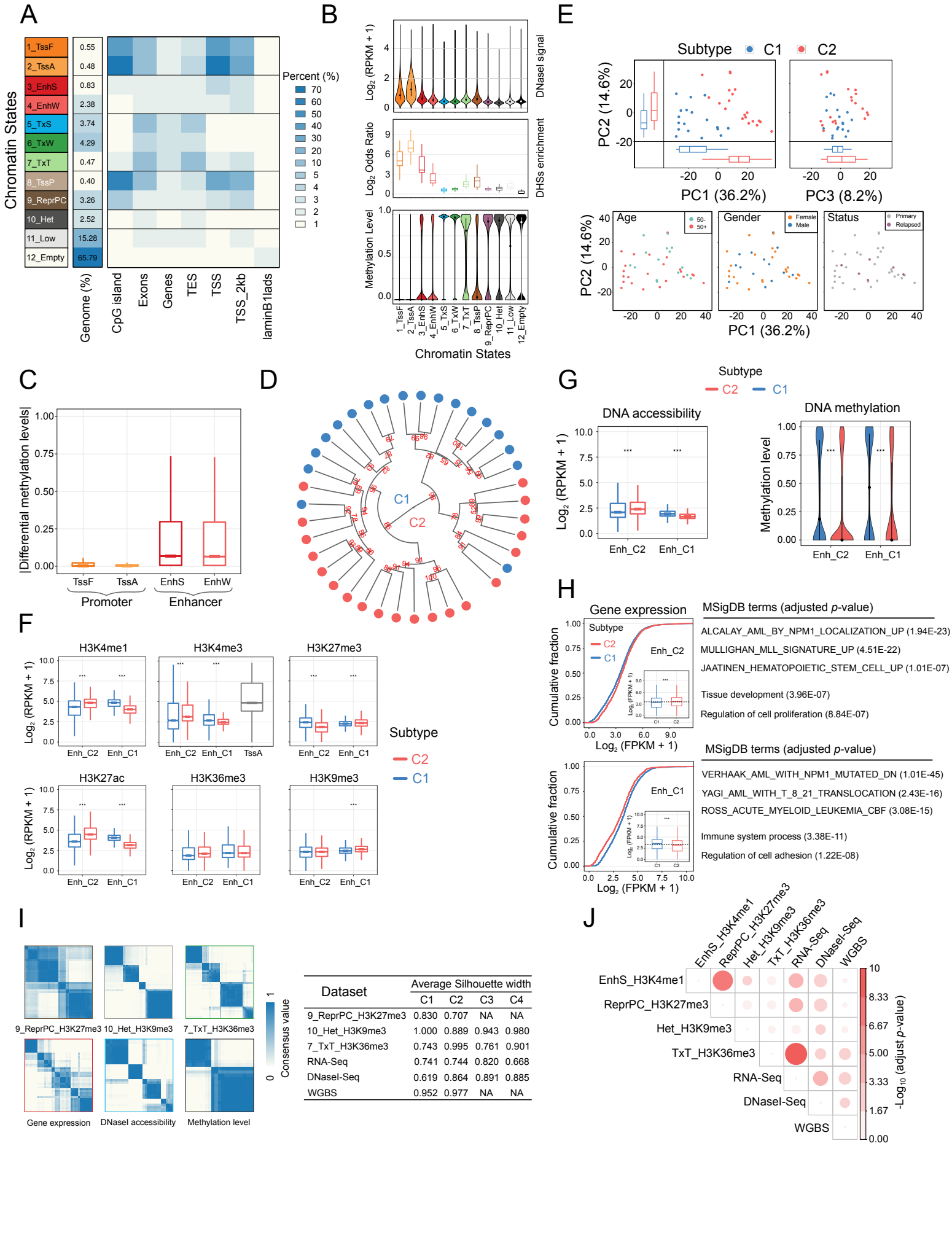
(D) Recurrent alteration spectrum for four genes with high frequency. Pairwise exclusivity and co-occurrence analysis are done using Fisher's exact test. In line with previous findings, our results revealed *NPM1* recurrence with *FLT3* and *DNMT3A*.

(E) Visual inspection of data quality by UCSC genome browser snapshots. All y-axis scales are RPKM-normalized units. Our experiments showed high signal-to-noise ratios, expected histone marking and accessibility (broad or peak shaped).

(F) Quality assessment by normalized strand cross-correlation coefficient (NSC) and relative strand cross-correlation coefficient (RSC) for six histone marks and DNA accessibility data. The red and blue dashed lines depict proposed thresholds respectively.

(G) The relationship between gene expression, DNA accessibility, six histone marks (H3K4me1, H3K4me3, H3K9me3, H3K27ac, H3K27me3 and H3K36me3) and DNA methylation level. Correlation between different datasets was evaluated by Spearman coefficient. The five active epigenetic marks (DNA accessibility, H3K4me1, H3K4me3, H3K27ac and H3K36me3) show greatest association with RNA-Seq.

(H) Correlation between gene expression and DNA accessibility, six histone marks (H3K4me1, H3K4me3, H3K9me3, H3K27ac, H3K27me3 and H3K36me3) and DNA methylation. All expressed genes are divided into four equal quartiles based on the expression levels.



Dataset	Average Silhouette width			
	C1	C2	C3	C4
9_ReprPC_H3K27me3	0.830	0.707	NA	NA
10_Het_H3K9me3	1.000	0.889	0.943	0.980
7_TxT_H3K36me3	0.743	0.995	0.761	0.901
RNA-Seq	0.741	0.744	0.820	0.668
DNaseI-Seq	0.619	0.864	0.891	0.885
WGBS	0.952	0.977	NA	NA

Figure S2, related to Figure 1. AML Epigenomes and Global Comparison among Subtypes from Different Marks.

(A) Average genomic coverage and annotation enrichment for each state in the AML population. Two promoter states, TssF (flanking active promoter) and TssA (active promoter), had higher coverage fractions at CpG island and transcription start site (TSS) regions than other states, and enhancer states (EnhS and EnhW) were found in intron and intergenic regions. The genomic coordinates for each feature are from UCSC database.

(B) DNaseI cleavage, DHS enrichment and DNA methylation level in each state. These four enhancer and promoter states displayed higher DNaseI enrichment/signal and lower DNA methylation level.

(C) Average absolute change of DNA methylation levels in the defined promoter (TssF and TssA) and enhancer (EnhS and EnhW) states.

(D) Circular dendrogram based on H3K27ac density in the EnhS state. Samples highlighted by blue correspond to putative C1 subtype, and samples in red indicate C2 subtype. Red numbers indicate approximately unbiased score from 1,000 bootstrap resampling.

(E) Unsupervised principal component analysis of top 1,000 most variable strong enhancer ranked by normalized H3K4me1 density. The first three principal components totally explain 59.0% variance. The panel below shows distribution of patient age, gender and clinical status in the two AML subtypes.

(F) Normalized profiles for six histone marks in 3,629 C1- (Enh_C1) and 4,400 C2-specific (Enh_C2) active enhancers. Asterisks indicate statistical significance (***$p < 0.001$).

(G) DNA accessibility and DNA methylation levels in differentially active enhancers between two subtypes. While for C2-active enhancers, DNA methylation is reduced in C2 and increased in C1, enhancers in C1 seem less variable although still significantly differentially methylated between C1 and C2 subtypes. Asterisks indicate statistical significance (***$p < 0.001$).

(H) Positive regulation and enriched functional terms of associated genes with subtype-specific enhancers.

(I) Independent consensus clustering results using three top dynamic states, RNA-Seq, DNaseI-Seq and WGBS data. A consensus score of 1 means that two items cluster together every time across all subsamples, whereas 0 means that two items never cluster together. Scaling is the same across all plots. The table (right) shows average silhouette values for each subtype identified from different datasets.

(J) Global comparison of AML classification based on different marks. Statistical significance of overlap among multiple clusters is assessed by Fisher's exact test, followed by Benjamini-Hochberg (B-H) correction.

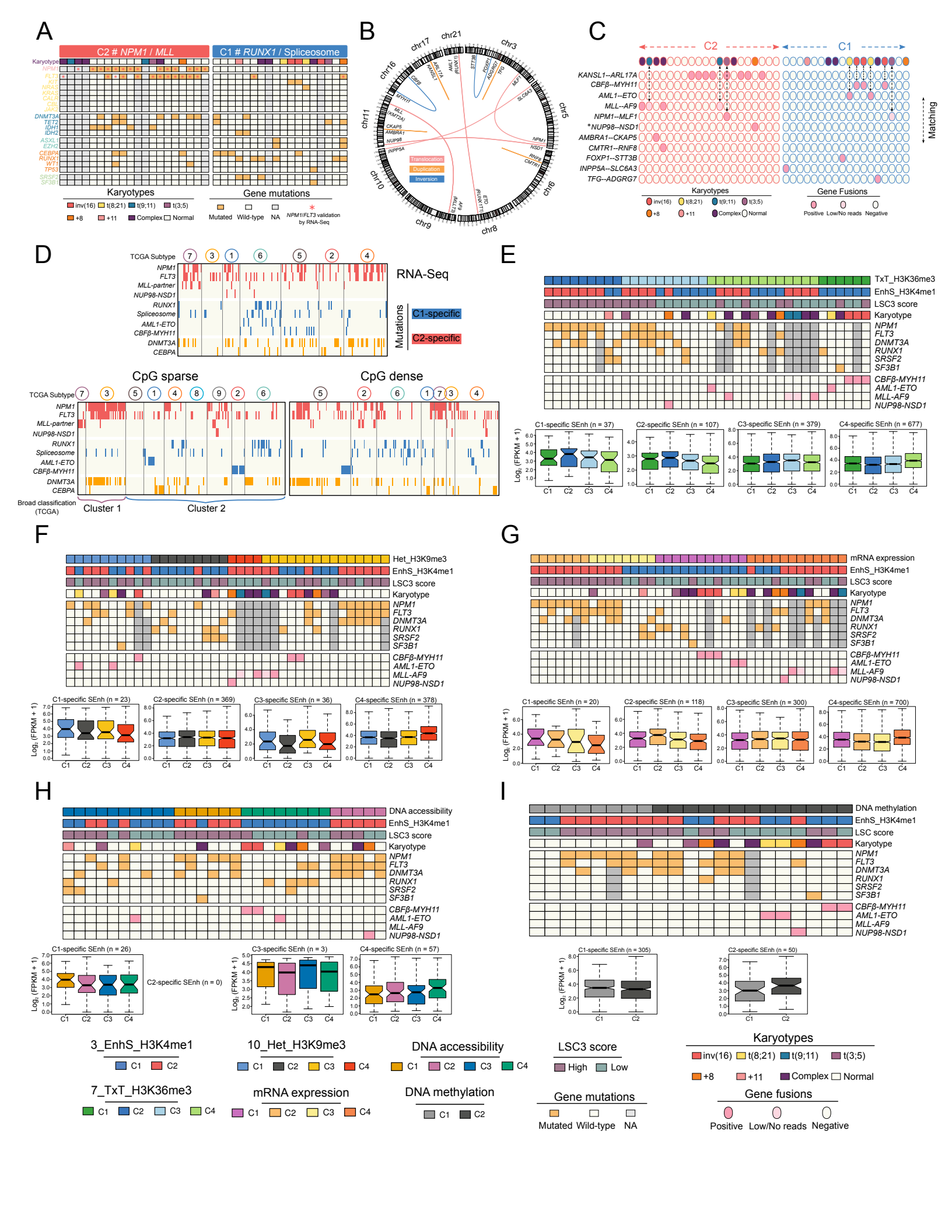

A
C2 # NPM1 / MLL
C1 # RUNX1 / Spliceosome
Karyotype
NPM1
FLT3
WT1
NRAS
KRAS
CALR
CBL
JAK2
DNMT3A
TET2
IDH1
IDH2
ASXL1
EZH2
CEBPA
RUNX1
WT1
TP53
SRSF2
SF3B1
Karyotypes
inv(16) t(8;21) t(9;11) t(3;5)
+8 +11 Complex Normal
Gene mutations
Mutated Wild-type NA
NPM1/FLT3 validation by RNA-Seq
B
Translocation
Duplication
Inversion
C
C2
C1
KANSL1--ARL17A
CBFβ--MYH11
AML1--ETO
MLL--AF9
NPM1--MLF1
*NUP98--NSD1
AMBRA1--CKAP5
CMTR1--RNF8
FOXP1--STT3B
INPP5A--SLC6A3
TFG--ADGRG7
Matching
Karyotypes
Gene Fusions
Positive Low/No reads Negative
D
TCGA Subtype
RNA-Seq
Mutations
C1-specific
C2-specific
CpG sparse
CpG dense
Broad classification (TCGA)
Cluster 1
Cluster 2
E
TxT_H3K36me3
EnhS_H3K4me1
LSC3 score
Karyotype
C1-specific SEnh (n = 37)
C2-specific SEnh (n = 107)
C3-specific SEnh (n = 379)
C4-specific SEnh (n = 677)
F
Het_H3K9me3
C1-specific SEnh (n = 23)
C2-specific SEnh (n = 369)
C3-specific SEnh (n = 36)
C4-specific SEnh (n = 378)
G
mRNA expression
C1-specific SEnh (n = 20)
C2-specific SEnh (n = 118)
C3-specific SEnh (n = 300)
C4-specific SEnh (n = 700)
H
DNA accessibility
C1-specific SEnh (n = 26)
C2-specific SEnh (n = 0)
C3-specific SEnh (n = 3)
C4-specific SEnh (n = 57)
I
DNA methylation
LSC score
C1-specific SEnh (n = 305)
C2-specific SEnh (n = 50)
3_EnhS_H3K4me1
10_Het_H3K9me3
DNA accessibility
LSC3 score
7_TxT_H3K36me3
mRNA expression
DNA methylation
Gene mutations
Karyotypes
Gene fusions

Figure S3, related to Figure 1. Comprehensive Landscapes of AML Subtypes Derived from Multiple Marks.

(A) Distribution and mutation status for all genomic lesions in classified AML subtypes. Samples in gray did not have target-exome sequencing data.

(B) Circos diagram depicting all fusion events identified from ultra-deep RNA-Seq data. These events are grouped into three categories based on the types of structural variations.

(C) Distribution patterns of fusion genes detected in AML samples. An up-down arrow indicates that the fusion event predicted from RNA-Seq can be validated by corresponding genomic data. * is reported to target HOX loci by interacting with MLL and nonspecific lethal (NSL) complexes, potentially explaining its clustering within this subgroup.

(D) Comparison between our subtype classification and AML clusters defined by TCGA RNA-Seq and DNA methylation data. To allow direct comparison of our ChIP-Seq-based subtypes with previous clusters defined by TCGA RNA-Seq or 450K DNA methylation analysis, we assigned the four mutations (*RUNX1*, spliceosome, inv(16) and t(8;21)) as C1-specific and another four aberrations (*NPM1*, *FLT3*, *MLL*-partner and *NUP98-NSD1*) as C2-specific. Examining the distribution of these mutations in the different TCGA defined clusters revealed several subclusters enriched for C1- or C2-specific mutations. Especially for the CpG sparse based clustering, it allowed broadly partitioning all samples into two groups, reminiscent of the chromatin-based clustering. Two broad groups shown in the DNA methylation section are from TCGA paper.

(E) - (I) (upper panels) Mutational spectrum and leukemia stem cell (LSC) score pattern within each cluster inferred from H3K36me3, H3K9me3, gene expression, DNA accessibility and DNA methylation levels separately. (lower panels) Association between involved genes and subtype-specific super enhancers (SEnh). Those associated genes are positively regulated by SEnh independent of marks used for AML classification.

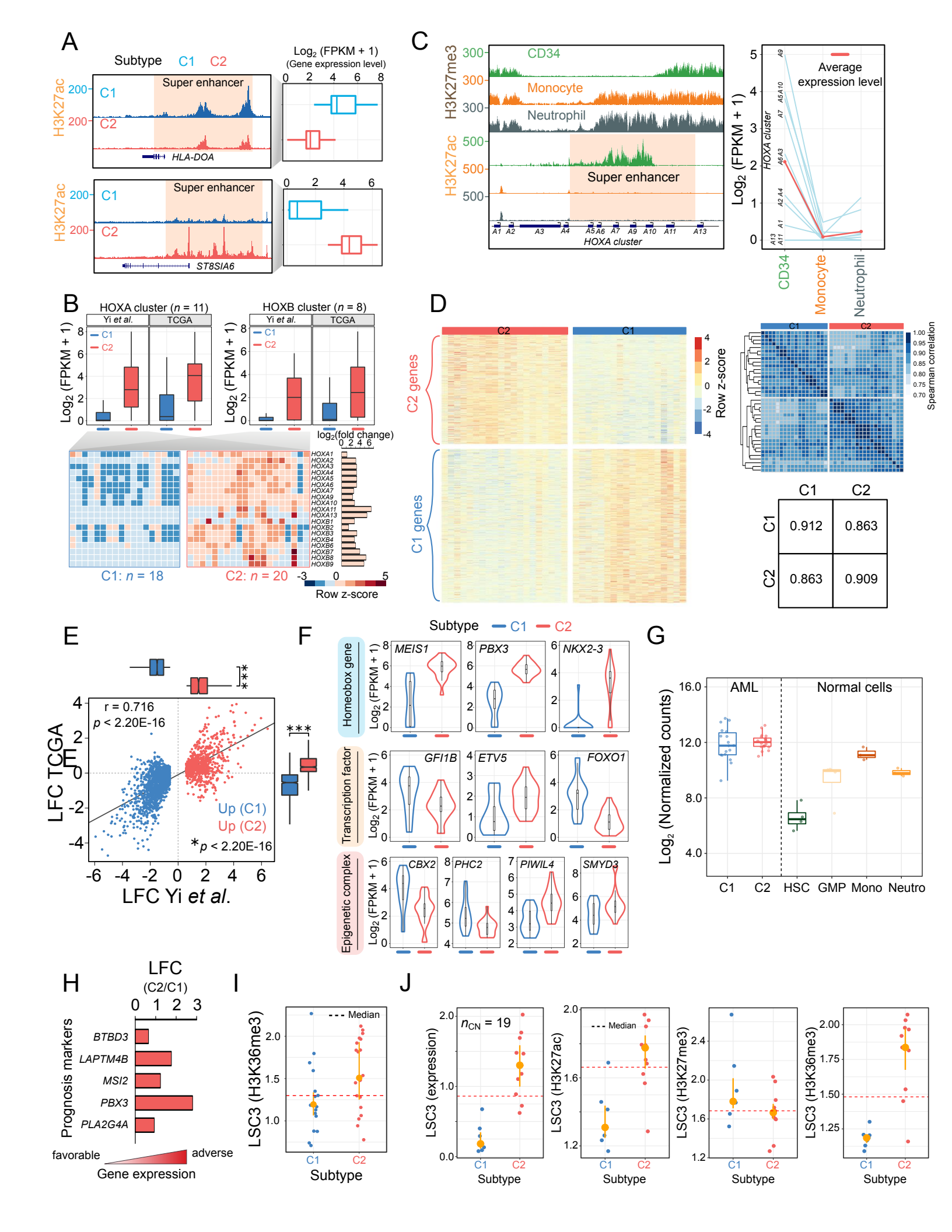

A
Subtype C1 C2
Log2 (FPKM + 1)
(Gene expression level)
H3K27ac
Super enhancer
HLA-DOA
ST8SIA6
B
HOXA cluster (n = 11)
HOXB cluster (n = 8)
Yi et al.
TCGA
Log2 (FPKM + 1)
log2(fold change)
C1: n = 18
C2: n = 20
Row z-score
C
H3K27me3
CD34
Monocyte
Neutrophil
Super enhancer
HOXA cluster
Average expression level
D
C2 genes
C1 genes
Row z-score
Spearman correlation
E
r = 0.716
p < 2.20E-16
Up (C1)
Up (C2)
*p < 2.20E-16
LFC TCGA
LFC Yi et al.
F
Subtype C1 C2
Homeobox gene
Transcription factor
Epigenetic complex
MEIS1
PBX3
NKX2-3
GFI1B
ETV5
FOXO1
CBX2
PHC2
PIWIL4
SMYD3
G
Log2 (Normalized counts)
AML
Normal cells
C1 C2 HSC GMP Mono Neutro
H
LFC
(C2/C1)
Prognosis markers
BTBD3
LAPTM4B
MSI2
PBX3
PLA2G4A
favorable
adverse
Gene expression
I
LSC3 (H3K36me3)
Median
Subtype
J
LSC3 (expression)
LSC3 (H3K27ac)
LSC3 (H3K27me3)
LSC3 (H3K36me3)
n_CN = 19
Median
Subtype

Figure S4, related to Figure 3 and 4. Subtype-specific Super Enhancers (SEs), Chromatin States, Transcriptional Levels and Clinical Conseqeunces.

(A) UCSC tracks and expression levels of two representative genes.

(B) Transcriptional levels of each *HOXA* gene in the two subtypes proposed by our study and TCGA project.

(C) Dynamic H3K27ac and H3K27me3 intensity as well as expression levels of *HOXA* families in normal $CD34^{+}$ progenitors, monocytes and neutrophils.

(D) The homogeneity of gene expression patterns within the same AML subtypes. The intra-subtype samples display more similar transcriptome profiles and higher Spearman correlation values.

(E) Correlation with the TCGA project shows consistent expression patterns for the differentially expressed genes derived from our study. LFC: $\log_2$ fold change. Three asterisks mean p-value < 0.001.

(F) Subtype-specific expression levels for homeobox families and cofactors, transcription factors and enzyme complexes.

(G) Expression of CEBPA transcription factors in two AML subtypes and four normal cell types. Transcriptional levels were normalized by DESeq2 package.

(H) Several known gene markers associated with clinical prognosis. Over-expression of these genes is significantly correlated with poor outcomes. LFC: $\log_2$ fold change.

(I) Leukemic stem cell score derived from 3 signature genes using H3K36me3 for two subtypes.

(J) Distribution of LSC3 score inferred from 19 cytogenetically normal AML samples.

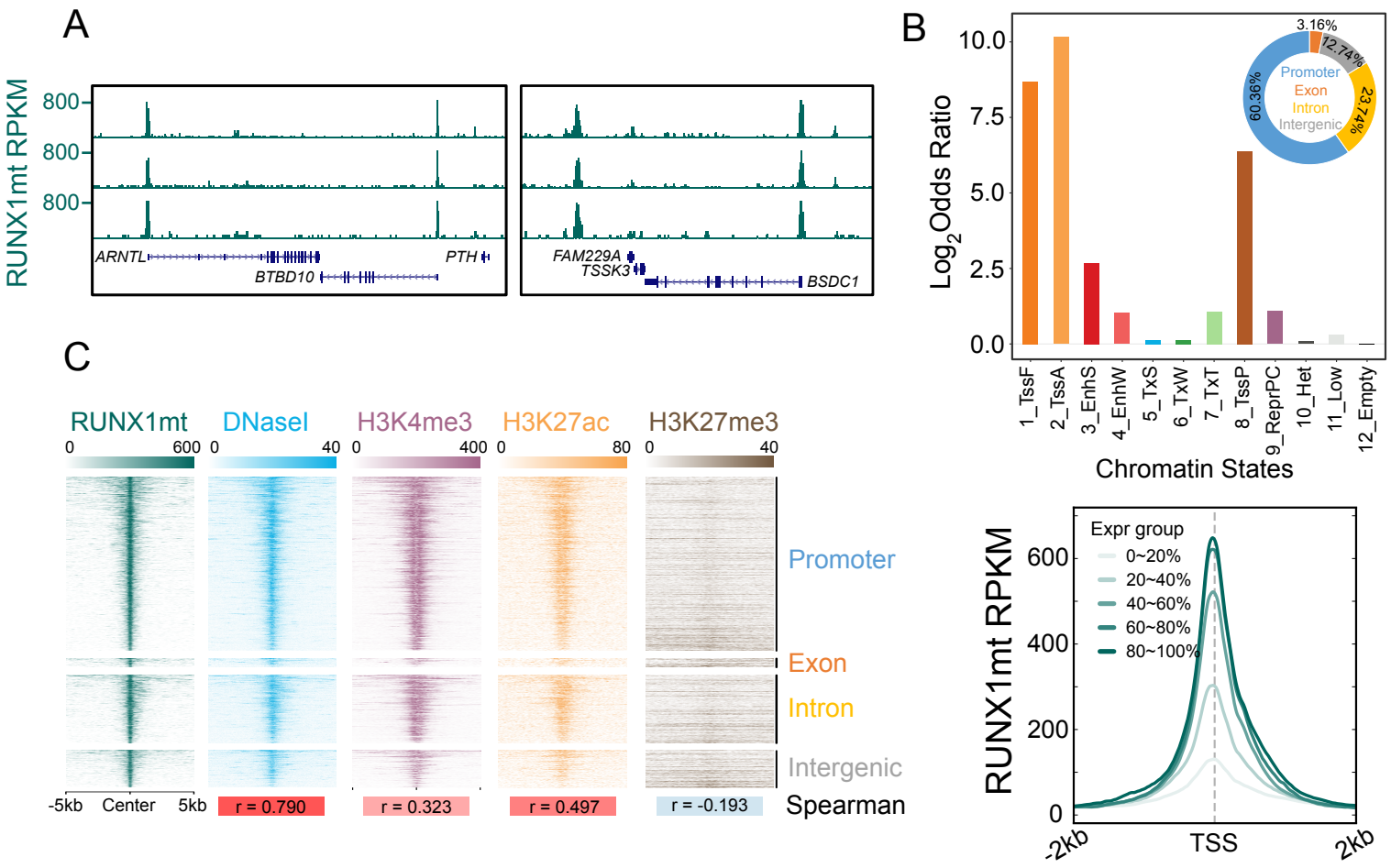


Figure S5, related to Figure 4. Genome-wide Binding Profiling of RUNX1 Protein in RUNX1 Mutant AMLs.

(A) High-quality binding patterns of RUNX1 ChIP-Seq data at several representative genes.

(B) RUNX1 occupancy at different chromatin states (upper), and its regulatory effect on gene expression (lower). All peaks are assigned to promoters (2 kb away from TSS), exons, introns and intergenic regions based on the physical distance. The priority rule of category assignment: promoter > exon > intron > intergenic region. All expressed genes are grouped into five classes based on expression levels.

(C) Co-localization of RUNX1 with DNA accessibility, H3K4me3, H3K27ac and H3K27me3 in promoter, exon, intron and intergenic region respectively. Corresponding spearman correlation for each mark is shown under the heatmap.

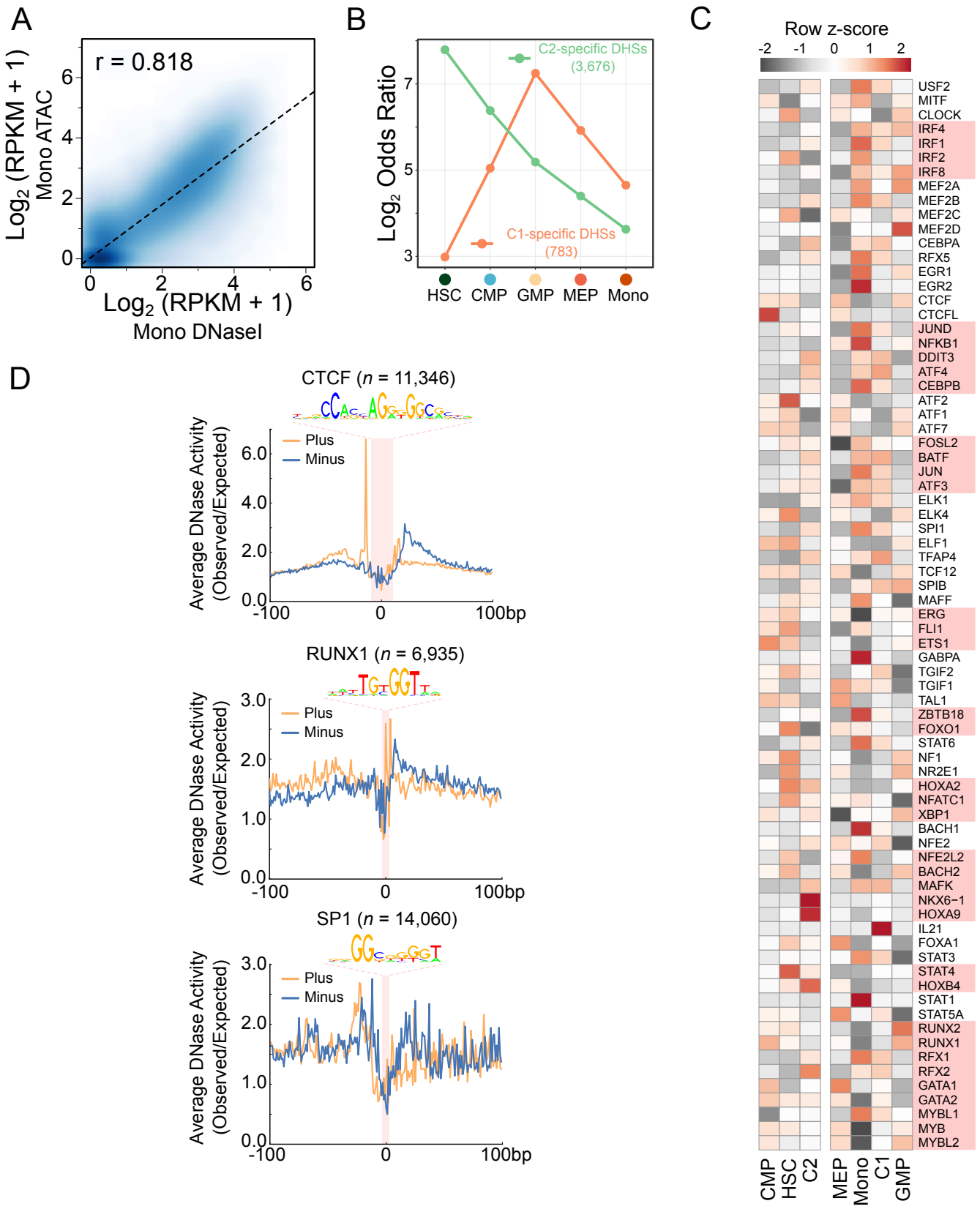


Figure S6, related to Figure 5. Chromatin Accessibility Signatures and Predicted Footprint Profiles based on DNaseI-Seq Data.

(A) Correlation between DNaseI-Seq data and ATAC-seq data in monocyte cells.

(B) Genomic overlap with signatures of five normal cell types for differentially regulated DNA hypersensitivity sites in each subtype.

(C) Expression patterns of corresponding transcriptional factors for identified motifs in two AML subtypes and four normal cell types.

(D) Putative footprint profiles for three transcription factors, CTCF, RUNX1 and SP1 after merging DNaseI-Seq data from patients in the same subtype.

Table S2, related to Figure 1. Detailed Information of 21 Targeted Genes for Mutation Detection.

Target genes	No. Exons targeted	Base-pairs targeted
ASXL1	12	2,912
CALR	9	205
CBL	8, 9	340
CEBPA	1	1,081
DNMT3A	8, 9, 13-15, 18-23	1,299
EZH2	2-20	2,242
FLT3	14, 15, 20	373
IDH1	4	296
IDH2	4	165
JAK2	12, 14	220
KIT	1-21	2,916
KRAS	2, 3	293
MPL	10	101
NPM1	12	43
NRAS	2, 3	292
RUNX1	2-9	1,489
SF3B1	12, 14-16	748
SRSF2	1	366
TET2	2-11	6,069
TP53	4-11	1,086
WT1	7, 9	247

Table S3, related to Figure 1. Global Landscape of Gene Mutations in Acute Myeloid Leukemia (AML).

Functional Group	Gene Name	NO. of Samples	Point Mutation*	INDEL
Shuttling protein	*NPM1*	13		c.863_864insTATG; c.863_864insCCTG; c.860_863dupTCTG
Activated Signaling	*FLT3*	13	c.2503G>C	c.1745_1780dupCGGCTCCTCAGATAATGAGTACTTCTACGTTGATTT; c.1769_1795dupTCTACGTTGATTTCAGAGAATATGAAT; c.1794_1795insGAATTTGGGGCTGCTTTTTTT; c.1756_1785dupGATAATGAGTACTTCTACGTTGATTTCAGA; c.1770_1793dupCTACGTTGATTTCAGAGAATATGA
	KIT	*2*	*c.2446G>T*	c.1249_1255delACTTACG c.1252insT
	NRAS	*1*	*c.34G>T*	
DNA Methylation	*DNMT3A*	9	c.2440G>T; c.2204A>G; c.2645G>A; c.2644C>T; c.2339T>C	
	TET2	5	c.1259C>A	c.2747delA; c.444dupA; c.3025_3028delCAAG; c.3880_3906dupTACTACAATGGATGTAAGTTTGCCAGA
	IDH1	5	c.395G>A; c.394C>A; c.394C>T; c.395G>T; c.394C>G	
	IDH2	2	c.419G>A	
Histone Modifiers	*ASXL1*	2	c.2077C>T	c.1934dupG
	EZH2	2	c.2084C>G; c.2197T>C	
Transcription Factors	*CEBPA*	7	c.898C>A; c.274A>T	c.247delC; c.949_950insGTC; c.937_939dupAAG; c.564_566delGCC; c.697_703dupACCCCGC
	RUNX1	6	c.724+1G>A; c.1103C>T; c.518C>G; c.235T>G; c.416G>A; c.415C>T	c.1109_1122delAAGCCAGCTCGCCC
	WT1	2		c.1108delC; c.1142dupC; c.1110dupT; c.1132_1139dupTGTACGGT
	TP53	1	c.536A>G	
Spliceosome Complex	*SRSF2*	3	c.284C>A	c.283CC>TT
	SF3B1	1	c.1873C>T	

*All positions are indexed to hg19 reference genome