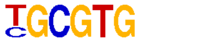
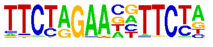
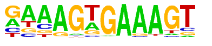
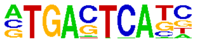
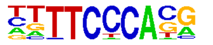
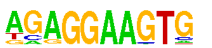
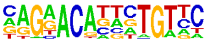
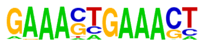
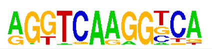
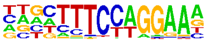
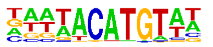
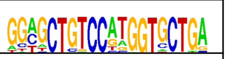
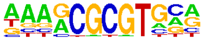
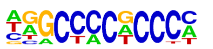
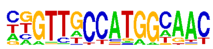
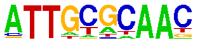
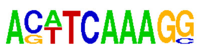
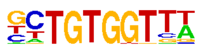
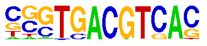
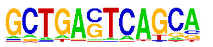
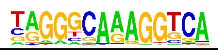
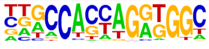
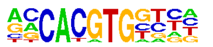
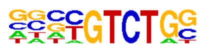
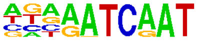
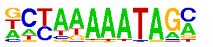
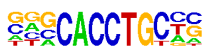
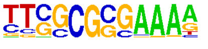
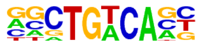
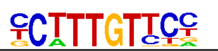
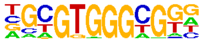
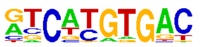
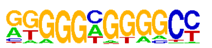
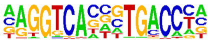
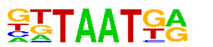
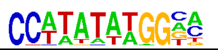
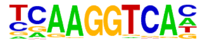
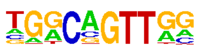
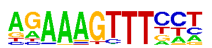
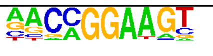
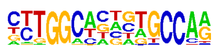
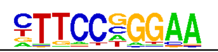
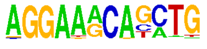
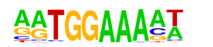
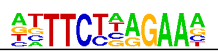
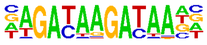
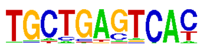
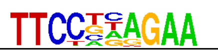
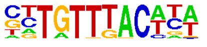
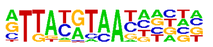
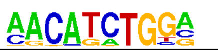
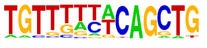
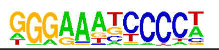
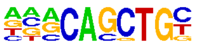
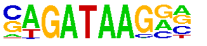
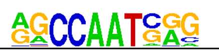
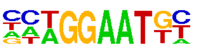
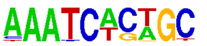
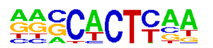
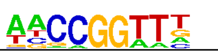
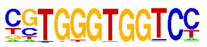
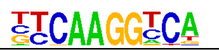
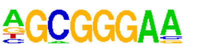
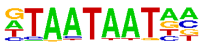
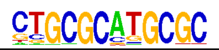
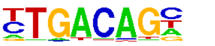
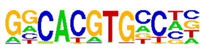
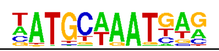
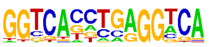
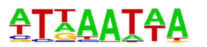
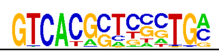
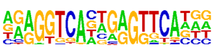
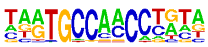
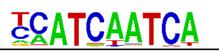
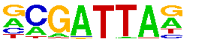
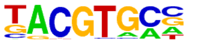
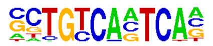
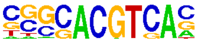
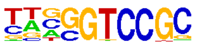
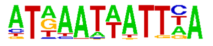
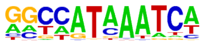
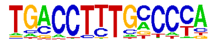
SUPPLEMENTARY MATERIALS

1. SUPPLEMENTARY FIGURES AND TABLES

patient code	Signalling	NPM1	Chrom	RUNX	Other mutations	DHS Seq	RNA Seq	Age	Sex	wbc	case
ITD-1	FLT3-ITD				DNMT3A, TET2x2, BCOR, TP53	Y	Y	45	F	56	Rel
ITD-2	FLT3-ITD		tri(13)		DNMT3A, TET2	Y	Y	68	F	2	Pres
ITD-3	FLT3-ITD				DNMT3A	Y	Y	80	F	143	Pres
ITD/NMP1-1	FLT3-ITD	NPM1			DNMT3A, WT1	Y	Y	45	F	32	Pres
ITD/NPM1-2	FLT3-ITD	NPM1				Y	Y	61	F	7	Rel
ITD/NPM1-3	FLT3-ITD	NPM1				Y	Y	66	F	91	Pres
ITD/NPM1-4	FLT3-ITD	NPM1			GATA2, DNMT3A	Y	Y	65	F	21	Pres
ITD/NPM1-5	FLT3-ITD	NPM1			DNMT3A, BCOR	Y	Y	68	M	190	Pres
ITD/NPM1-6	FLT3-ITD	NPM1			WT1, DNMT3A, TET2, PHF6	Y	Y	58	F	195	Pres
NPM1-1		NPM1			IDH1	Y	Y	37	M	60	Pres
NPM1-2		NPM1			DNMT3A, TET2x2	Y	Y	75	M	94	Pres
t(8;21)-1			t(8;21)		TET2	Y	Y	72	M	29	Pres
t(8;21)/KIT-2	KIT		t(8;21)		NOTCH1	Y	Y	48	M	36	Pres
t(8;21)-3	FLT3-TK		t(8;21)			Y	Y	53	M	6	Pres
t(8;21)-4			t(8;21)			Y	Y	45	M	2	Pres
inv(16)-1	KIT		inv(16)			Y	Y	40	M	22	Pres
inv(16)-2			inv(16)			Y	Y	26	M	63	Pres
inv(16)-3			inv(16)		ASXL1	Y	Y	75	M	54	Pres
RUNX1-DT-1	FLT3		tri(13)	RUNX1	CREBBP, DNMT3A, SF3B1	Y	Y	68	M	112	Rel
RUNX1-DT/CEBPA-2	FLT3-ITD			RUNX1	CEBPA, WT1x2, SF3B1, TP53	Y	Y	83	M	68	Pres
RUNX1-DT-3				RUNX1		Y	Y	58	M	37	Pres
RUNX1(x2)-D&T-4				RUNX1x2	SRSF2, DNMT3A, IDH2	Y	Y	82	M	55	Pres
RUNX1-D-5				RUNX1	IDH1, BCORL1x2, SRSF2x2	Y	Y	65	M	8	Pres
RUNX1-T/CEBPA-6	NRAS		tri (8)	RUNX1	CEBPA, EZH2	Y	Y	75	M	107	Pres
CEBPA(x2)-1					CEBPAx2	Y	Y	76	F	238	Pres
CEBPA(x2)-2					CEBPAx2, GATA2	Y	Y	21	F	10	Pres
CEBPA(x2)-3					CEBPAx2, GATA2, TET2	Y	Y	75	M	106	Pres
ITD(2x)/NPM1-1	FLT3-ITDx2	NPM1			DNMT3A, IDH2	Y	Y	78	F	26	Pres
ITD(2x(/NPM1-2	FLT3-ITDx2	NPM1			CEBPA, IDH2	Y	Y	72	F	68	Pres
NPM1/RAS-3	NRAS	NPM1			PTPN11, DNMT3A, IDH1	Y	Y	30	F	4	Rel
inv(3)/RAS-3	NRAS		inv(3)		ETV6, SF3B1	N	Y	54	M	104	Pres
inv(3)/RAS-1	NRAS		inv(3)		GATA2, SF3B1	Y	Y	59	M	4	Rel
inv(3)/CBL-2	CBL		inv(3)		SF3B1	Y	N	34	F	21	Rel
t(8;21)/ITD(x2)-5	FLT3-ITD		t(8;21)		SMC1A	N	Y	43	M	86	Pres
RUNX1-D/JAK-1	JAK2		tri (21, 9)	RUNX1	IDH2, SRSF2	Y	Y	79	M	12	Pres
RUNX1-T/JAK-2	JAK2			RUNX1	TET2x2, TP53	Y	Y	77	F	79	Pres
RUNX1-T-7 (NHL)			tri (21)	RUNX1	TET2x2, PHF6	Y	Y	73	F	NA	Pres
CEBPA-5					CEBPA, DNMT3A	N	Y	79	F	40	Pres
SRSF2-1					IDH2, SRSF2	N	Y	67	M	2	Pres
SRSF2-2					SOCS1, DNMT3A, IDH2, SRSF2	N	Y	71	M	2	Pres
t(8;21)-1R	KIT		t(8;21)		TET2	Y	Y	72	M	29	Rel

Supplementary Table 1: Patient groups, mutation data, and clinical data. Patient codes depicted in color represent samples included in the seven major defined mutation groups, or which have either 2 FLT-ITD mutations or a mutation in either CBL or NRAS. This table also indicates samples where DNaseI-Seq and RNA-Seq data are either available (Y) or not available (N). Further details can be found in Supplementary Data-set 1.

Logos of position weight matrices used for motifs shared within transcription factor families

motif	logo	motif	logo	motif	logo
AHR		HSF1		PRDM1	
AP-1		IKZF		PU.1	
AR		IRF		RAR	
BCL6		IRX		REST	
CAMTA		KLF		RFX	
C/EBP		LEF1		RUNX	
CREB/ATF		MAF		RXR	
CTCF		MYC/MAX		SMAD	
CUT		MEF2		SNAI	
E2F		MEIS		SOX	
EGR		MITF		Sp1	
ESR1		MNX1		SRF	
ESRRA		MYB		ST18	
ETS		NF1		STAT3	
ETS:E-box		NFAT		STAT5	
EVI1		NFE2		STAT6	
FOX		NFIL3		TAL1	
FOX:E-box		NF-kB		TCF3	
GATA		NFY		TEAD	
GFI1B		NKX		TFCP2	
GLI		NR		TFDP1	
HBP1		NRF1		TGIF	
HES		OCT		THR	
HHEX		PAX5		VDR	
HIC1		PBX		VENTX	
HIF1A		PKNOX1		XBP1	
HINFP		POU4F1			
HOX		PPAR			

Supplementary Table 2: List of representative position weight matrices for TF families.

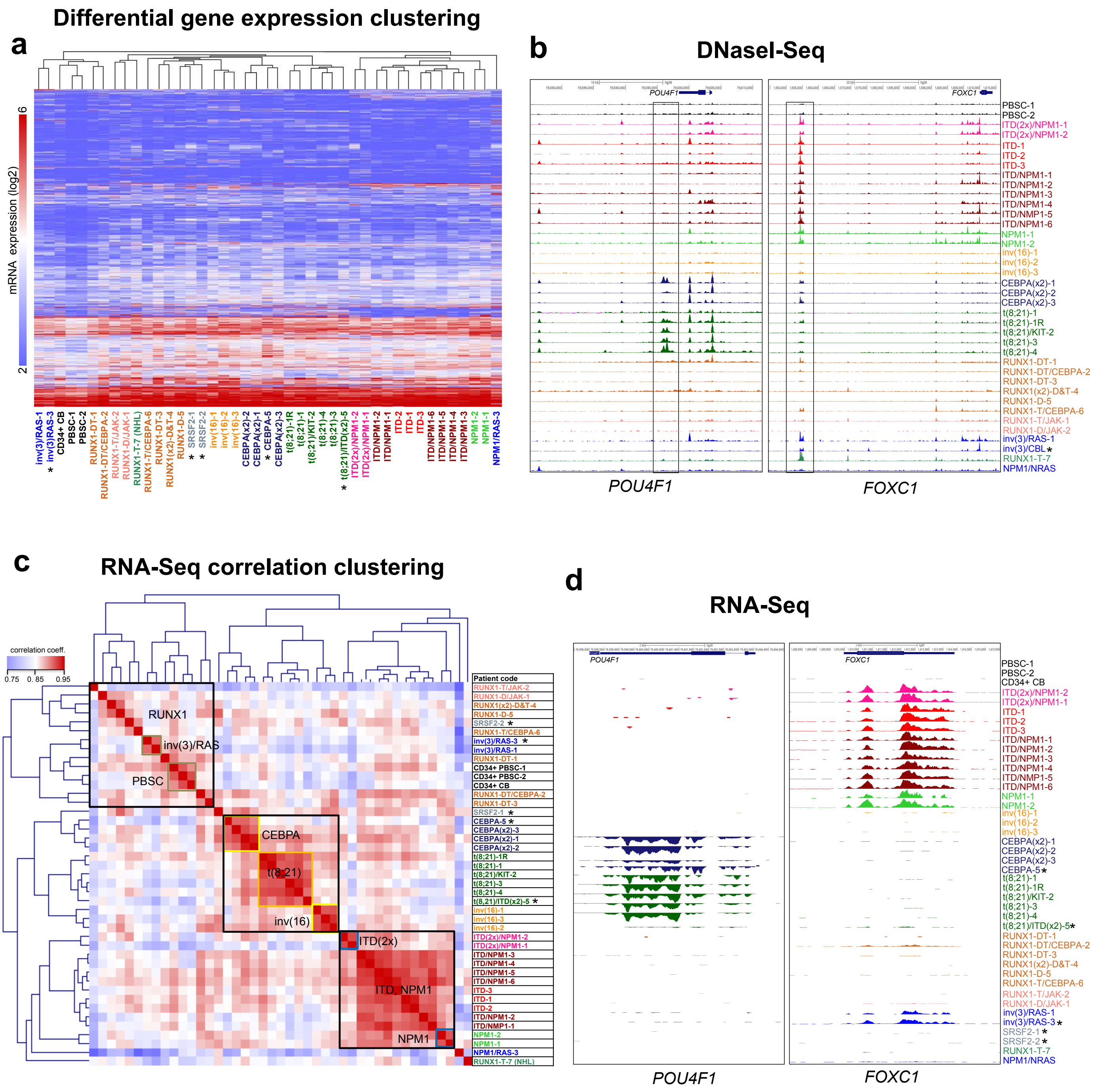
To improve the process of linking regulatory factors with their binding sites on DNA, we consolidated the different versions of transcription factor consensus binding sequences for closely related family members where the motif signatures are indistinguishable. For most transcription factor families there are typically various alternate subtly different versions of position weight matrices for not just different family members but also for the same factor from different data sets.

The prevalence of so many different related consensus sequences is a major impediment to the construction of regulatory networks from genome-wide analyses of DNA elements. For the current study, we first identified a subset of almost 300 transcription factor genes that are expressed in one or more of our AML samples. We then inspected the motifs listed on either the HOMER or JASPER databases (see URLs), motifs defined in a recent large-scale study of recombinant proteins[66], or motifs described in various other publications. We grouped together those factors where the motifs are essentially the same, and chose the best representative example for further analysis. These selections were often validated by referring to the large body of literature which is devoted to defining specific motifs, which also informed the choices of which orientation of motifs represented the conventional form used in publications. The JASPAR motifs were viewed via their web link (see URL section).

shRNA	**Target sequence**
shFOXC1_B	GTCACAGAGGATCGGCTTGAA
shFOXC1_C	GCCGCACCATAGCCAGGGCTT
shNFIX_B	GGAATCCGGACAATCAGAT
shNFIX_C	GCAGTCTCAGTCCTGGTTCCT
shPOU4F1_C	GCCGAGAAACTGGACCTCAAA
shPOU4F1_E	GCCGATTAACAAGACTGAAAT
shMM	GCGCGATAGCGCTAATAATTT

Supplementary Table 3: List shRNA target sequences.

Target sequences of shRNA sequences used in validation experiments.

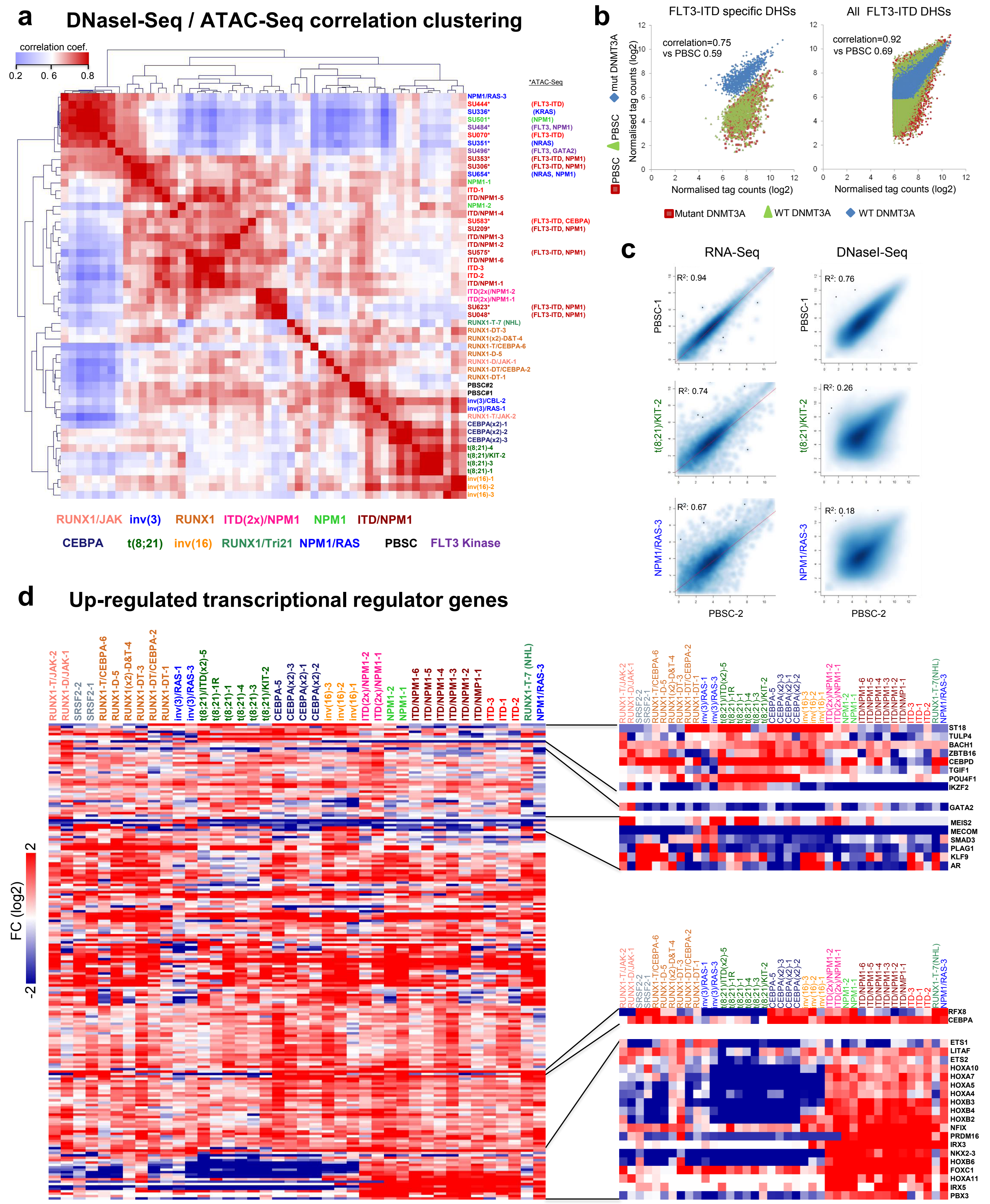


Supplementary Figure 1: Different types of AML adopt unique transcriptomes. (a) Hierarchical clustering of gene expression as determined by RNA-Seq of all patient samples. Clustering of log2 FPKM values for all differentially expressed genes changing expression at least 2 fold in at least one patient as compared to normal CD34+ PBSC.

(b) UCSC genome browser screenshots of DNaseI-Seq in all AML patients with different classes of mutations and normal CD34+ PBSC at *POU4F1* (left panel) and *FOXC1* (right panel) locus.

(c) Hierarchical clustering of Pearson correlation coefficient between all patient samples of RNA-Seq data: (left panel), right panel: list of mutations in cells from each patient. The correlation between any two patients was obtained with log2 FPKM expression values over all genes.

(d) UCSC genome browser screenshots of RNA-Seq reads in AML patients at *POU4F1* (left panel) and *FOXC1* (right panel) locus. Asterisks denote samples for which the matching RNA-Seq or DNaseI-Seq data are unavailable.

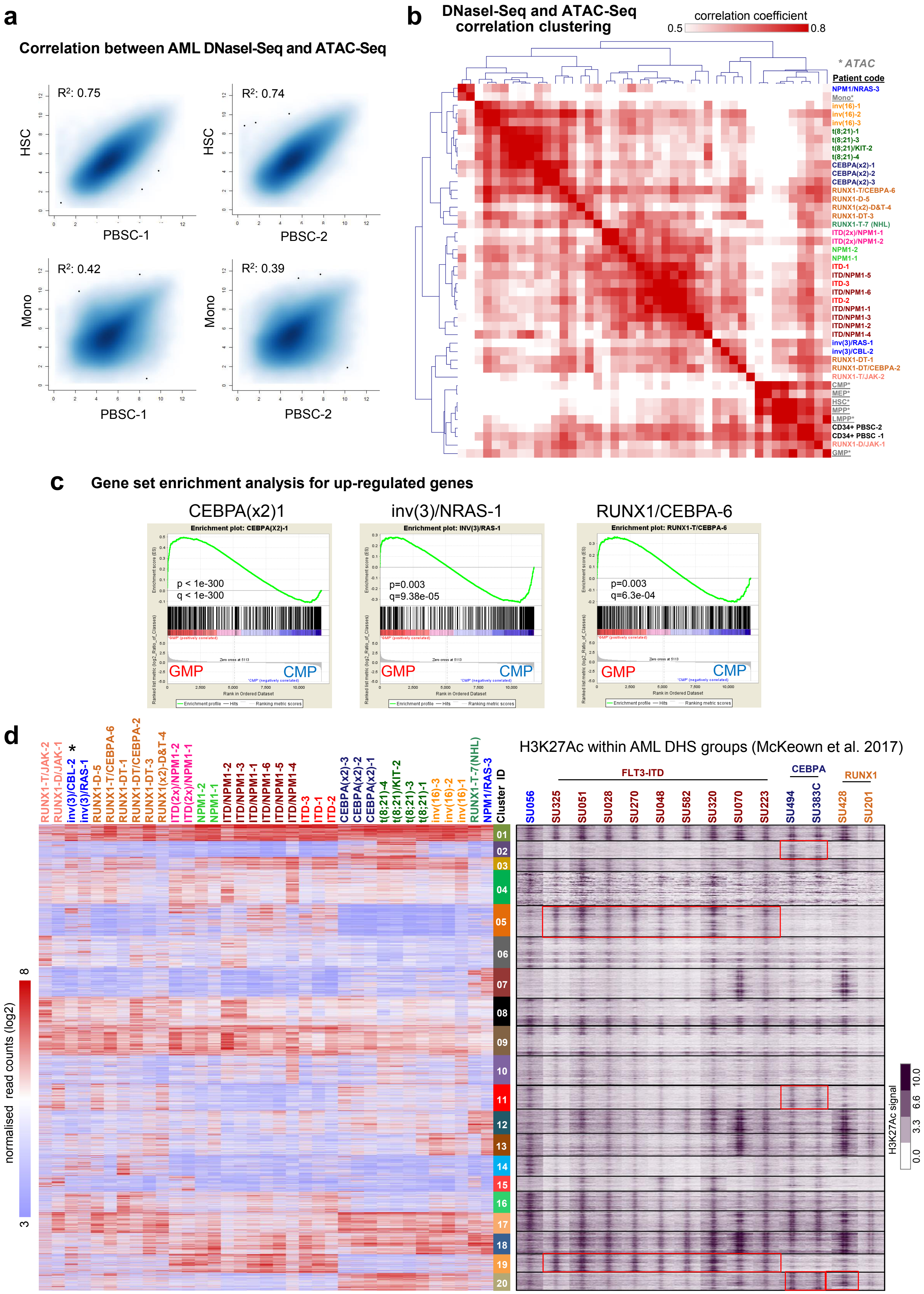


Supplementary Figure 2: Different types of AML adopt unique transcriptome and chromatin landscapes. (a) Hierarchical clustering of Pearson correlation coefficients of DNaseI accessible sequences from all our patient samples with normalized read counts of DNase-Seq data for the different classes of mutations also including ATAC-Seq data from Corces et al.[67], with similar mutations (SU(nnn), mostly FLT3-ITD). The mutation class is highlighted to the right of the panel and by a color code below the heatmap, again showing that specific elements from specific AML-types cluster together. Note the tight clustering of FLT3 and RAS mutant AML.

(b) Scatter plots comparing the DNaseI tag count signals of patients with (11) and without (8) DNMT3 mutations against each other and against PBSCs as indicated by colored shapes.

(c) Smooth scatter plots showing the correlation between DNase-Seq and RNA-Seq data from AML patients. Shown are CD34$^+$ PBSC cells from individual #1 versus individual #2 (left plot), CD34+ PBSC from individual #2 versus a patient with NPM1 and NRAS mutation (right plot). RNA-Seq plots (top panel) and DNaseI-Seq plots (bottom panel). The coefficients of determination (R-squared) highlighting the significance of correlation are shown on each panel. Other comparisons can be retrieved from the webserver.

(d) Hierarchical clustering of log2 gene expression fold difference for all differentially transcription factor (TFs) and transcriptional regulator genes changing expression at least 2 fold in at least one patient as compared to normal CD34+ PBSC. Clustering was done only on rows (i.e., genes) while samples were ranked based on the clustering in **Fig.1c.** The heatmap colour is related to the degree of differential expression (fold-change (FC)). Red is up-regulated compared to normal CD34+ and blue is a down regulated TF.

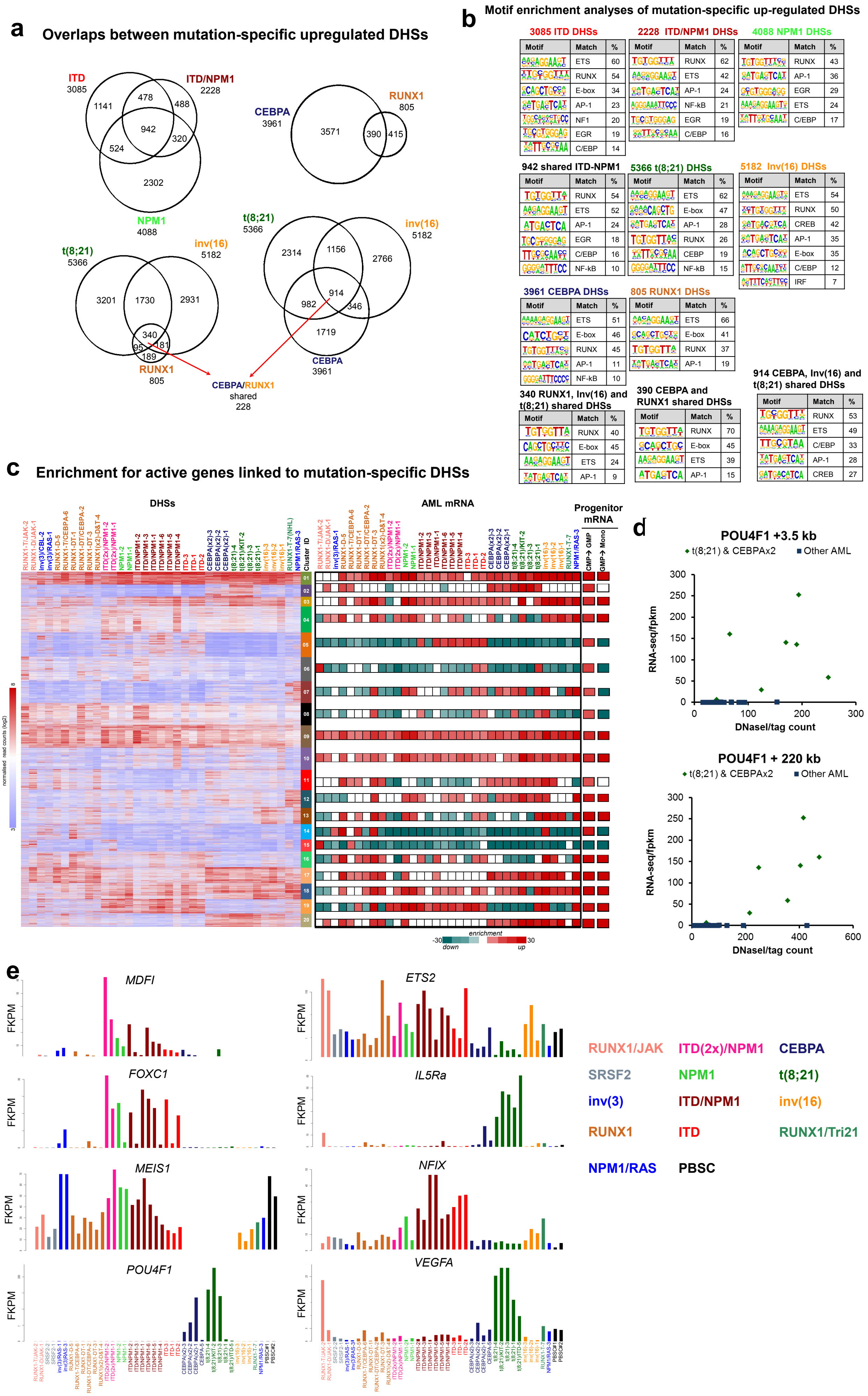


Supplementary Figure 3: Different types of AML are blocked at different stages of differentiation and correlation with publicly available data-sets. (a) Smooth scatter plots show the correlation between AML DNase-Seq and ATAC-Seq data. Top panel shows the DNAseI-Seq from normal CD34+ PBSC patient #1 & #2 versus the ATAC-Seq from Hematopoietic stem cells (HSC) and lower panel shows the DNaseI-Seq from normal CD34+ PBSC patient #1 & #2 versus the ATAC-Seq from Monocytes (Mono). The coefficients of determination (R-squared) highlighting the significance of correlation are shown on each panel.

(b) Hierarchical clustering of Pearson correlation coefficient between DNaseI-Seq data from all patient samples together with ATAC-seq data from Corces et al.[66] The correlation between any two patients was obtained with normalized read counts calculated with +/- 200 bases from the peak summit.

(c) Gene set enrichment analysis for the up-regulated genes that are at least 2 fold difference compared to the normal CD34+. The AML up-regulated genes were tested for enrichment against the common myeloid progenitors (CMP) versus Granulocyte Macrophage Precursors (GMP) taken from Corces et al[66] RNA-seq data. p and q(FDR) values highlighting the significance of enrichment are shown on each panel.

(d) Heatmap showing density enrichment of H3K27Ac peaks from McKeown et al., 2017 ranked according to the same coordinates of the DNase-Seq within the clusters (left heatmap), the H3K27Ac densities were plotted with a window size of +/- 2 kb around the DNaseI-Seq peaks summit. Selected AML-specific blocks of peaks are highlighted. The asterisk highlights samples inv(3)/CBL-2 for which RNA-Seq data are unavailable.



b Motif enrichment analyses of mutation-specific up-regulated DHSs

3085 ITD DHSs

Motif	Match	%
AAGAGGAAGT	ETS	60
TTGCGGTTT	RUNX	54
GCAGCTGCCA	E-box	34
GATGAGTCAT	AP-1	23
TGGCAGACTGCC	NF1	20
TGCGTGGGAG	EGR	19
TATTGCGCAA	C/EBP	14

2228 ITD/NPM1 DHSs

Motif	Match	%
TGTGGTTT	RUNX	62
AAGGAGGAAGT	ETS	42
GATGAGTCAT	AP-1	24
AGGGAAATTCCC	NF-kB	21
TGCGTGGGAG	EGR	19
GGTTGCGCAA	C/EBP	16

4088 NPM1 DHSs

Motif	Match	%
TGTGGTTTCG	RUNX	43
GATGAGTCAT	AP-1	36
GCGTGGGCGG	EGR	29
AAAGAGGAAGTG	ETS	24
TATTGTGCAAT	C/EBP	17

942 shared ITD-NPM1

Motif	Match	%
TGTGGTTA	RUNX	54
AAGAGGAAGT	ETS	52
ATGACTCA	AP-1	24
TGCGGGGGAG	EGR	18
TTGCGCAACC	C/EBP	16
GGGGATTTCC	NF-kB	10

5366 t(8;21) DHSs

Motif	Match	%
AAGAGGAAGT	ETS	62
GAAACAGCTG	E-box	47
GATGAGTCAT	AP-1	28
TGTGGTTAC	RUNX	26
GGTTGCGCAA	CEBP	19
GGGGATTTCC	NF-kB	15

5182 Inv(16) DHSs

Motif	Match	%
AAAGAGGAAGTG	ETS	54
TCTGTGGTTT	RUNX	50
GATGACGTCA	CREB	42
GATGAGTCAT	AP-1	35
ACAGCTGCGG	E-box	35
ATTGCGCAATCT	C/EBP	12
AGTTTCAGTTCC	IRF	7

3961 CEBPA DHSs

Motif	Match	%
AAAAGAGGAAGT	ETS	51
CATCTGCT	E-box	46
TGTGGTTTCG	RUNX	45
AATGAGTCAT	AP-1	11
GGGGATTTCCCC	NF-kB	10

805 RUNX1 DHSs

Motif	Match	%
AACAGGAAGT	ETS	66
GCAGCTGCTG	E-box	41
TGTGGTTA	RUNX	37
TATGAGTCAT	AP-1	19

340 RUNX1, Inv(16) and t(8;21) shared DHSs

Motif	Match	%
TGTGGTTA	RUNX	40
CAGCTGCTTC	E-box	45
AAGAGGAAGT	ETS	24
TATGAGTCAT	AP-1	9

390 CEBPA and RUNX1 shared DHSs

Motif	Match	%
TGTGGTTA	RUNX	70
GCAGCTGC	E-box	45
AAGAGGAAGT	ETS	39
ATGAGTCA	AP-1	15

914 CEBPA, Inv(16) and t(8;21) shared DHSs

Motif	Match	%
TGCGGTTT	RUNX	53
AAAAGAGGAAGT	ETS	49
TTGCGTAA	C/EBP	33
TATGAGTCAT	AP-1	28
GATGACATCA	CREB	27

Supplementary Figure 4: AML-specifically active cis-regulatory elements cluster into common and unique chromatin landscapes and correlate with the upregulation of expression of the nearest genes. (a) Venn diagrams depicting the overlaps of subsets of DHSs which are up regulated compared to CD34+ve PBSCs within each of the 7 mutation groups. These groups were generated as the average log2 values for 7 distinct subsets of AMLs that carried the same specific mutations in key regulators. These 7 mutation groups are defined on the basis of average values derived from 3 ITD patients, 6 ITD/NPM1 patients, 2 NPM1 patients, 4 t(8;21) patients, 3 inv(16) patients, 6 RUNX1 patients, and 3 patients with 2 CEBPA mutations. These groups are defined in Table S1 (note colour code). Up-regulated DHSs are defined as being at least 3-fold greater than in PBSCs, and have a DHS signal spanning a 400 bp window of at least 64.

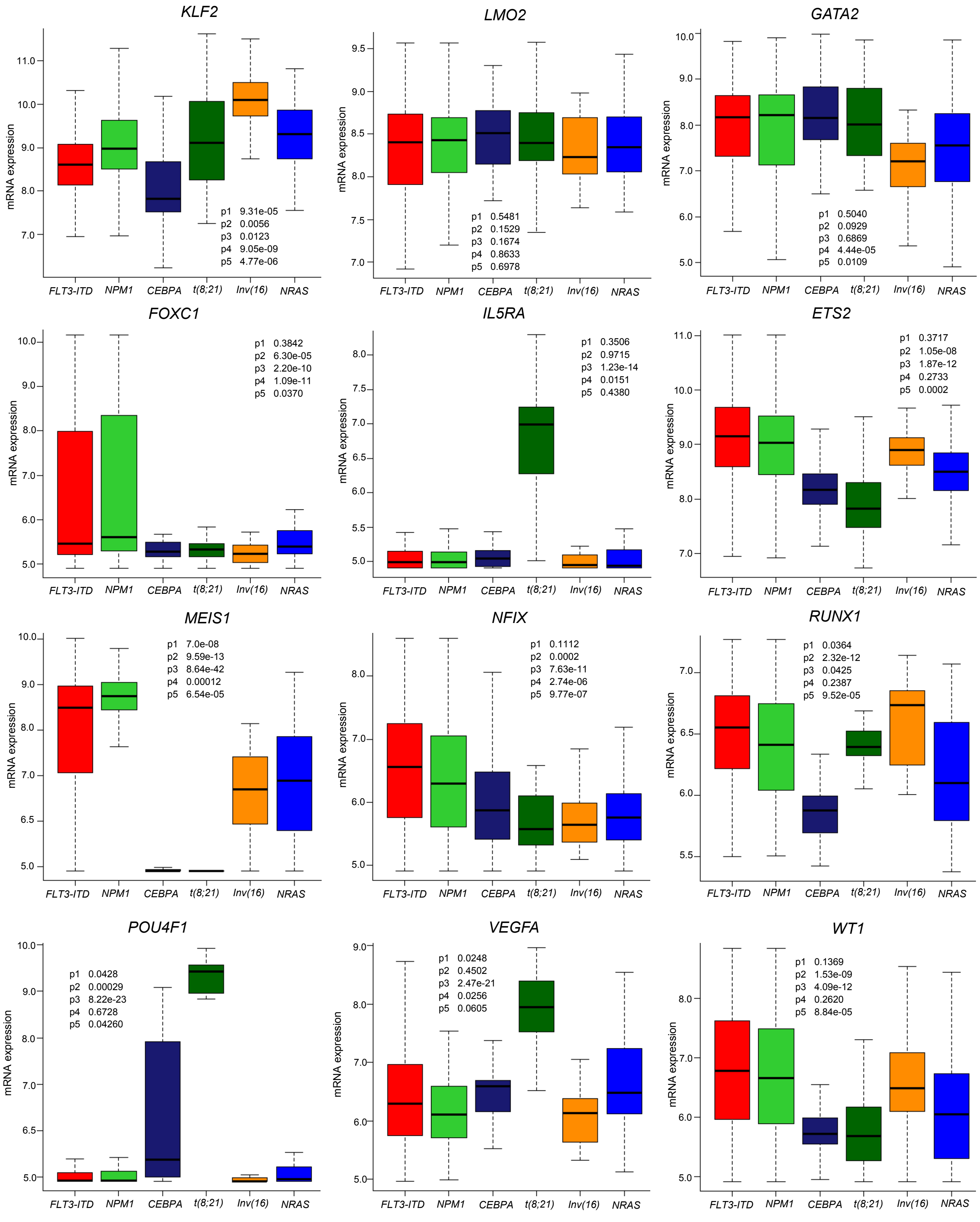
(b) De novo motif search results using Homer for the up regulated DHSs groups and in overlapping deregulated DHSs for ITD and/or NPM1 and for CEBPA and RUNX1 that are shown in (a) the numbers indicate of percentage of each subset that contains the identified motif.

(c) Gene set enrichment analysis for all expressed genes that are annotated to the DHSs identified in each of the AML specific 20 clusters, the enrichment scores (right panel) are aligned against each of the 20 clusters (left panel) that was initially described in **Fig. 2a**. The target genes were tested for enrichment against all AML RNA-seq data; red color indicates that these genes are enriched with up-regulated genes compared to CD34+ PBSC.

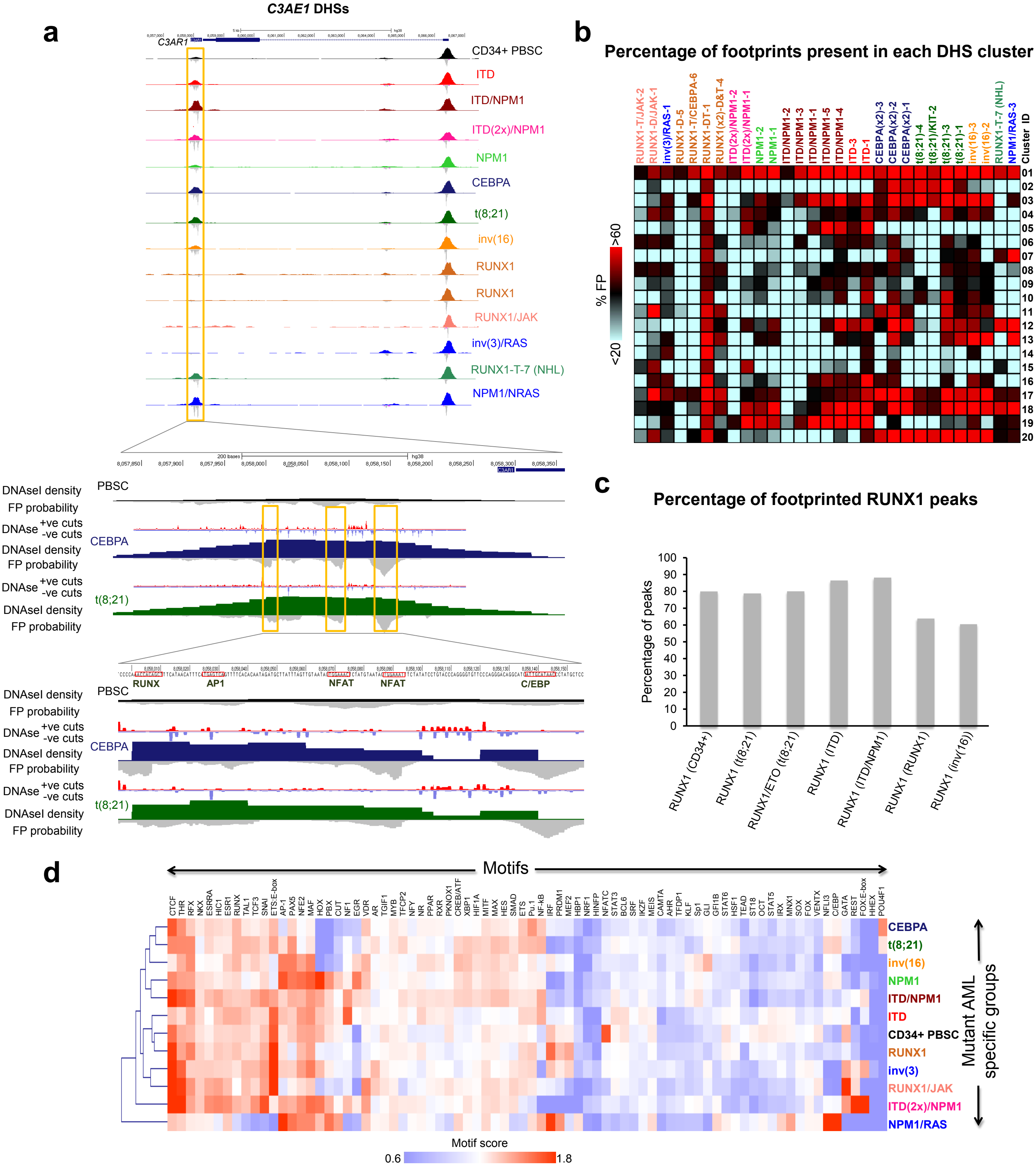
(d) Correlation of gene expression with DNaseI tag count as exemplified for the *POU4F1* gene (see also **Supplementary Fig.1 b,d)*.***

(e) Bar graphs depicting the expression level for some of the targets differentially expressed genes, the FPKM values were plotted on the y-axis for each AML samples used in this study, the color code identified each of the mutation groups.

Validation of gene expression patterns (data from Verhaak et al)



Supplementary Figure 5: Common and group-specific DHS associate with genes belonging to different functional groups. Boxplots validating gene expression patterns for some of the differentially expressed genes using gene expression data from Verhaak et al (2009). P-values highlighting the significance of differences are shown on each panel; the t-test was used to calculate the p-values.

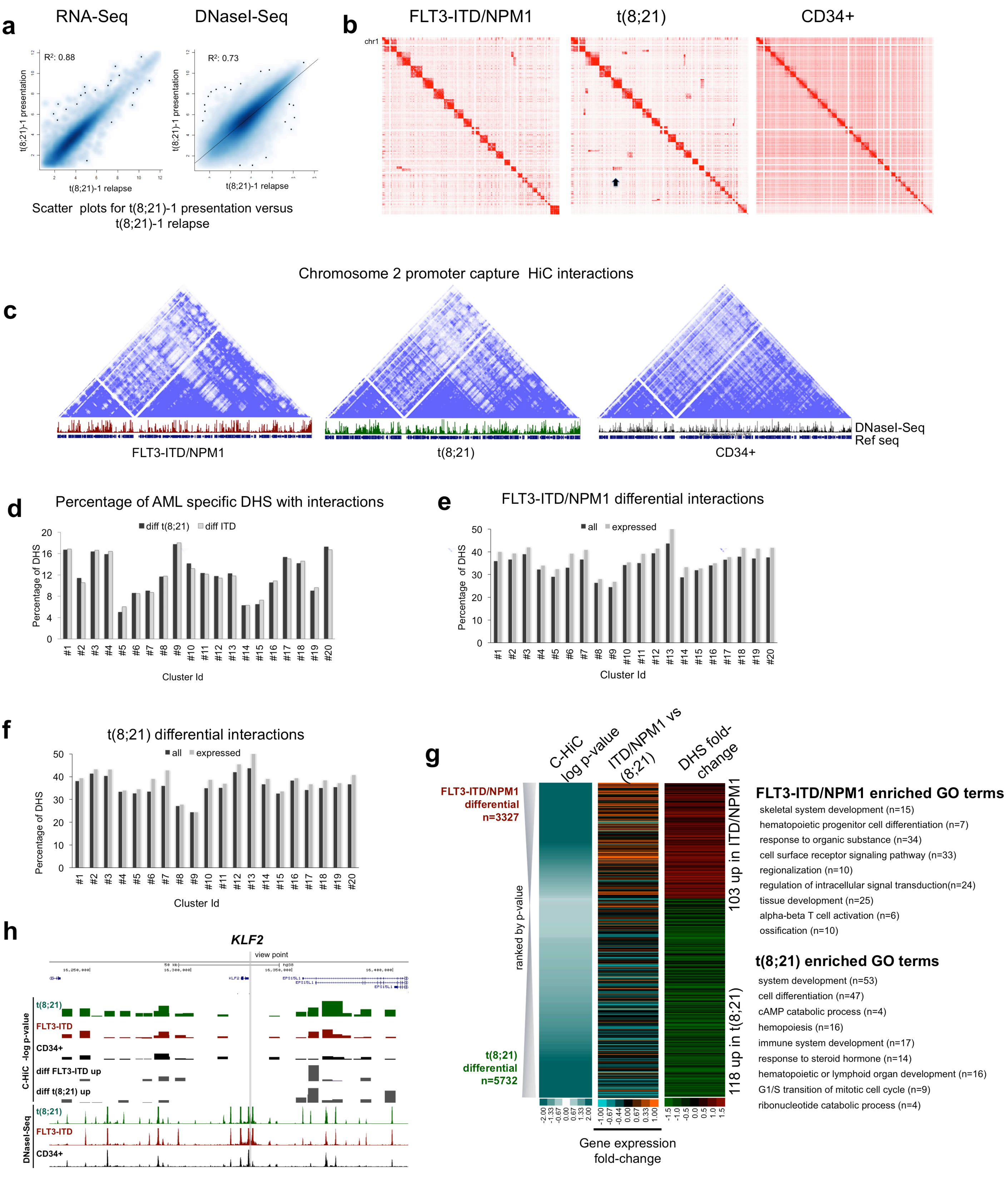


Supplementary Figure 6: AML-specifically active cis-regulatory elements are characterized by specific transcription factor binding patterns. (a) UCSC genome browser screenshot of DNAseI-Seq data aligned with digital footprints at the *C3AE1* locus. The screenshot shows the DHSs for one patient from each group. Footprint probabilities as calculated by Wellington are indicated as grey density below the lines. The bottom indicates the precise location of occupied RUNX, C/EBP and AP-1 footprints.

(b) Percentage of the footprints in the AML specific DHSs for the 20 clusters identified in **Fig. 2a**. The footprints were identified using the Wellington algorithm. We first identified differential footprints for each AML sample compared to the CD34+ PBSCs and then the percentage of these differential footprints in the DHS subsets in the 20 clusters was calculated.

(c) Percentage of footprints with RUNX motifs in the indicated AML-types peaks which are bound by RUNX1 or RUNX1-ETO in ChIP assays from [22,33,66].

(d) Heatmap depicting the degree of motif enrichment after hierarchical clustering of all (not just the specific) motif enrichments for each of the mutation-specific AML groups. Enrichment scores were calculated by the level of motif enrichment in all the footprints of all Hi-read depth samples for each group, as compared to the union of footprints in all experiments.



Supplementary Figure 7: Capture HiC shows differences in cis-regulatory interactions between different types of AML and normal cells. (a) Smooth scatter plots show the correlation between t(8;21)-1 presentation and t(8;21)-1 relapse AML DNAseI-Seq data. The coefficients of determination (R-squared) highlighting the significance of correlation are shown on each panel.

(b) Heatmaps show the raw overall inter and intra interactions of the promoter capture HiC data for all chromosomes for FLT3-ITD (ITD/NPM1-2, left), t(8;21) (middle) and CD34+ (right) across all chromosomes.

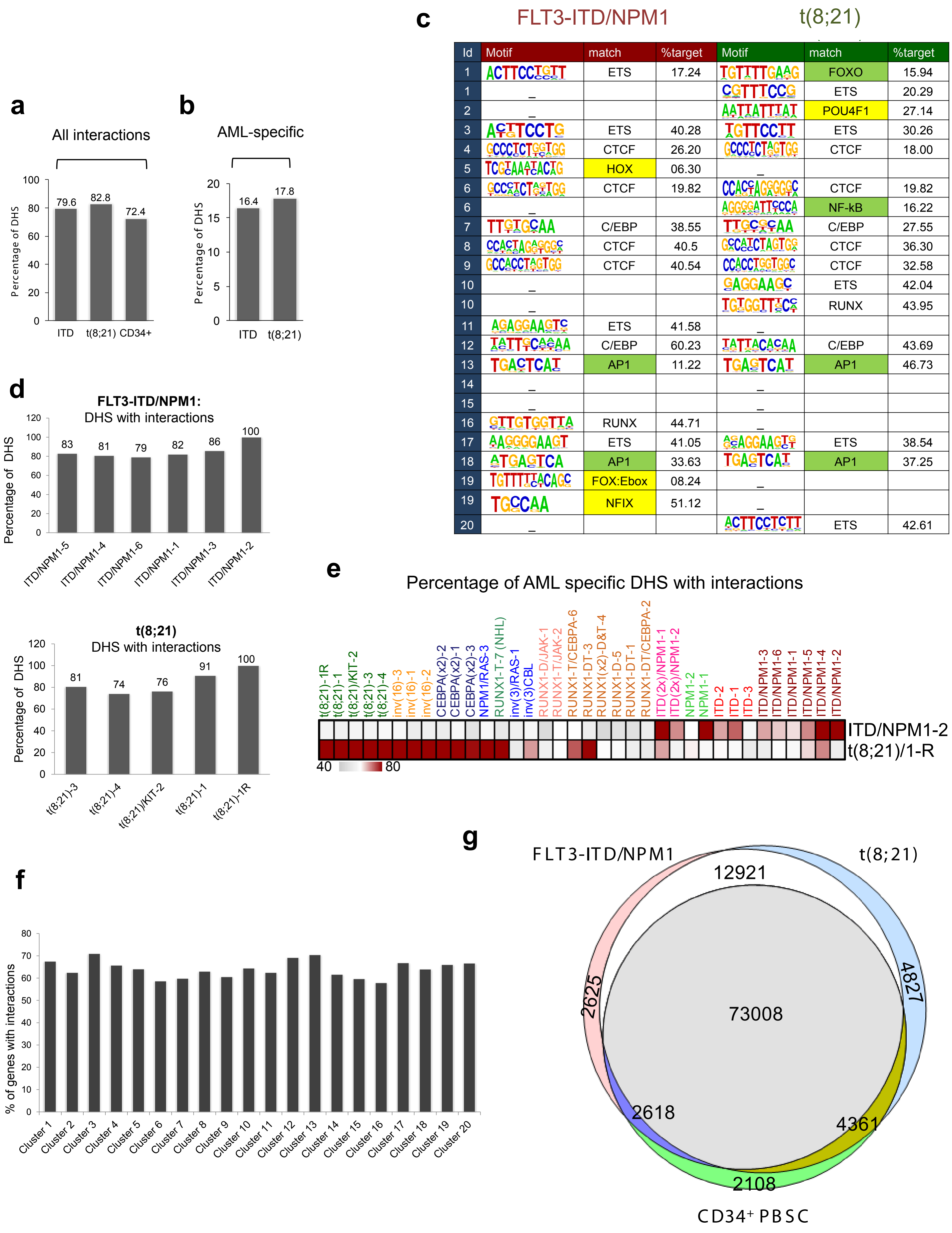
(c) Heatmaps showing the raw interactions of the promoter capture HiC data using purified patient blasts on chromosome 2 for the FLT3-ITD (FLT3-ITD/NPM1 patient) (left), t(8;21) (middle) and CD34+ (right), a UCSC tracks is shown below each heatmap.

(d) Bar figure showing the percentage of DHSs within each of the 20 clusters identified in **Fig. 2a** that have differential interactions compared to CD34+ cells.

(e, f) percentage of DHSs within each of the 20 clusters interacting with the nearest gene within differential interactions for all genes expressed genes as identified by the RNA-Seq data. e: FLT3-ITD and f: t(8;21).

(g) Heatmap of differential interactions ranked by the strength of interaction (-log p-value) from highly significant to less significant for the FLT3-ITD and from less significant to more significant for the t(8;21) (outer left panel). Plotted along-side is the gene expression fold-difference for the FLT3-ITD compared to the t(8;21) (middle panel) and the DHS fold difference FLT3-ITD versus the t(8;21) (right panel). Outmost right panel top: The top enriched GO terms for up regulated genes associated with differentially interacting DHSs in the FLT3-ITD compared the t(8;21). Bottom: the top enriched GO terms for the up regulated genes in the t(8;21) compared to the FLT3-ITD.

(h) UCSC genome browser showing a screenshot of *KLF2*. The top two tracks display the log p-value of the capture HiC interaction for *KLF2* promoter as viewpoint, the following two tracks display log p-value of the differential interaction of the t(8;21) and the FLT3-ITD compared to the CD34+. Shown are also the DNaseI-Seq and RNA-Seq data of t(8;21), FLT3-ITD and CD34+ PBSC.



Id	Motif	match	%target	Motif	match	%target
1	ACTTCCTGTT	ETS	17.24	TGTTTTGAAG	FOXO	15.94
1	_			CGTTTCCG	ETS	20.29
2	_			AATTATTTAT	POU4F1	27.14
3	ACTTCCTG	ETS	40.28	TGTTCCTT	ETS	30.26
4	GCCCTCTGGTGG	CTCF	26.20	GCCCTCTAGTGG	CTCF	18.00
5	TCGTAAATACTG	HOX	06.30	_		
6	GCCCACTAGTGG	CTCF	19.82	CCACCAGGGGC	CTCF	19.82
6	_			AGGGGATTCCCA	NF-kB	16.22
7	TTGTGCAA	C/EBP	38.55	TTGCGCAA	C/EBP	27.55
8	CCACTAGAGGGC	CTCF	40.5	GCCATCTAGTGG	CTCF	36.30
9	GCCACCTAGTGG	CTCF	40.54	CCACCTGGTGGC	CTCF	32.58
10	_			GAGGAAGC	ETS	42.04
10	_			TGTGGTTTCC	RUNX	43.95
11	AGAGGAAGTC	ETS	41.58	_		
12	TATTGCACAA	C/EBP	60.23	TATTACACAA	C/EBP	43.69
13	TGACTCAT	AP1	11.22	TGAGTCAT	AP1	46.73
14	_			_		
15	_			_		
16	GTTGTGGTTA	RUNX	44.71	_		
17	AAGGGGAAGT	ETS	41.05	AGAGGAAGTG	ETS	38.54
18	ATGAGTCA	AP1	33.63	TGAGTCAT	AP1	37.25
19	TGTTTTTACAGC	FOX:Ebox	08.24	_		
19	TGCCAA	NFIX	51.12	_		
20	_			ACTTCCTCTT	ETS	42.61

Supplemental Figure 8: Interactions are representative for their patient groups and the majority of interactions are shared. (a) Bar diagram showing the percentage of DHSs involved in significant interactions. (I) Bar diagram showing the percentage of DHSs involved in significant differential interactions compared to CD34+ cells. (b) Bar diagram showing the percentage of DHSs involved in significant differential interactions for DHSs unique to FLT3-ITD or t(8;21) DHSs compared to CD34+ cells, with DHS common to FLT3-ITD and t(8;21) being excluded.

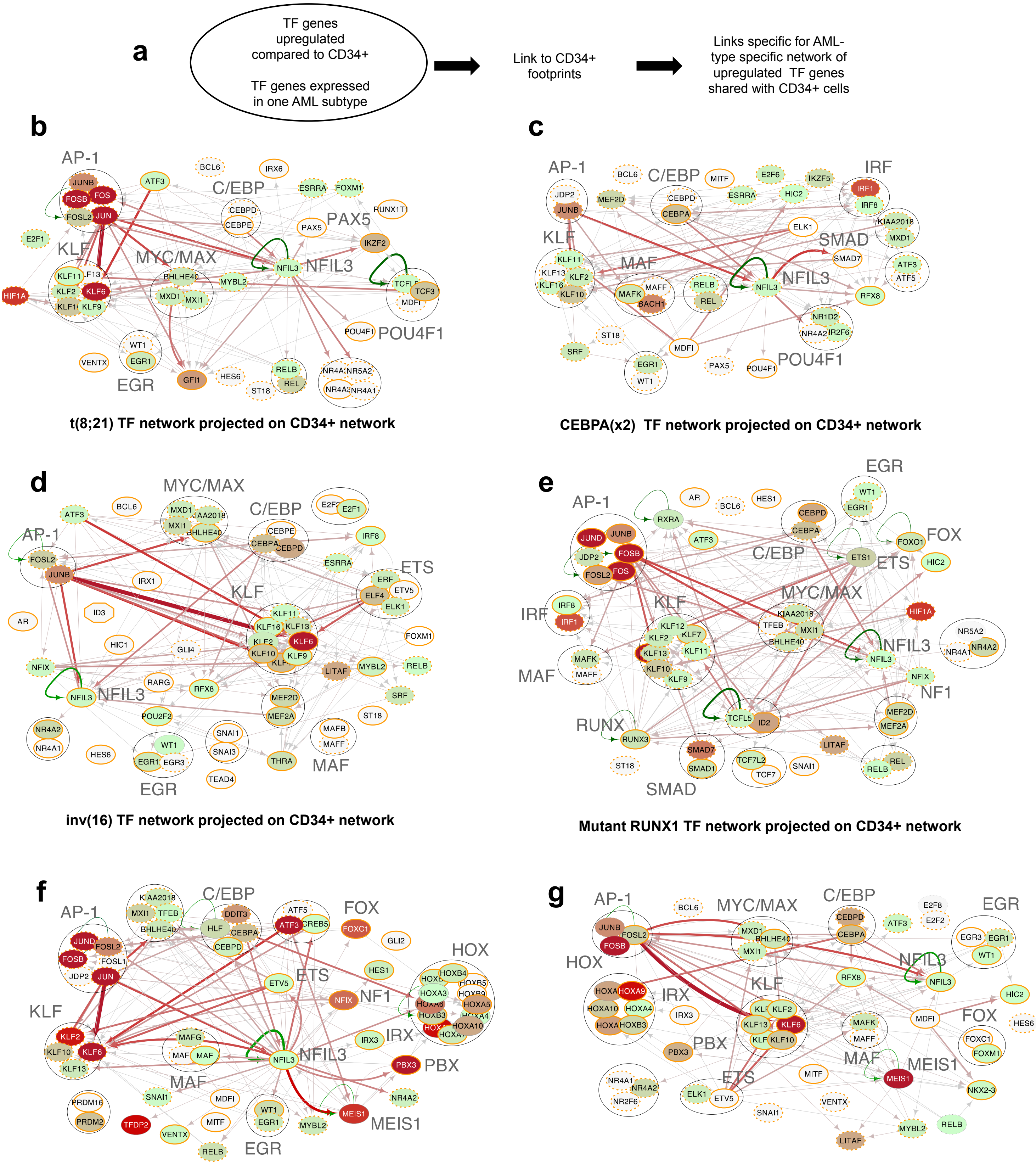
(c) Enriched footprinted motifs in DHS associated each of the 20 clusters involved in differential interactions for the two patients. Motifs for transcription factors normally not expressed in myeloid cells are highlighted in yellow, motifs for inducible factors are marked in green.

(d) Percentage of all DHSs with interactions present in each dataset of each individual patients,

(e) Heatmap highlighting the percentage of AML-type specific DHS with interactions found in the different patient groups, indicating that the patient chosen for the Chi-C experiment are representative for each patient group.

(f) Percentage of up-regulated gens associated with DHS clusters that have significant interactions in any of the three Chi-C experiments.

(h) Overlap of all DHSs underlying interactions in all three samples as indicated demonstrating that the majority of interactions are the same in all three samples.



t(8;21) TF network projected on CD34+ network

CEBPA(x2) TF network projected on CD34+ network

inv(16) TF network projected on CD34+ network

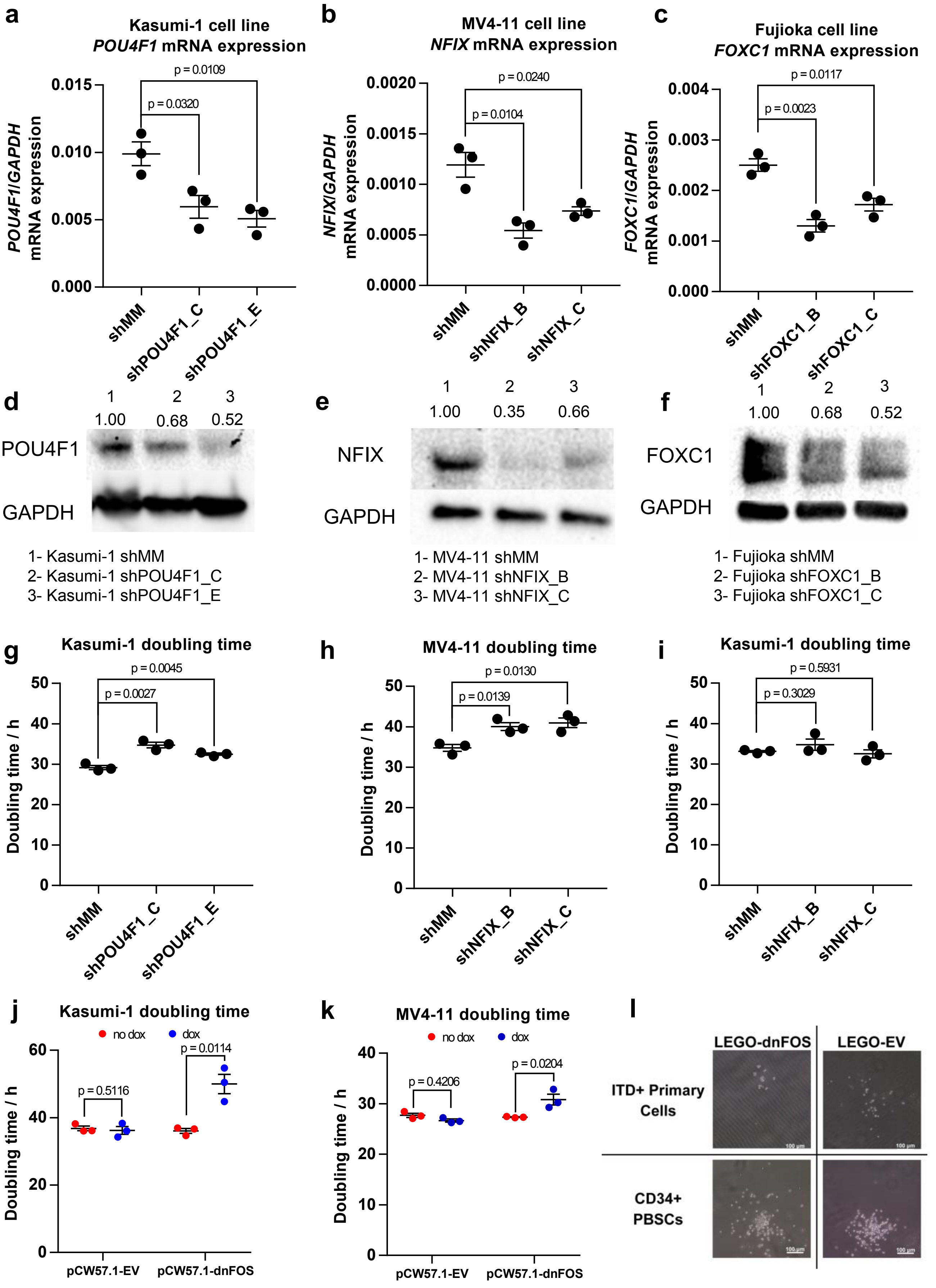
Mutant RUNX1 TF network projected on CD34+ network

FLT3-ITD/NPM1 TF network projected on CD34+ network

NPM1 TF network projected on CD34+ network

Supplementary Figure 9: Identification of transcription factor network components driving the expression of TF genes in each AML subtype which are shared with CD34+ cells. Here we projected the links from the indicated AML subtypes onto the CD34+ footprints.

(a) Analysis strategy. (b) Shared t(8;21) TF network, (c) Shared CEBPA(x2) TF, (d) Shared Inv(16) TF network, (e) Shared Mutant RUNX1 TF network, (f) Shared FLT3-ITD/NPM1 TF network, (g) Shared NPM1- TF network. Factor families binding to the same motif as shown in Table S2 form a node contained within a circle. Arrows going outwards from the entire node highlight footprinted motifs in individual genes generated by any member of this factor family whereby the footprint was annotated to the gene using the Chi-C data where possible, otherwise to the nearest gene. The expression level (FKPM) for the individual genes is depicted in white (low)/red (high) colour. An orange smooth ring around the circle indicates that this gene is specifically up-regulated in this type of AML compared to CD34+ PBSCs and/or other AML types, a dotted circle indicates a gene that is up-regulated as compared to CD34+ cells. Genes with no outgoing arrows due to a lack of know binding motifs are highlighted by an octagon shape. For a detailed guide to node and edge attributes: See legend of **Fig 6**.



Supplementary Figure 10: AML type-specifically expressed transcription factors are required for leukemic growth. (a, b, c) Dot plots showing *POU4F1* (a), *FOXC1* (b) and *NFIX*

(c) mRNA expression after transduction with the indicated shRNA and control lentiviruses in Kasumi-1, MV4-11 and Fujioka cell lines, respectively. Note that Fujioka cells express high levels of *FOXC1* and were only used to test the functionality of our lentiviral construct. *FOXC1* is not highly expressed in MV4-11 cells.

(d-f) Western Blots showing the efficiencies of shRNA knock-down for FOXC1 (d), NFIX (e) and POU4F1 (f), images are representative of three independent experiments.

(g - i) Dot plots showing doubling time of t(8;21) Kasumi-1 cells after transduction with sh*POU4F1* (g)*,* MV4-11 cells after transduction with *shNFIX* (h) and of Kasumi-1 cells after transduction with *shNFIX* (i).

(j, k): doubling times of Kasumi-1 (j) and MV4-11 cells (k) expressing a DOX inducible version of a dominant negative FOS peptide (dnFOS) (k,m) as well as an empty vector control (l,o).

All experiments were performed in triplicate as detailed in Online Methods. In all dot plots n=3 independent biological replicates with p values calculated using a two-tailed Student's t-test. Error bars show standard error of the mean. (l) Pictures of representative colonies derived from FLT3-ITD patient cells and CD34+ PBSCs transduced with the indicated lentiviral vectors.

2. SUPPLEMENTARY NOTES AND DISCUSSION

The complexity of our data allows extensive integration with published data-sets, and allows us to address specific questions of specialist interest which cannot be extensively discussed in the main paper. Here we show examples of such analyses.

Determining the stage of the differentiation block in different AML subtypes

We used our DHS data to examine whether the mutation class and its associated DHS pattern correlated with a block at a specific stage of the differentiation. Here we used published ATAC-Seq data[67] describing the open chromatin landscape of normal stem and progenitor cells **(Fig 2a)**. Our DNaseI-Seq data correlated well with these data (**Fig S3a** and **S3**b), whereby CD34+ PBSC sequences clustered with hematopoietic stem cells (HSCs) and early progenitors but not monocytic cells. When compared to the various types of progenitor cells, t(8;21), inv(16), CEBPA(x2) and NPM1-mutated AML displayed distal element patterns most similar to those of normal GMPs, with some differentiation into monocytes (**Fig SN1b**). In contrast, RUNX1 and FLT3-ITD/NPM1 mutated AML displayed a spread of lineage-specific patterns with little or no monocytic differentiation (**Fig SN1b**). Gene-set enrichment analysis comparing the gene expression patterns of AML cells with the various progenitor stages confirmed that the mutation-group specific cistrome was mirrored by the gene expression pattern (**Fig SN1c**). However, although AML subtypes showed some characteristics of normal progenitor cells, they still clustered away from normal cells (**Suppementary Fig 3b**). These data indicate how AML cells reprogram their chromatin and that cell differentiation goes “sideways”, meaning that cells adopt a separate identity compared to all stages of normal myeloid cells.

Figure SN1

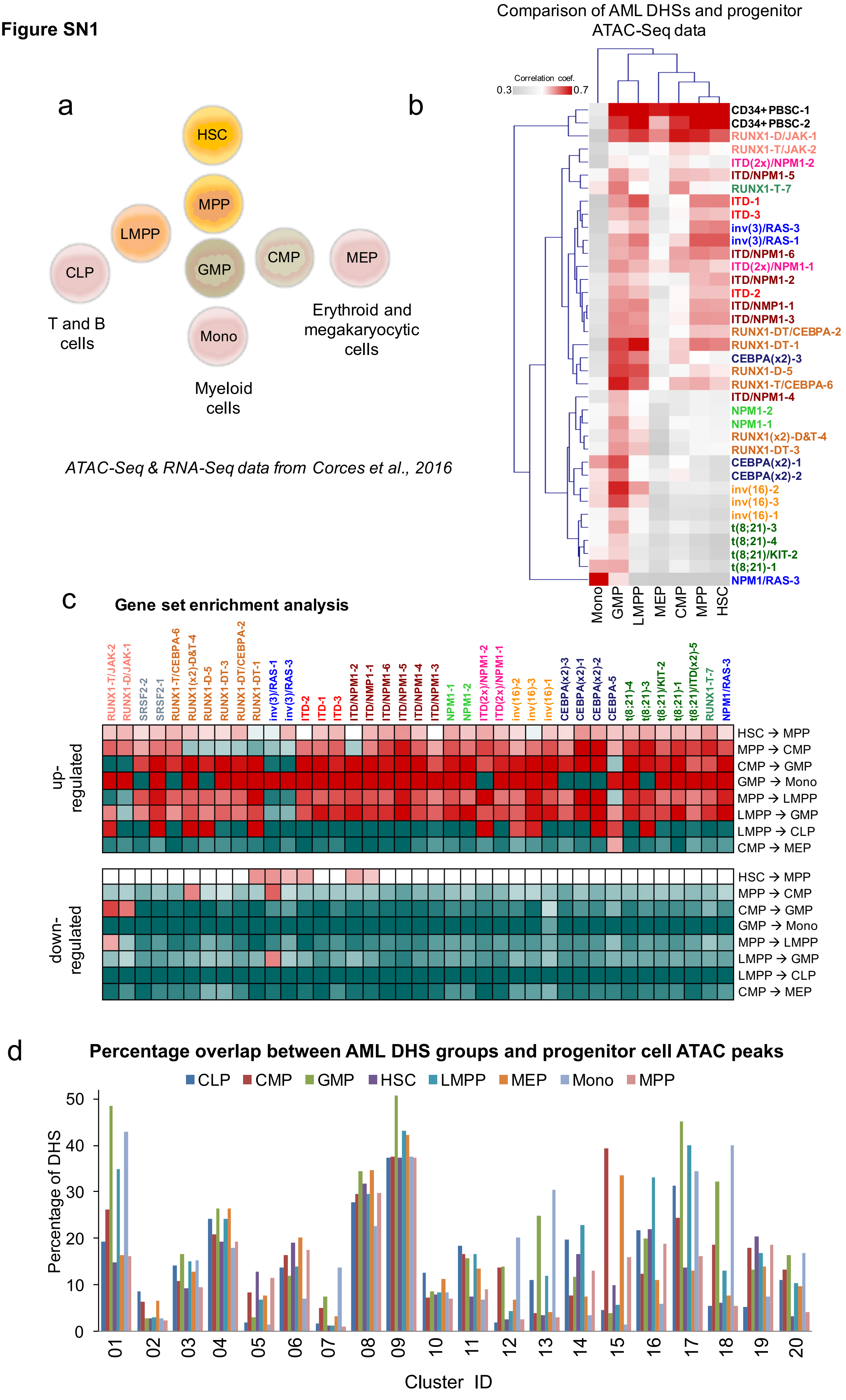

a
HSC
MPP
LMPP
CLP
GMP
CMP
MEP
Mono
T and B cells
Myeloid cells
Erythroid and megakaryocytic cells
ATAC-Seq & RNA-Seq data from Corces et al., 2016
b
Comparison of AML DHSs and progenitor ATAC-Seq data
Correlation coef.
0.3
0.7
CD34+PBSC-1
CD34+PBSC-2
RUNX1-D/JAK-1
RUNX1-T/JAK-2
ITD(2x)/NPM1-2
ITD/NPM1-5
RUNX1-T-7
ITD-1
ITD-3
inv(3)/RAS-3
inv(3)/RAS-1
ITD/NPM1-6
ITD(2x)/NPM1-1
ITD/NPM1-2
ITD-2
ITD/NMP1-1
ITD/NPM1-3
RUNX1-DT/CEBPA-2
RUNX1-DT-1
CEBPA(x2)-3
RUNX1-D-5
RUNX1-T/CEBPA-6
ITD/NPM1-4
NPM1-2
NPM1-1
RUNX1(x2)-D&T-4
RUNX1-DT-3
CEBPA(x2)-1
CEBPA(x2)-2
inv(16)-2
inv(16)-3
inv(16)-1
t(8;21)-3
t(8;21)-4
t(8;21)/KIT-2
t(8;21)-1
NPM1/RAS-3
Mono
GMP
LMPP
MEP
CMP
MPP
HSC
c
Gene set enrichment analysis
RUNX1-T/JAK-2
RUNX1-D/JAK-1
SRSF2-2
SRSF2-1
RUNX1-T/CEBPA-6
RUNX1(x2)-D&T-4
RUNX1-D-5
RUNX1-DT-3
RUNX1-DT/CEBPA-2
RUNX1-DT-1
inv(3)/RAS-1
inv(3)/RAS-3
ITD-2
ITD-1
ITD-3
ITD/NPM1-2
ITD/NMP1-1
ITD/NPM1-6
ITD/NPM1-5
ITD/NPM1-4
ITD/NPM1-3
NPM1-1
NPM1-2
ITD(2x)/NPM1-2
ITD(2x)/NPM1-1
inv(16)-2
inv(16)-3
inv(16)-1
CEBPA(x2)-3
CEBPA(x2)-1
CEBPA(x2)-2
CEBPA-5
t(8;21)-4
t(8;21)-3
t(8;21)/KIT-2
t(8;21)-1
t(8;21)/ITD(x2)-5
RUNX1-T-7
NPM1/RAS-3
up-regulated
HSC → MPP
MPP → CMP
CMP → GMP
GMP → Mono
MPP → LMPP
LMPP → GMP
LMPP → CLP
CMP → MEP
down-regulated
HSC → MPP
MPP → CMP
CMP → GMP
GMP → Mono
MPP → LMPP
LMPP → GMP
LMPP → CLP
CMP → MEP
d
Percentage overlap between AML DHS groups and progenitor cell ATAC peaks
CLP
CMP
GMP
HSC
LMPP
MEP
Mono
MPP
Percentage of DHS
0
10
20
30
40
50
01
02
03
04
05
06
07
08
09
10
11
12
13
14
15
16
17
18
19
20
Cluster ID

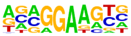
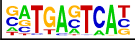
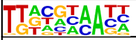
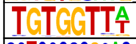
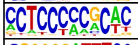
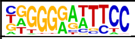
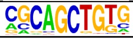
Figure SN1: Different types of AML are blocked at different stages of differentiation and are regulated by different transcriptional network. (a) Hematopoietic hierarchy; shown are some of the precursor stages from which ATAC-seq and RNA-seq data were generated in Corces et al., 2016: Hematopoietic stem cells (HSC), common myeloid progenitors (CMP), common lymphoid progenitors (CLP), Megakaryocyte Erythrocyte Precursors (MEP) and Granulocyte Macrophage Precursors (GMP). (b) Clustering of the correlation of percentage of peak overlap between DNaseI-Seq and ATAC-seq data by first generating a matrix with all overlap percentages between all DHS peaks, and ATAC-seq peaks and then hierarchically clustering. (c) Gene set enrichment analysis for the differentially expressed genes that are at least 2-fold different compared to the normal CD34+ PBSCs. Up and down regulated gene expression patterns were tested for their similarity to specific pairs of progenitor RNA-seq data from Corces et al. 2016, representing different steps of differentiation. Up-regulated genes are shown in top panel and the bottom panel shows the down-regulated genes. (d) The percentage of DHS peaks that overlap with ATAC-Seq data from different progenitor types, DHS clusters from **Fig. 2a** was overlapped with each of the progenitor ATAC peaks; these include CLP, CMP, GMP, MPP, LMPP, MEP and Monocyte populations.

Comprehensive motif analysis of occupied sequences within DHS in AML patients

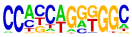
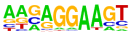
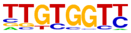
The full complement of all DHSs in all cell types occupies a much larger sequence space within the genome than the DHSs present in any one cell type. To define the full complement of ~128,000 distal DHSs present in either PBSCs or the AML samples included in this study we created a merged data set. To complement our digital footprinting analyses of these DHSs we identified enriched DNA-binding motifs in each of the 20 clusters in **Fig. 2a** using HOMER[68] (**Figure SN1a**). The vast majority of all DHS clusters show an enrichment of occupied RUNX1 and ETS motifs which forms the backbone of each AML TF network. The same was true for AP-1 motifs which were enriched in most AML-specific clusters, consistent with the presence of signalling mutations in most samples. In general, this analysis confirmed the results of our footprinting analysis. We show that (i) the enrichment of a POU4F1 motif in DHSs in cluster 2 linked to t(8;21) and CEBPA double mutant AML, (ii) the E-box in cluster 17 is shared by RUNX1, CEBPA, t(8;21) and inv(16) AML, and (iii) the HOX and Nuclear Factor I (NFI) motif signatures in the FLT3-ITD/NPM1–specific Cluster 19.

The mutation-specific DHS subgroups defined in **Supplementary Fig 4** were also distinct from a subset of 5460 DHSs which were up-regulated in GMPs compared to PBSCs (**Fig. SN1b**). Just 146 of the 942 DHSs shared by the ITD and NPM1 subgroups, and 71 of the 228 DHSs shared by the RUNX1, CEBPA, t(8;21) and inv(16) subgroups were up-regulated in GMPs, the population which was otherwise the most similar to AML blasts. These AML-specific subsets of DHSs maintained a common AP-1/ETS signature, while the t(8;21), inv(16), RUNX1 group maintained an E-box signature. When analysed as scatter blots (**Fig. SN1b**), the 942 ITD/NPM1 group-specific DHSs show strong upregulation in the ITD/NPM1 AML cells, but there is much less variation between these DHSs in PBSCs compared to ATAC peaks at these regions in GMPs.

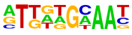
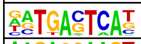
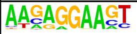
Cluster 01 (common)

Motif	Match	P-value	%
	ETS	1e-161	15.06
	AP-1	1e-146	06.61
	C/EBP	1e-124	11.92
	RUNX	1e-104	13.40
	EGR	1e-36	02.59
	NF-kB	1e-30	02.19
	E-box	1e-24	02.72

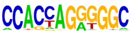
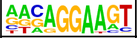
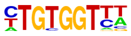
Cluster 04

Motif	Match	p-value	%
	CTCF	1e-469	08.13
	ETS	1e-184	06.39
	RUNX	1e-110	06.37

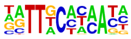
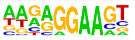
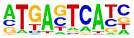
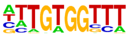
Cluster 07

Motif	Match	p-value	%
	C/EBP	1e-675	22.29
	AP-1	1e-280	08.46
	ETS	1e-136	10.45

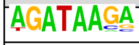
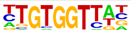
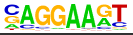
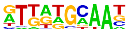
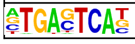
Cluster 09

Motif	Match	P-value	%
	CTCF	1e-755	10.98
	ETS	1e-115	3.96
	RUNX	1e-85	04.21

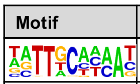
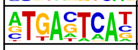
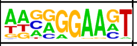
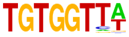
Cluster 12

Motif	Match	p-value	%
	C/EBP	1e-691	23.09
	ETS	1e-169	12.61
	AP-1	1e-169	05.14
	RUNX	1e-58	05.48

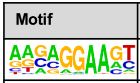
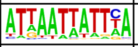
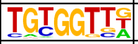
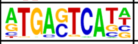
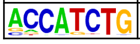
Cluster 15

Motif	Match	p-value	%
	GATA	1e-153	09.55
	RUNX	1e-101	06.64
	ETS	1e-99	09.60
	C/EBP	1e-78	05.24
	AP-1	1e-78	03.16

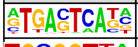
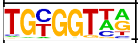
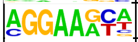
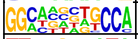
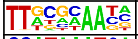
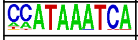
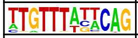
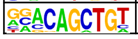
Cluster 18

Motif	Match	P-value	%
	C/EBP	1e-400	17.16
	AP-1	1e-363	10.10
	ETS	1e-121	09.32
	RUNX	1e-74	7.54

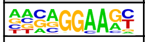
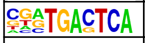
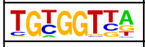
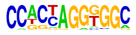
Cluster 02 (CEBPA and t(8;21)

Motif	Match	p-value	%
	ETS	1e-185	15.30
	**POU4F1**	1e-158	03.51
	RUNX	1e-50	07.20
	AP-1	1e-38	02.06
	E-box	1e-31	06.07

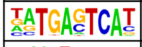
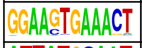
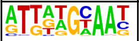
Cluster 05

Motif	Match	p-value	%
	AP-1	1e-248	05.00
	RUNX	1e-157	06.92
	ETS	1e-128	08.79
	NF1	1e-120	03.64
	C/EBP	1e-117	04.32
	HOX	1e-98	01.55
	FOX:Ebox	1e-73	00.52
	E-box	1e-38	02.89

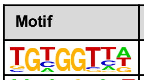
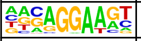
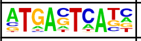
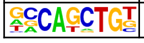
Cluster 10

Motif	Match	P-value	%
	ETS	1e-184	09.77
	AP-1	1e-139	02.71
	RUNX	1e-89	04.55
	CTCF	1e-88	01.90

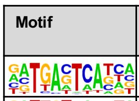
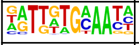
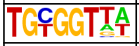
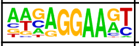
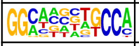
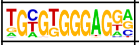
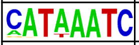
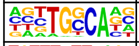
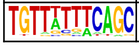
Cluster 13

Motif	Match	p-value	%
	AP-1	1e-452	12.44
	IRF	1e-212	05.43
	C/EBP	1e-212	06.96

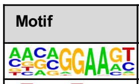
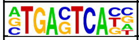
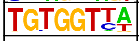
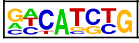
Cluster 16

Motif	Match	P-value	%
	RUNX	1e-190	12.61
	ETS	1e-186	15.92
	AP-1	1e-68	03.02
	E-box	1e-35	05.06

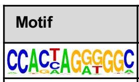
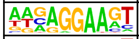
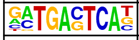
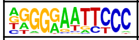
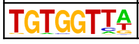
Cluster 19 FLT3-ITD and NPM1

Motif	Match	P-value	%
	AP-1	1e-208	8.23
	C/EBP	1e-135	6.41
	RUNX	1e-130	8.62
	ETS	1e-109	8.43
	NF1	1e-72	5.81
	EGR	1e-68	1.90
	HOX	1e-35	1.96
	NF1-half	1e-29	4.81
	FOX::Ebox	1e-26	0.65

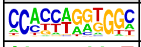
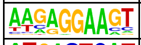
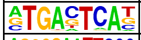
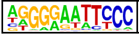
Cluster 03

Motif	Match	p-value	%
	ETS	1e-120	17.68
	AP-1	1e-90	04.34
	RUNX	1e-80	08.18
	E-box	1e-36	08.88

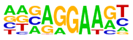
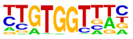
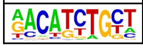
Cluster 06

Motif	Match	p-value	%
	CTCF	1e-726	11.38
	ETS	1e-157	07.65
	AP-1	1e-86	03.11
	NF-kB	1e-67	01.53
	RUNX	1e-34	03.64

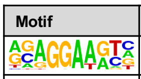
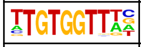
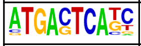
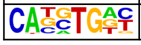
Cluster 08

Motif	Match	p-value	%
	CTCF	1e-694	14.38
	ETS	1e-164	07.09
	AP-1	1e-51	02.63
	NF-kB	1e-40	01.66

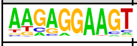
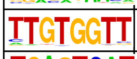
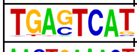
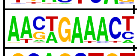
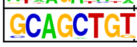
Cluster 11

Motif	Match	p-value	%
	ETS	1e-303	15.00
	RUNX	1e-106	08.19
	E-box	1e-40	07.38

Cluster 14

Motif	Match	p-value	%
	ETS	1e-223	13.29
	RUNX	1e-118	06.07
	AP-1	1e-75	03.08
	E-box	1e-45	11.37

Cluster 17 (CBF; CEBPA, RUNX1)

Motif	Match	P-value	%
	ETS	1e-314	15.67
	RUNX	1e-84	12.25
	AP-1	1e-31	3.05
	IRF	1e-28	1.61
	Ebox	1e-25	6.35

Cluster 20 (CBF; CEBPA)

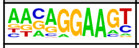
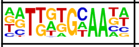
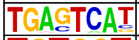
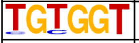
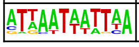
Motif	Match	P-value	%
	ETS	1e-275	15.34
	C/EBP	1e-92	8.12
	AP-1	1e-65	5.12
	RUNX	1e-60	10.2
	POU4F1	1e-22	0.49

Figure SN2a: Comprehensive motif analysis of footprinted sequences within DHS clusters 1-20 (related to Figure 3A). The major differentially enriched motifs are depicted in red.

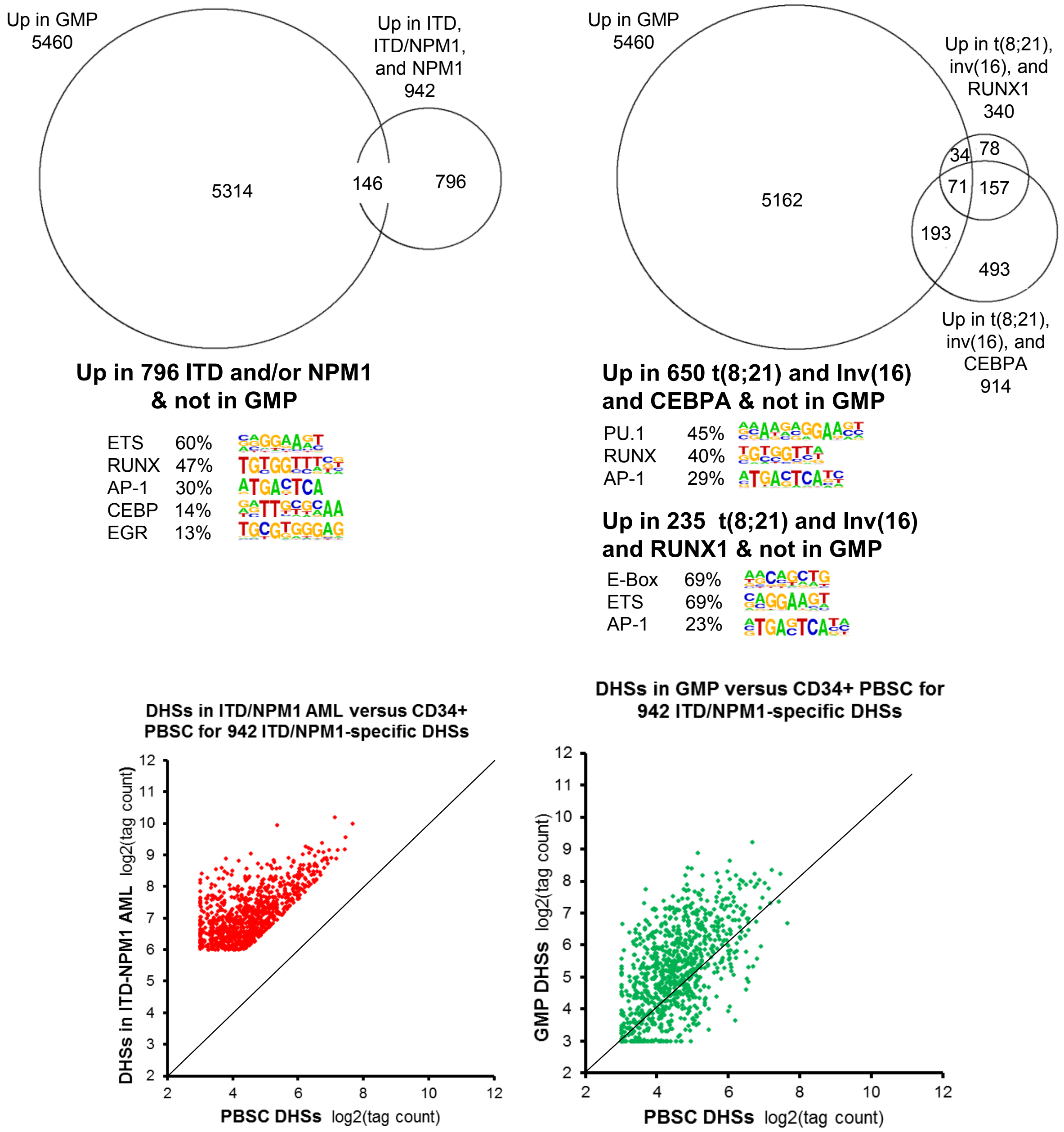
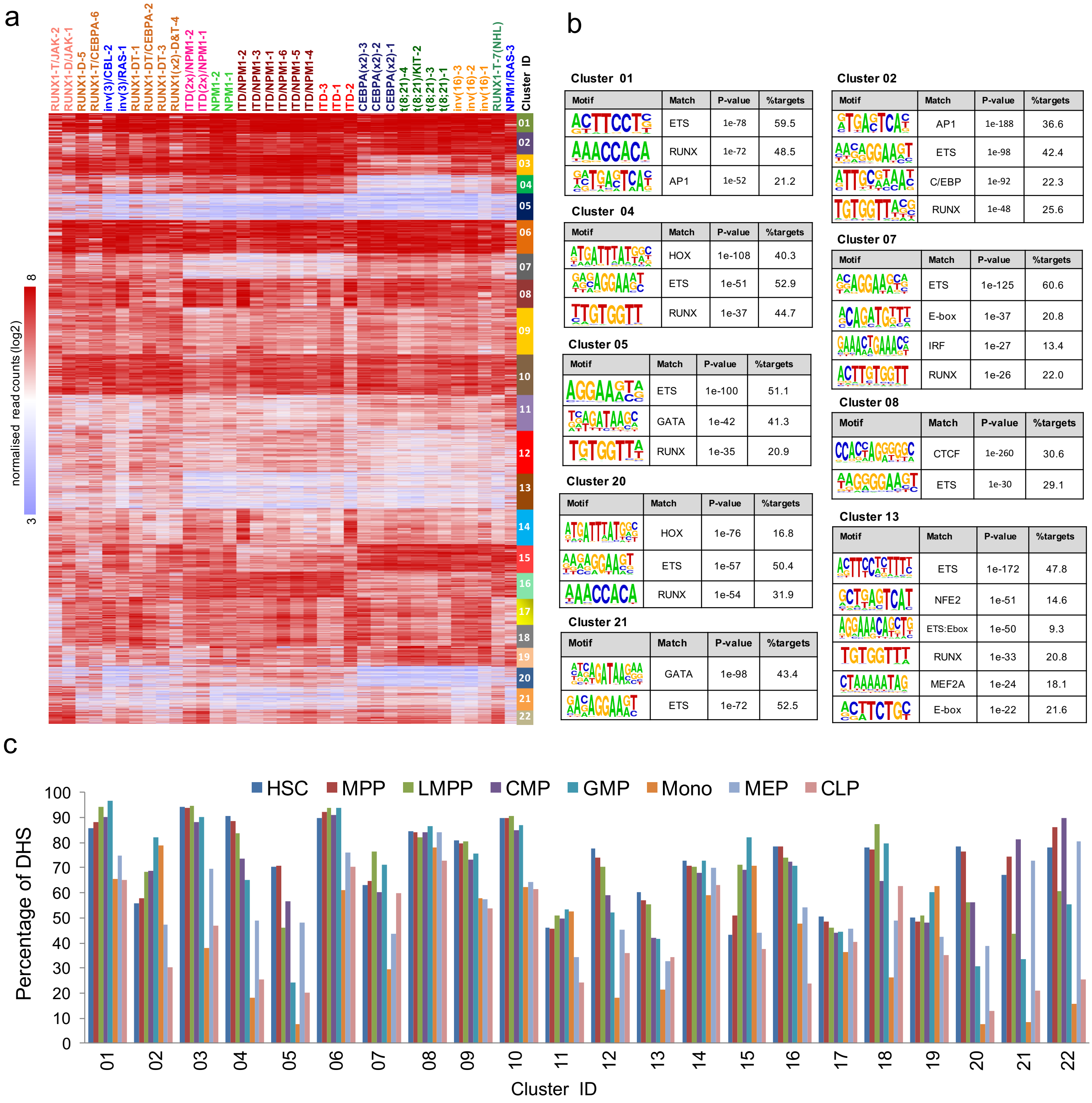


Figure SN2b: Top: Venn diagrams depicting the overlaps of subsets of DHSs which are upregulated in specific AML mutation classes, and DHSs which are upregulated in GMPs, relative to PBSCs. This analysis evaluated 942 DHSs which are upregulated in FLT-ITD, FLT3-ITD/NPM1 and NPM1 sub-types of AML relative to CD34+ PBSCs, and found that just 15% of these DHSs were also upregulated in GMPs relative to PBSCs (146 DHSs). We also evaluated 914 DHSs upregulated in the t(8;21), inv(16) and CEBPA subtypes of AML and another overlapping group of

340 DHSs upregulated in the t(8;21), inv(16) and RUNX1 subtypes of AML. The majority of these DHSs were also AML-specific as they were not upregulated in GMPs. Furthermore, the above two major groupings of upregulated DHSs were also unrelated to each other as just 179 DHSs were shared between the above groups of 942 and 914 DHSs that define these groupings. The main features distinguishing the major groupings were the enrichment for EGR and C/EBP motifs in the ITD/NPM1 grouping, and E-box motifs in the t(8;21/inv(16)/RUNX1 grouping. **Bottom:** Scatter plots showing DHS peak sizes for the 942 ITD/NPM1-specific DHSs in ITD/NPM1 AML cells relative to PBSCs, and ATAC peaks for the same regions in GMPs relative to DHSs in PBSCs. This analysis also confirms that it is the FLT3-ITD and not the DNMT3A mutation that underlies differences to CD34+ PBSCs.

Our detailed analysis of AML-type-specific DHSs also allowed us to identify distal elements that were active in normal CD34$^+$ cells but which were specifically lost in in AML cells. When investigated in parallel with the analyses of the progenitor-specific DHS profiles, and the block in differentiation seen in the AML cells, these analyses can shed additional light on the precise nature of the defect seen in the AML cells. Similar to the analysis of up-regulated DHSs shown in **Fig. 3a**, **Fig. SN2a** shows a supervised clustering analysis of down-regulated DHSs together with a search for enriched motifs in such sites (**Fig. SN2b**) as in **Fig SN1**, with the DHSs present in CD34$^+$ cells but not in AML cells highlighted in yellow. The motif analyses of the cluster-specific groups (Fig. SN2B) revealed a loss of HOX motifs in clusters 4 and 20 which are specifically down-regulated in the closely related t(8;21), Inv(16) and CEBPA groups of AML samples, confirming that HOX factors play no role in programming these sub-types of AML-type. It also shows that GATA-motifs were enriched in cluster 5 spanning all of the mutation-specific subgroups of DHSs that are lost in AML cells compared to PBSCs, thereby explaining why no GATA motifs were enriched in the AML-type specific DHSs (**Fig.4**; **Supplementary Fig 4b**). In combination, these results indicate that such cells are past the stage where they are dependent on GATA2 and are unrelated to the GATA1 dependent erythroid lineage [69,70]. This is in contrast to another type of CBF AML, the t(3;21) which carries a translocation fusing the DNA binding domain of RUNX1 to the EVI1 oncogene, whose survival is dependent on GATA2 expression[3.]

In parallel with the above clustering analysis, we also directly identified mutation group-specific down-regulated DHSs (**Fig. SN3a**) by the same methodology as used for upregulated DHSs depicted in **Supplementary Fig 4a**. The motif analyses of these groups ((**Fig. SN3b**) confirmed the enrichment of GATA motifs in down-regulated DHSs in AML, together with a frequent loss of E-box motifs, which typically co-localize with GATA motifs in undifferentiated HSCs. DHSs lost in t(8;21) and inv(16) were enriched in HOX motifs, while IRF and/or NF-E2 motifs were enriched in several clusters. Interestingly, ETS, RUNX and AP-1 motifs were just as likely to be enriched in down-regulated DHSs (**Fig.SN3b**) as in up-regulated DHSs (**Supplementary Fig. 4b**).



Cluster 01

Motif	Match	P-value	%targets
ACTTCCTC	ETS	1e-78	59.5
AAACCACA	RUNX	1e-72	48.5
GATGAGTCAC	AP1	1e-52	21.2

Cluster 04

Motif	Match	P-value	%targets
ATGATTTATGGC	HOX	1e-108	40.3
GAGAGGAAAT	ETS	1e-51	52.9
TTGTGGTT	RUNX	1e-37	44.7

Cluster 05

Motif	Match	P-value	%targets
AGGAAGTA	ETS	1e-100	51.1
TCAGATAAGC	GATA	1e-42	41.3
TGTGGTTA	RUNX	1e-35	20.9

Cluster 20

Motif	Match	P-value	%targets
ATGATTTATGGC	HOX	1e-76	16.8
AAGAGGAAGT	ETS	1e-57	50.4
AAACCACA	RUNX	1e-54	31.9

Cluster 21

Motif	Match	P-value	%targets
ATCAGATAAGAA	GATA	1e-98	43.4
GACAGGAAGT	ETS	1e-72	52.5

Cluster 02

Motif	Match	P-value	%targets
GTGAGTCAC	AP1	1e-188	36.6
AACAGGAAGT	ETS	1e-98	42.4
ATTGCGTAAT	C/EBP	1e-92	22.3
TGTGGTTACG	RUNX	1e-48	25.6

Cluster 07

Motif	Match	P-value	%targets
ACAGGAAGCA	ETS	1e-125	60.6
ACAGATGTTT	E-box	1e-37	20.8
GAAACTGAAACC	IRF	1e-27	13.4
ACTTGTGGTT	RUNX	1e-26	22.0

Cluster 08

Motif	Match	P-value	%targets
CCACCAGGGGGC	CTCF	1e-260	30.6
AAGGGGAAGT	ETS	1e-30	29.1

Cluster 13

Motif	Match	P-value	%targets
ACTTCCTCTTTT	ETS	1e-172	47.8
GCTGAGTCAT	NFE2	1e-51	14.6
AGGAAACAGCTG	ETS:Ebox	1e-50	9.3
TGTGGTTT	RUNX	1e-33	20.8
CTAAAAATAG	MEF2A	1e-24	18.1
ACTTCTGC	E-box	1e-22	21.6

Figure SN3: Analysis of DHSs lost in AML-type specific DHS clusters as compared toprogenitor cells

(a) Heatmap depicting unsupervised K-mean clustering of the DNase-Seq log2 signals of DHS peaks in each AML sample which are either lost as compared to PBSCs (yellow) or shared with

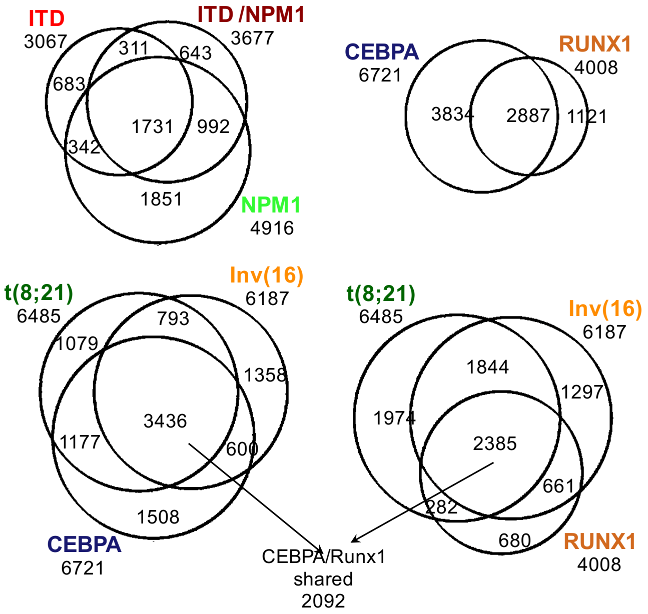
PBSC (blue). Clustering was done only on rows (DHS peaks) while samples were ranked based on the clustering in **Fig. 1c.**

(b) De novo motif search results for selected clusters as indicated.

(c) The percentage of DHS peaks that overlap with ATAC-Seq data from different progenitor types, the overlap was done between the DHS clusters from **Fig.SN2a** and each of the primary ATAC peaks; these include CLP, CMP, GMP, MPP, LMPP, MEP and Monocyte populations.

Overlaps between mutation-specific down-regulated DHSs

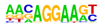
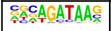
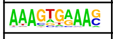
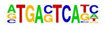
a



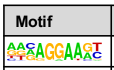
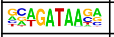
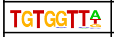
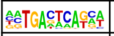
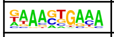
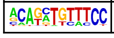
Motif enrichment analyses of mutation-specific down-regulated DHSs

b

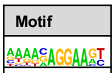
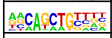
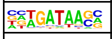
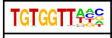
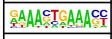
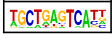
3067 ITD DHSs

Motif	Match	%
	ETS	64
	GATA	31
	IRF	17
	AP-1	14

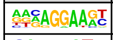
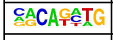
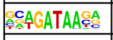
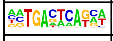
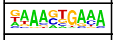
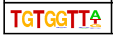
3677 ITD/NPM1 DHSs

Motif	Match	%
	ETS	64
	GATA	30
	RUNX	19
	AP-1	17
	IRF	14
	E-box	9

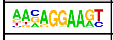
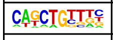
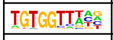
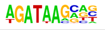
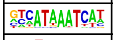
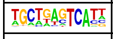
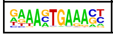
4916 NPM1 DHSs

Motif	Match	%
	ETS	72
	E-box	49
	GATA	29
	RUNX	21
	IRF	11
	NF-E2	10

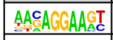
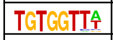
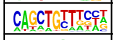
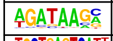
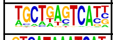
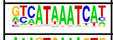
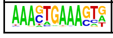
1731 shared ITD & NPM1

Motif	Match	%
	ETS	64
	E-box	63
	GATA	36
	AP-1	17
	IRF	14
	RUNX	9

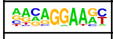
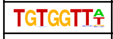
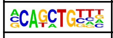
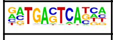
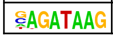
6485 t(8;21) DHSs

Motif	Match	%
	ETS	61
	E-box	53
	RUNX	40
	GATA	32
	HOX	21
	NF-E2	20
	IRF	7

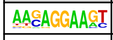
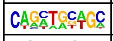
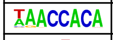
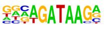
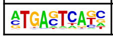
6187 Inv(16) DHSs

Motif	Match	%
	ETS	70
	RUNX	36
	E-box	31
	GATA	25
	NF-E2	17
	HOX	16
	IRF	8

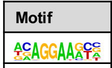
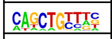
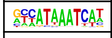
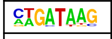
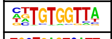
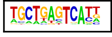
6721 CEBPA DHSs

Motif	Match	%
	ETS	50
	RUNX	36
	E-box	27
	AP-1	26
	GATA	13

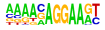
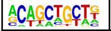
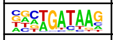
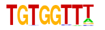
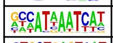
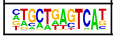
4008 RUNX1 DHSs

Motif	Match	%
	ETS	62
	E-box	33
	RUNX	39
	GATA	28
	AP-1	16

3436 CEBPA, Inv(16) and t(8;21) DHSs

Motif	Match	%
	ETS	69
	E-box	43
	HOX	31
	GATA	30
	RUNX	27
	NF-E2	19

2385 RUNX1, Inv(16) and t(8;21) DHSs

Motif	Match	%
	ETS	65
	E-box	52
	GATA	34
	RUNX	28
	HOX	22
	NF-E2	19

2887 CEBPA and RUNX1 DHSs

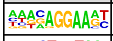
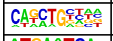
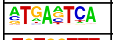
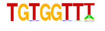
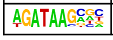
Motif	Match	%
	ETS	60
	E-box	60
	AP-1	37
	RUNX	32
	GATA	24

Figure SN4: **Analysis of down-regulated DHS groups.**

(a) Venn diagrams depicting the overlaps between mutation groups for DHSs that are down regulated compared to CD34+ve PBSCs. The down-regulated DHSs are defined as being at least 3-fold less than in PBSCs, for DHSs where the signal spanning a 400 bp window is at least 64 in PBSCs.
(b) De novo motif search results for the down regulated DHSs classes and in overlapping down-regulated DHSs for ITD and/or NPM1 and for CEBPA and RUNX1 that are shown in **Fig. SN2a**.

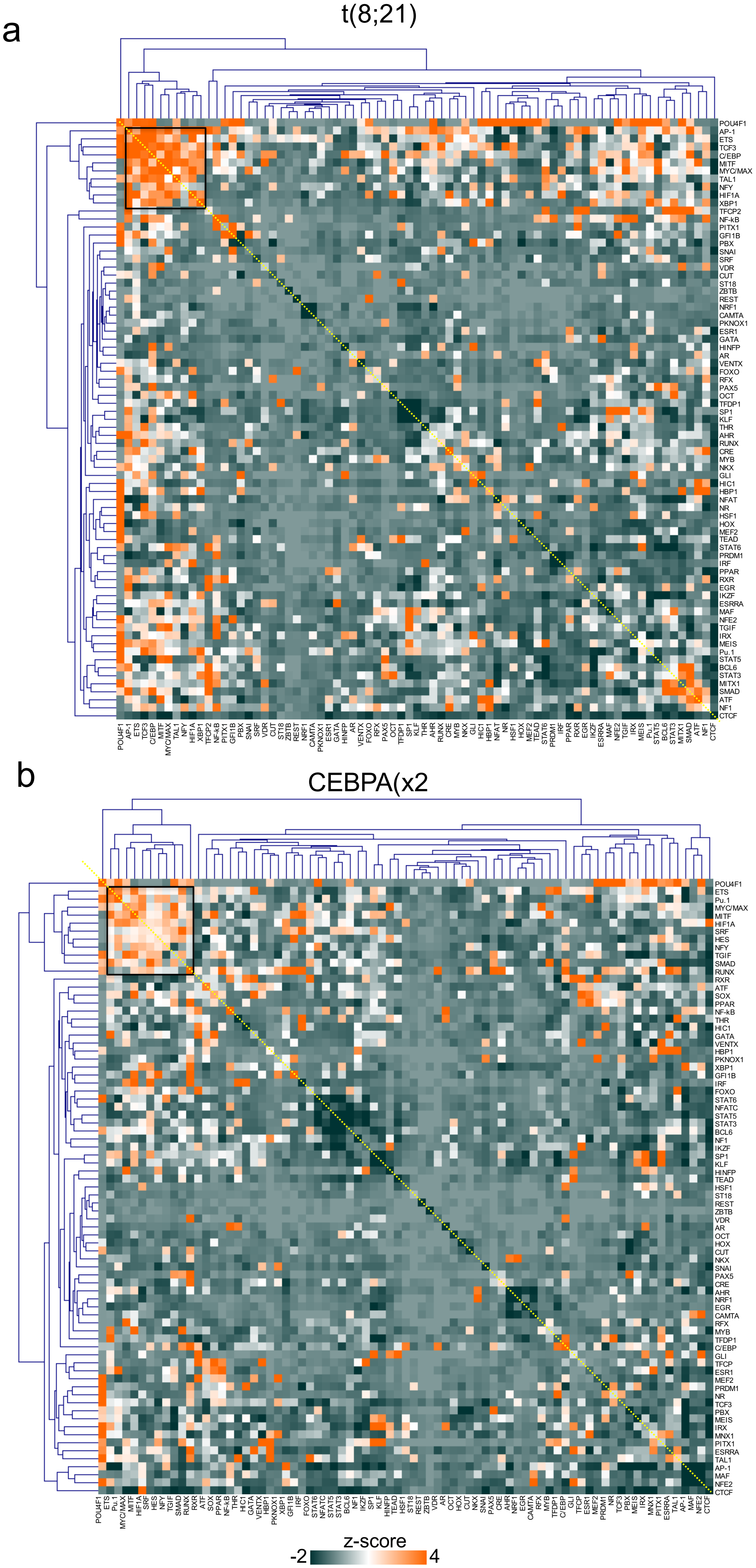
Analysis of transcription factor cooperation in different types of AML

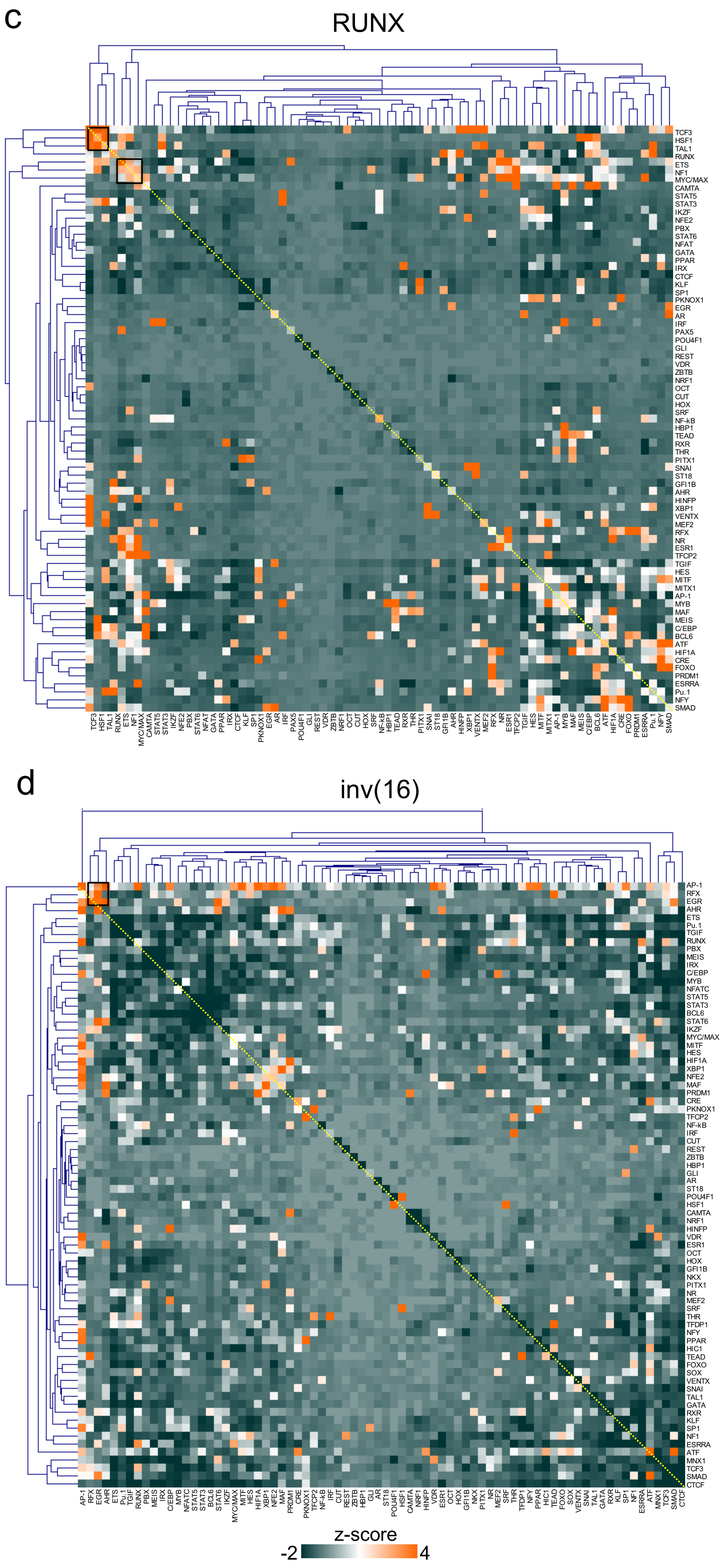
TFs do not work alone, but cooperate to form large cooperating complexes which may be potential target of therapeutic intervention using small molecule approaches as exemplified in the case of RUNX1 [71]. To examine whether it was possible to identify AML-type specific complexes in an unbiased way, we identified footprints within each mutation-specific AML group, but not in PBSCs, and used these sites to perform bootstrapping analyses searching for significant co-localizing occupied motifs within 50 bp for the main patient groups (**Fig SN4**, for the identity of motifs see **Table S**2). We have recently shown that such an analysis highlights the presence of co-localizing TF complexes identified by biochemical means such as pull-down or ChIP experiments. Such experiments showed that the product of the t(3;21) translocation, RUNX1-EVI1 associates with GATA2 and AP-1[72] or the complex assembled by RUNX1-ETO associates with ETS and E-box binding factors[73,74]. In both types of AML, knockdown of one component of the complex led to a reduction the growth of AML cells. We confirmed the co-localization of specific occupied motifs in the larger t(8;21) patient group analysed here (**Fig SN4a**). The occupied motifs for POU4F1 were not part of this cluster, suggesting that this factor is not part of a larger complex although its motif significantly co-localizes with a number of other factors individually. The CEBPA double mutant group (**Fig SN1b)** is characterized by the co-localization of occupied C/EBP and RUNX motifs, indicating that such cells contain still C/EBP binding activity. Again, occupied motifs for POU4F1 were not part of this cluster and co-localized with individual motifs. The RUNX1 group contained two co-localizing motif clusters, one consisting of E-box (TCF3) and ETS motifs and the other of NF1 and E-box (MYC/MAX) motifs (**Fig SN4c**), respectively.

Inv(16) deregulates CBF function and, similar to RUNX1-ETO, RUNX1/CBFβ-MYH11 also co-localizes with ETS/RUNX1 and E-box factors as shown by ChIP in cell lines [75]. However, our bootstrapping analysis of the inv(16) group in patient cells did not feature co-localisation of occupied RUNX motifs with other factors, but was associated with co-localising AP-1 motifs (**Fig SN3d**). This result could be explained by a different stability of the RUNX1/CBFβ-MYH11 complex in the nucleus as the ability to generate a footprint is directly correlated to the TF residence time [76].

The analysis of the ITD/NPM1 group revealed co-localizing NF1, AP-1, C/EBP and RUNX motifs (**Fig SN4e**) together with a number of co-localizing motif pairs which highlights the fact that

AML blasts from this type of AML show a signature specific for different differentiation stages (**Fig SN1b**). RUNX1 appears to occupy centre stage in this group. Its expression is up-regulated[77] and it cooperates with FLT3-ITD to establish AML in mice [78], indicating that blocking RUNX1 may be a valuable strategy for therapy for FLT3-ITD AML. Preliminary experiments indeed show that primary FLT3-ITD cells are indeed highly sensitive to small molecule inhibition[71] of RUNX1 (data not shown). Occupied AP-1 motifs were prominently present and clustered with C/EBP motifs, providing an additional explanation for the sensitivity of FLT3-ITD cells to AP-1 inhibition via dnFOS. Similarly, normal CD34+ PBSCs showed multiple enriched co-localizing TF motifs (such as RUNX, ETS and GATA motifs), highlighting the multi-lineage differentiation capacity and precursor composition of these cells (**Fig SN4f**). In summary, our analysis highlights that different combinations of TFs cooperate to drive the activity of AML subtype-specific regulatory modules and highlight those combinations that may be of therapeutic relevance. Our data are preliminary and need to be confirmed by biochemical studies, but we believe that such analyses could form the foundation of more detailed experiments investigating the relevance of specific TF interactions for cis-element activity, similar to what has been described in[3,5].


a
t(8;21)
POU4F1
AP-1
ETS
TCF3
C/EBP
MITF
MYC/MAX
TAL1
NFY
HIF1A
XBP1
TFCP2
NF-kB
PITX1
GFI1B
PBX
SNAI
SRF
VDR
CUT
ST18
ZBTB
REST
NRF1
CAMTA
PKNOX1
ESR1
GATA
HINFP
AR
VENTX
FOXO
RFX
PAX5
OCT
TFDP1
SP1
KLF
THR
AHR
RUNX
CRE
MYB
NKX
GLI
HIC1
HBP1
NFAT
NR
HSF1
HOX
MEF2
TEAD
STAT6
PRDM1
IRF
PPAR
RXR
EGR
IKZF
ESRRA
MAF
NFE2
TGIF
IRX
MEIS
Pu.1
STAT5
BCL6
STAT3
MITX1
SMAD
ATF
NF1
CTCF
b
CEBPA(x2
POU4F1
ETS
Pu.1
MYC/MAX
MITF
HIF1A
SRF
HES
NFY
TGIF
SMAD
RUNX
RXR
ATF
SOX
PPAR
NF-kB
THR
HIC1
GATA
VENTX
HBP1
PKNOX1
XBP1
GFI1B
IRF
FOXO
STAT6
NFATC
STAT5
STAT3
BCL6
NF1
IKZF
SP1
KLF
HINFP
TEAD
HSF1
ST18
REST
ZBTB
VDR
AR
OCT
HOX
CUT
NKX
SNAI
PAX5
CRE
AHR
NRF1
EGR
CAMTA
RFX
MYB
TFDP1
C/EBP
GLI
TFCP
ESR1
MEF2
PRDM1
NR
TCF3
PBX
MEIS
IRX
MNX1
PITX1
ESRRA
TAL1
AP-1
MAF
NFE2
CTCF
z-score
-2
4


c
RUNX
TCF3
HSF1
TAL1
RUNX
ETS
NF1
MYC/MAX
CAMTA
STAT5
STAT3
IKZF
NFE2
PBX
STAT6
NFAT
GATA
PPAR
IRX
CTCF
KLF
SP1
PKNOX1
EGR
AR
IRF
PAX5
POU4F1
GLI
REST
VDR
ZBTB
NRF1
OCT
CUT
HOX
SRF
NF-kB
HBP1
TEAD
RXR
THR
PITX1
SNAI
ST18
GFI1B
AHR
HINFP
XBP1
VENTX
MEF2
RFX
NR
ESR1
TFCP2
TGIF
HES
MITF
MITX1
AP-1
MYB
MAF
MEIS
C/EBP
BCL6
ATF
HIF1A
CRE
FOXO
PRDM1
ESRRA
Pu.1
NFY
SMAD
d
inv(16)
AP-1
RFX
EGR
AHR
ETS
Pu.1
TGIF
RUNX
PBX
MEIS
IRX
C/EBP
MYB
NFATC
STAT5
STAT3
BCL6
STAT6
IKZF
MYC/MAX
MITF
HES
HIF1A
XBP1
NFE2
MAF
PRDM1
CRE
PKNOX1
TFCP2
NF-kB
IRF
CUT
REST
ZBTB
HBP1
GLI
AR
ST18
POU4F1
HSF1
CAMTA
NRF1
HINFP
VDR
ESR1
OCT
HOX
GFI1B
NKX
PITX1
NR
MEF2
SRF
THR
TFDP1
NFY
PPAR
HIC1
TEAD
FOXO
SOX
VENTX
SNAI
TAL1
GATA
RXR
KLF
SP1
NF1
ESRRA
ATF
MNX1
TCF3
SMAD
CTCF
z-score
-2
4

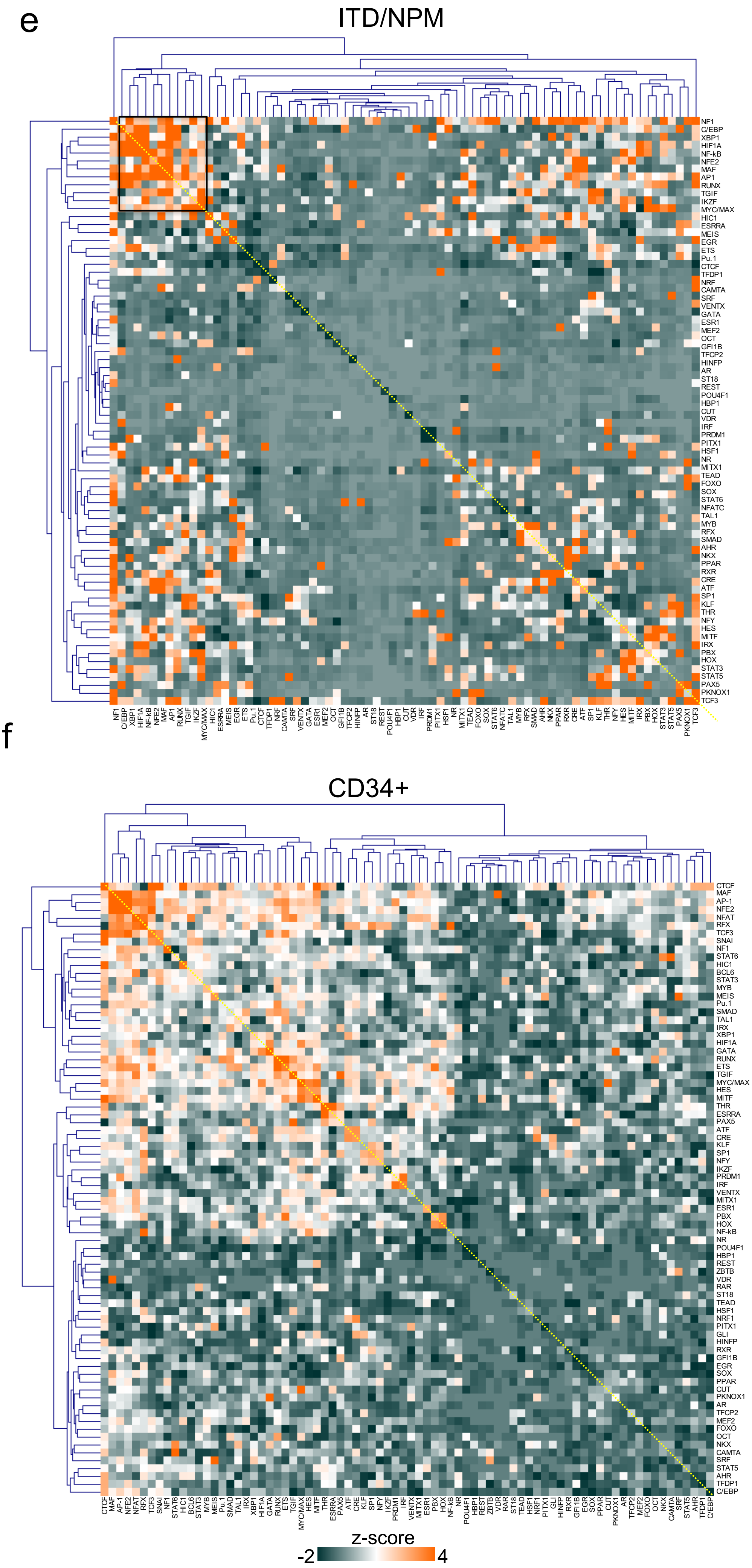
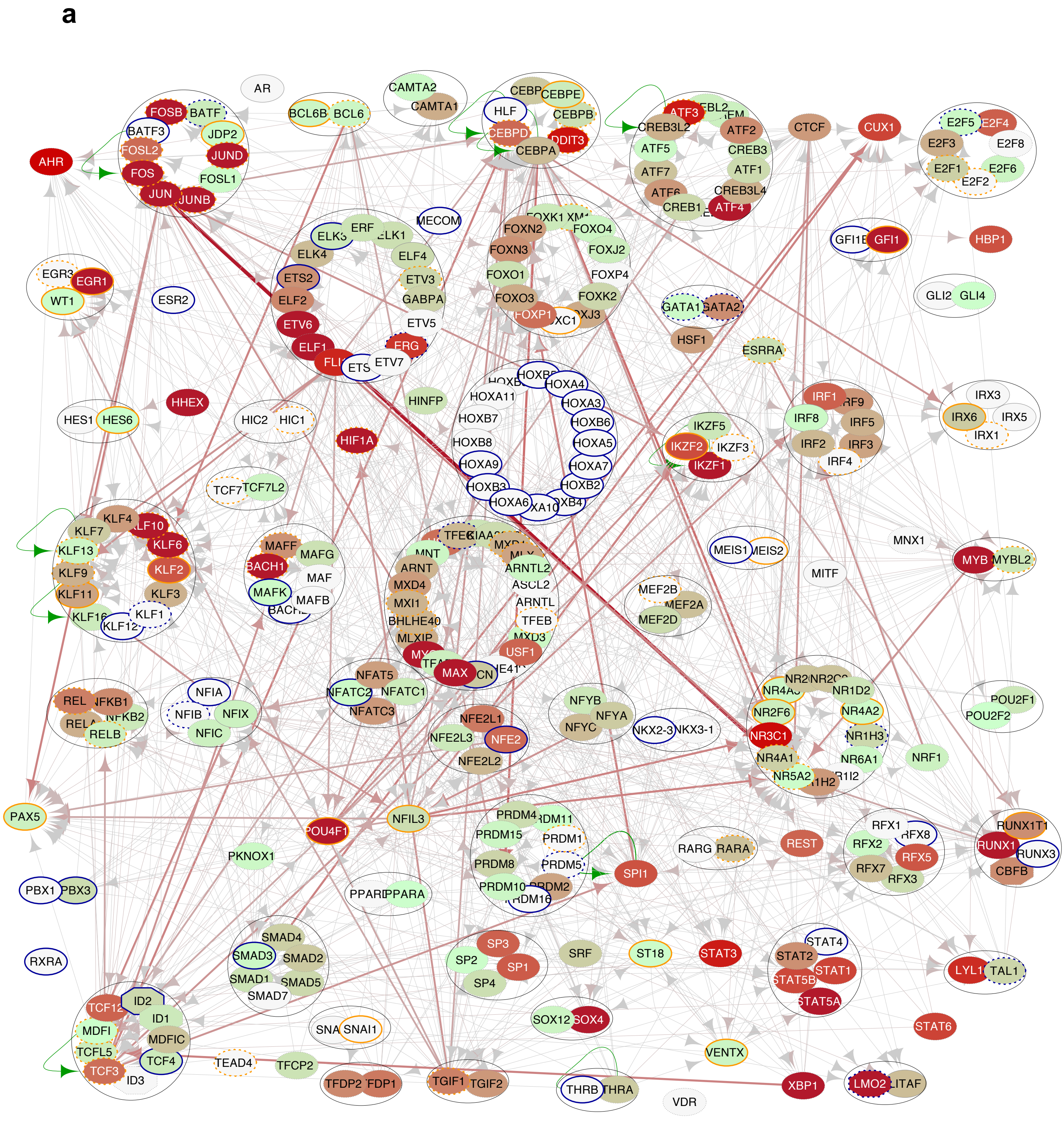

e
ITD/NPM
NF1
C/EBP
XBP1
HIF1A
NF-kB
NFE2
MAF
AP1
RUNX
TGIF
IKZF
MYC/MAX
HIC1
ESRRA
MEIS
EGR
ETS
Pu.1
CTCF
TFDP1
NRF
CAMTA
SRF
VENTX
GATA
ESR1
MEF2
OCT
GFI1B
TFCP2
HINFP
AR
ST18
REST
POU4F1
HBP1
CUT
VDR
IRF
PRDM1
PITX1
HSF1
NR
MITX1
TEAD
FOXO
SOX
STAT6
NFATC
TAL1
MYB
RFX
SMAD
AHR
NKX
PPAR
RXR
CRE
ATF
SP1
KLF
THR
NFY
HES
MITF
IRX
PBX
HOX
STAT3
STAT5
PAX5
PKNOX1
TCF3
f
CD34+
CTCF
MAF
AP-1
NFE2
NFAT
RFX
TCF3
SNAI
NF1
STAT6
HIC1
BCL6
STAT3
MYB
MEIS
Pu.1
SMAD
TAL1
IRX
XBP1
HIF1A
GATA
RUNX
ETS
TGIF
MYC/MAX
HES
MITF
THR
ESRRA
PAX5
ATF
CRE
KLF
SP1
NFY
IKZF
PRDM1
IRF
VENTX
MITX1
ESR1
PBX
HOX
NF-kB
NR
POU4F1
HBP1
REST
ZBTB
VDR
RAR
ST18
TEAD
HSF1
NRF1
PITX1
GLI
HINFP
RXR
GFI1B
EGR
SOX
PPAR
CUT
PKNOX1
AR
TFCP2
MEF2
FOXO
OCT
NKX
CAMTA
SRF
STAT5
AHR
TFDP1
C/EBP
z-score
-2
4

Figure SN5: Significance of co-localizing occupied TF binding motifs as indicated by bootstrapping analysis of the indicated mutation groups (a – f). In this analysis we determine the significance of co-localization of the indicated occupied factor binding motifs within 50 bp of sequence within specific DHSs as compared to the union of all DHSs. The heatmap shows hierarchical clustering of footprinted motif co-occurrences by z-score within AML specific DHSs. The motif search was done within footprint coordinates and motif frequencies were calculated within a window of 50 bp. Orange and green colours indicate statistically over- and underrepresented motif co-occurrences, respectively. Only motifs occurring >50 times were considered. Motifs in AML cells forming a cluster underlying a potential complex are boxed in.

Identification of regulators of leukemogenesis and leukemia maintenance using footprinting-based TF network construction

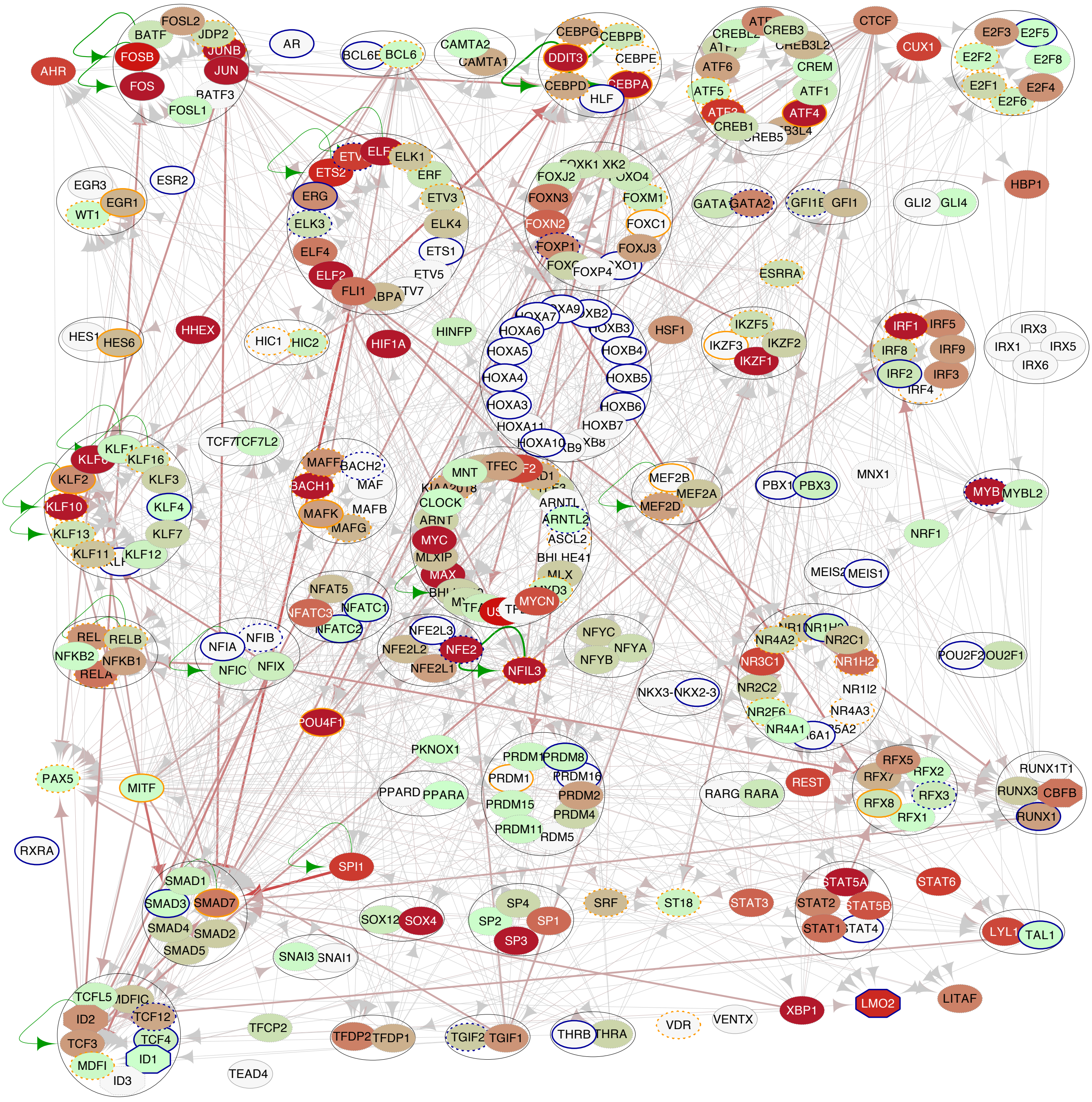
Besides the regulators and regulator families described in the main text (POU4F1, NFIX, FOXC1 and AP-1), our analysis highlighted a number of transcription factor genes that form regulatory nodes in AML subtype specific networks of up-regulated TF genes and appear to have AML-specific roles (**Fig 6, Fig. SN6 a-f**). Some of these factors had already been identified to play a role in AML in general, but had not yet been assigned to a specific type of AML. For example, the homeobox gene *VENTX* which forms a node in the t(8;21) and FLT3-ITD/NPM1 TF networks had been shown to promote myeloid proliferation and also to cooperate with RUNX1-ETO, the product of the t(8;21)[79]. *VENTX* forms a node in FLT3 ITD/NPM1 cells as well, but not in the CEBPA (x2) AML where this gene is not expressed. Another node in the t(8;21) is *PAX5* which is responsible for the up-regulation of the B-cell marker CD19 in these cells[80]. It has previously been shown that POU4F1 activated *PAX5*[81] together with other B cell specific gens*,* but our network analysis shows a number of additional links to this gene. The most prominent link is coming from the AP-1 factor family which could explain why the expression of this gene responds to signalling modulation[82]. This factor family is also responsible for the up-regulation of the *NRF5A2* gene in the t(8;21) and RUNX1 mutant AML which has been shown to be required to restrict the inflammatory response in pancreatic cells[83]. This again highlights the activation of multiple signalling pathways in this type of AML and at the same time indicates the potential activation of a feedback mechanism that may dampen excessive inflammatory signals in these cells. Last, but not least the FLT3-ITD/NPM1 network shows highly specific *IRX3* expression as a major node whose expression which is again linked to the AP-1 family and to the expression of the homeobox gene PBX3. Similar to *FOXC1, a*ctivation of this gene has been shown to be linked to a differentiation block in myeloid cells [84]. Another interesting observation concerns the role of the *NFIL3* gene encoding the bZIP protein NFIL3 which is up-regulated in the majority of all our AML patients and forms a major node with multiple, but differential links in all our networks. NFIL3 plays multiple roles in the immune system, in particular in lymphoid development[85], where it up-regulates cytokine genes such as IL-3[86] but also in other tissues. So far no connection was seem to AML, but our data would predict that this gene is a major coordinator of signalling processes in the myeloid lineage as it is linked with members of the AP-1 factor family as well as with CEBP family members,

Our analysis of the larger AML-type specific TF networks where we did not restrict our analysis to up-regulated genes (**Fig. SN5** depicting links specific for AML subtype specific genes) or the full network highlighted a number of additional interesting observations (links between all TF genes in all AML subtypes and CD34+ PBSCs, are shown on our webserver (see URL section)). We observed that the Aryl-hydrocarbon-receptor (*AHR*) formed a major node in all AML subtypes and shows a link to the AP-1 family. AHR has been implicated to control the balance between stem cell proliferation and quiescence[87] and inhibitors of this factor have been used to maintain proliferating stem cells in culture[88]. It is tempting to speculate that this factor performs the same control function in leukemic cells. In summary, our analyses provide ample opportunities for hypothesis testing. The next step will be, to link the transcription factor networks to the up- or down-regulation of effector genes and exploit this information in systems pharmacology approaches.



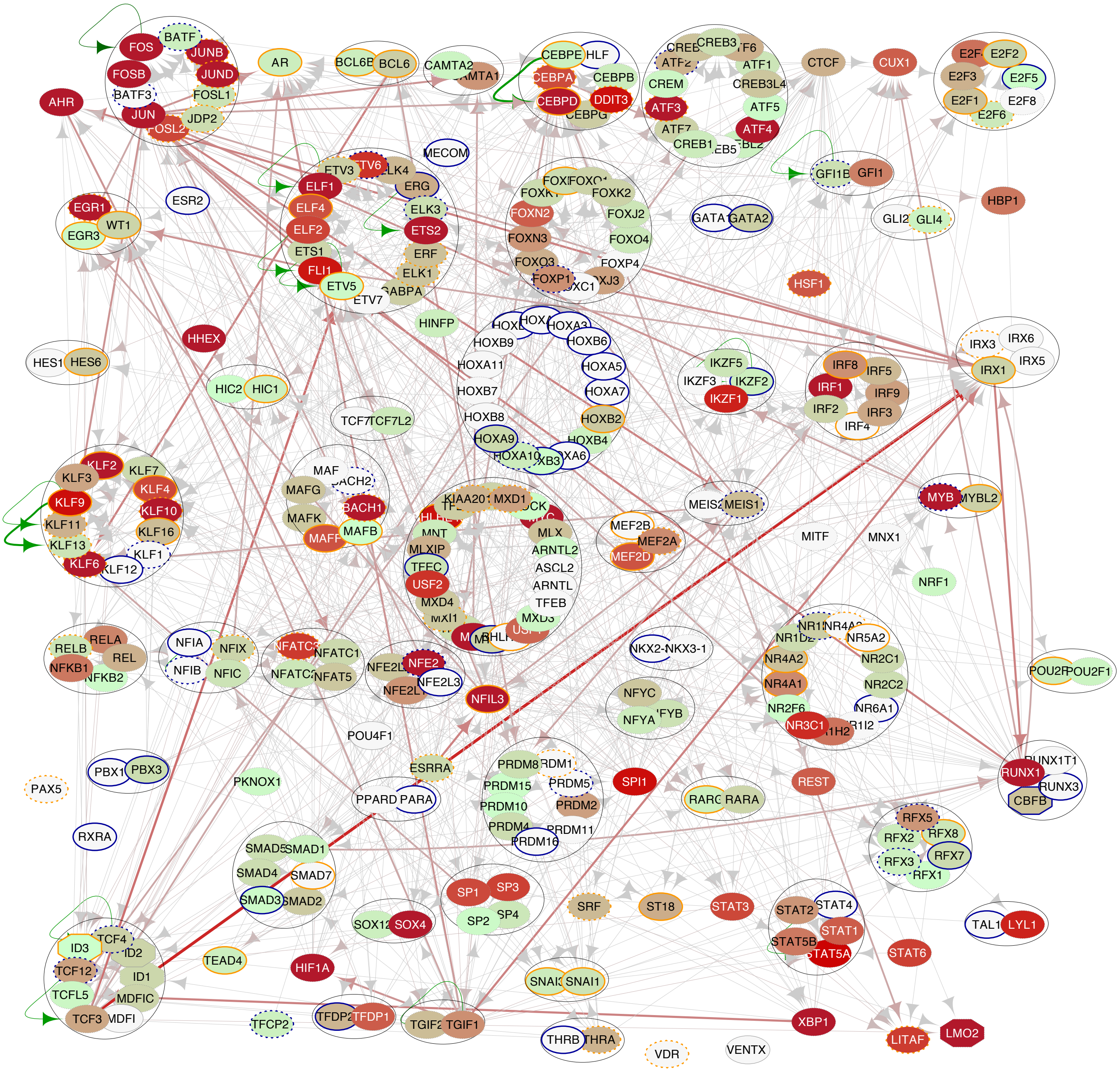
t(8;21)-specific TF network

b

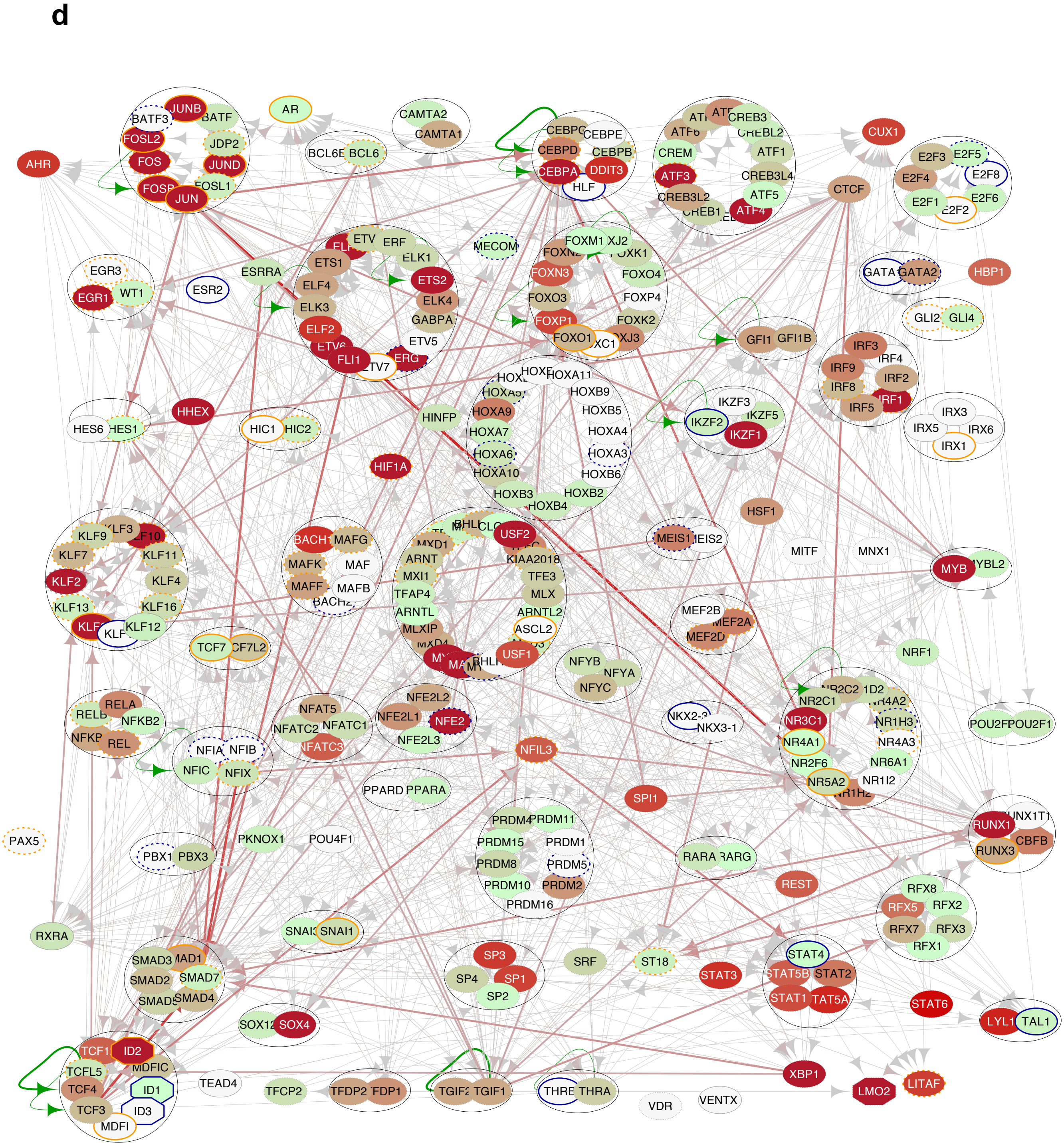


CEBPA(x2)-specific TF network

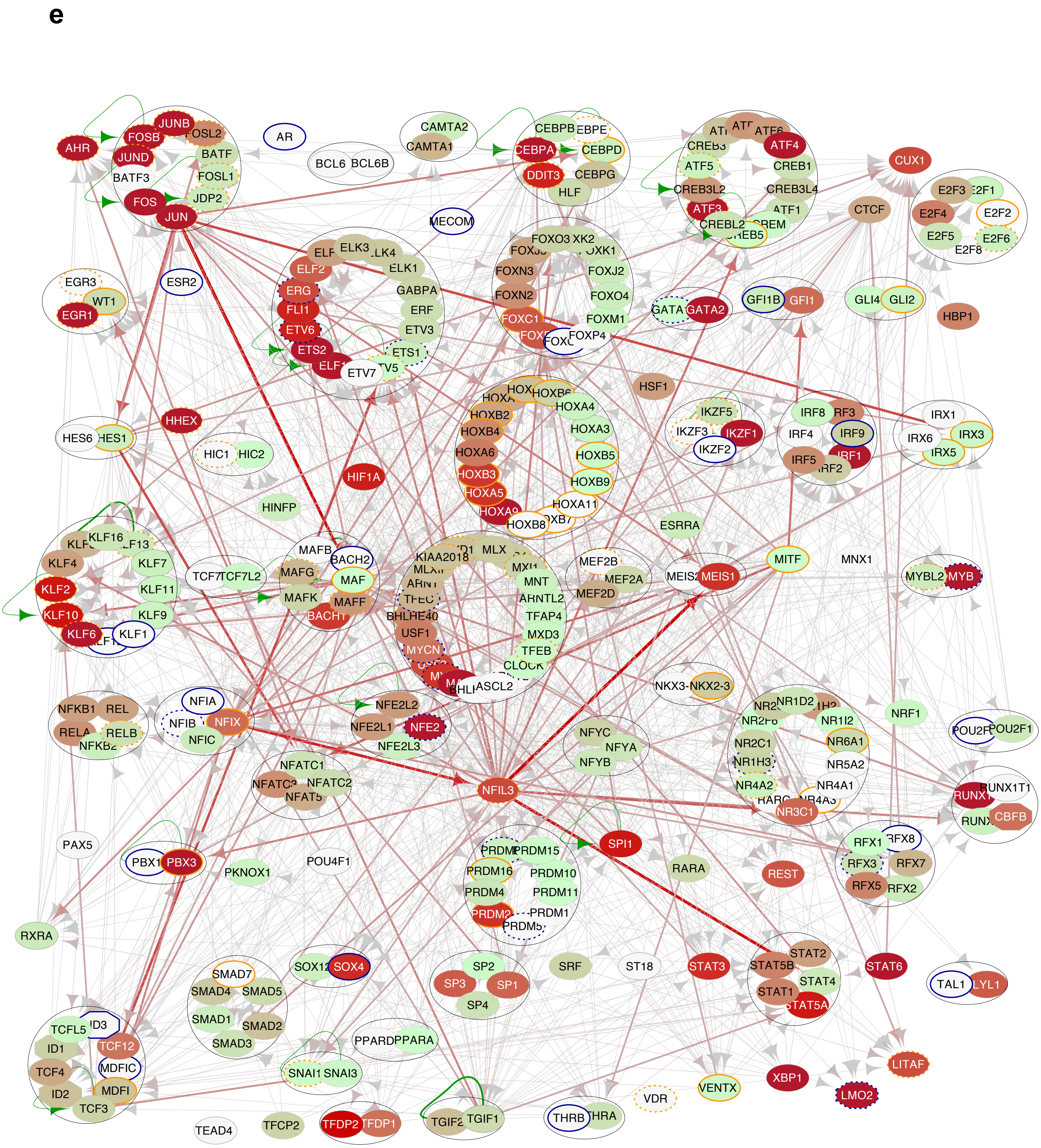
c



Inv(16) specific TF network

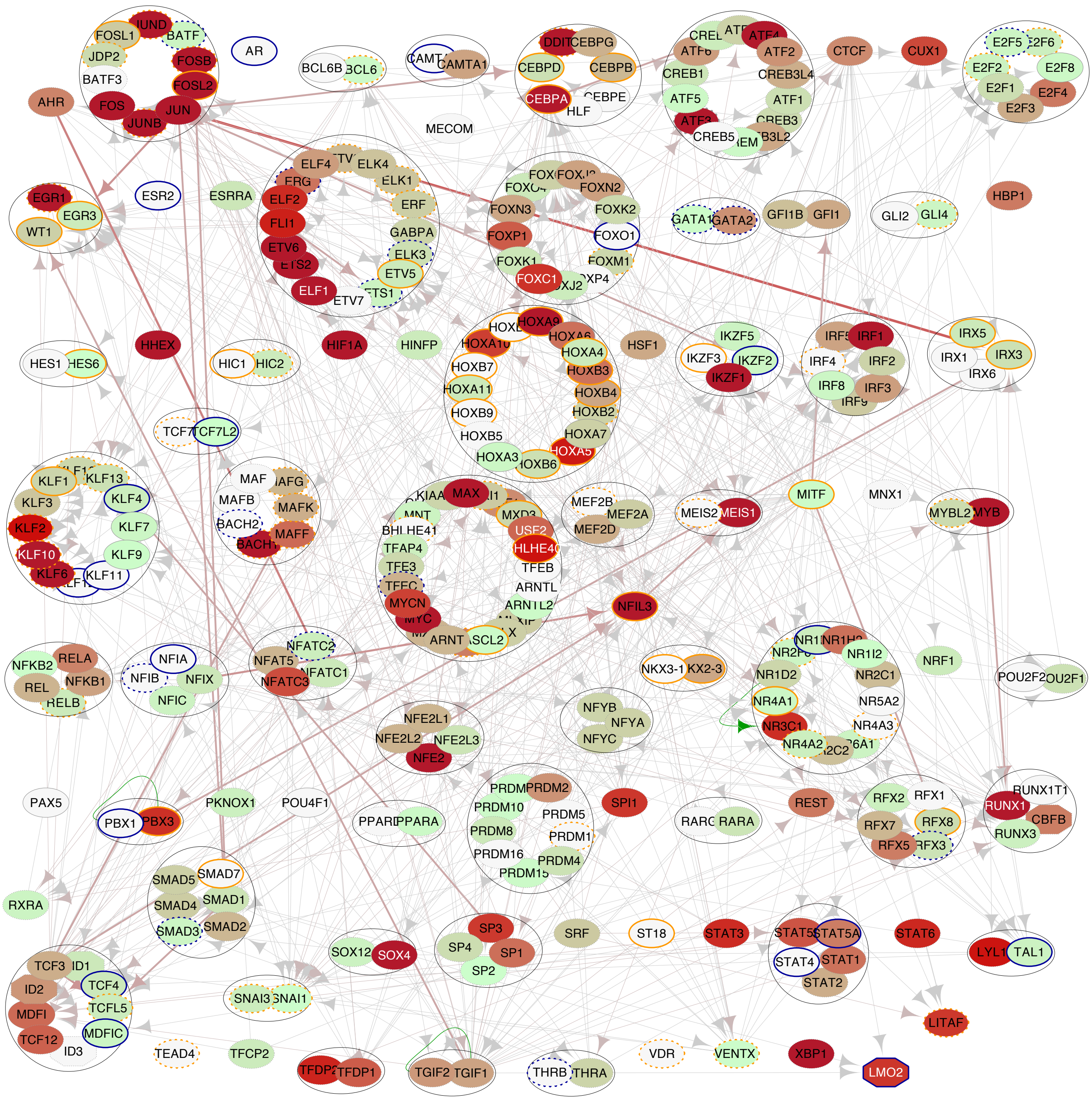


Mutant RUNX1-specific TF network



FLT3-ITD/NPM1 specific TF network

f



NPM1-specific TF network

Figure SN6: Identification of transcription factor networks driving the expression of TF genes in each AML subtype. Here we plotted links connecting all TF encoding genes which are specific for the respective AML subtype irrespective whether they were up-regulated or not. Top panels: Analysis strategy. (a) t(8;21)-specific TF network, (b) CEBPA(x2)-specific TF network, (c) Inv(16) specific TF network, (d) Mutant RUNX1-specific TF network, e) FLT3-ITD/NPM1 specific TF network, (f) NPM1-specific TF network. Factor families binding to the same motif as shown in Table S2 form a node contained within a circle. Arrows going outwards from the entire node highlight footprinted motifs in individual genes generated by any member of this factor family whereby the footprint was annotated to the gene using the CHiC data where possible, otherwise to the nearest gene. The expression level (FKPM) for the individual genes is depicted in white (low)/red (high) colour. An orange smooth ring around the circle indicates that this gene is specifically up-regulated in this type of AML compared to CD34+ PBSCs and/or other AML types, a dotted circle indicates a gene that is up-regulated as compared to CD34+ cells. Genes with no outgoing arrows due to a lack of know binding motifs are highlighted by their octagon shapes. For a detailed guide to node and edge attributes: See **Fig. 6**.

3. ADDITIONAL SUPPLEMENTARY BIOINFORMATICS METHODS

Further DNaseI-Seq data analysis

K-mean clustering of AML specific DHSs: A combined set of up-regulated distal DHSs that defined as being at least 3-fold greater than in PBSCs was used to perform unsupervised k-mean clustering. The number of reads that mapped to these peaks was counted in a 400bp window centered on the DHS summit, and subsequently normalized to total sample size using DEseq2[89]. Clustering was done on rows (DHSs) while samples (columns) were ranked based on the hierarchical clustering in **Fig. 1c**. Initially the read counts output from DEseq2 [89] was further quartile normalised using the "preprocessCore" package in R, The log2 of the normalised reads were clustered using *k*-means clustering with Euclidean distances (*stats* package in R) and the optimal number of clusters was determined to be 20 based on the lowest Bayesian Information Criterion (BIC) scores was generated using R. Each of the 20 clusters was then hierarchically clustered using the "complete linkage" agglomeration method.

ATAC sequencing data analysis

ATAC-seq profiles of hematopoietic and leukemic cell types taken [67] were downloaded from GEO with accession number GSE74912*.* ATAC-seq data of HSC, MPP, CMP, CLP, MEP, GMP and Monocytes were downloaded and aligned to the human genome version hg38. Aligned reads with the same cell line were merged and then ATAC peaks were obtained using MACS2 with default parameter. Overlaps between DHS and ATAC peaks were defined by requiring the summits of two peaks to lie within +/-200 bp. Pair-wise peak overlaps between DHSs and ATAC peaks of hematopoietic *i* and *j* were performed in order to calculate the fraction (M_{ij})

$M_{ij} = \frac{N_{ij}}{N_i}$ where *N_{ij}* is the total peaks that overlap, *N_i* is the total number peaks in set i (DHSs) and *N_j* is the total peaks in j (ATAC). A matrix with the calculated fraction multiply by 100 was generated and a heatmap was plotted (**Fig. 2b**) after hieratically clustered in R. Clustering of DNaseI-seq and ATAC-seq samples (**Fig. S2a** and **Fig. S3b**) was carried out using the merged distal DHSs as described earlier using the DNaseI-seq only.

ChIP sequencing data analysis

ChIP-Seq sequencing reads were downloaded from GEO with accession numbers (GSM1581788, GSM1693378, GSM1466000)[77] (GSM722705, GSM722704)[90], the reads were aligned to the human genome version hg38 with Bowtie version 2.3.1[91]. Reads that mapped uniquely to the genome were retained and duplicated reads were removed using the MarkDuplicates function in Picard tools (http://broadinstitute.github.io/picard/). Peaks were identified with MACS version 1.4.2 [92] and DFilter software[93] with recommended parameters (-bs=100 -ks=50 –refine). Peaks common to both peak calling methods were considered for further analysis.

H3K27Ac ChIP data analysis

H3K27Ac ChIP data from[94] were downloaded from NCBI with accession number SRP103200. The raw reads were aligned to the human reference genome hg38 and density profiles were generated using *bedtools. The bedGraph* files were used to generate the H3K27Ac average coverage plotted a long side the DHSs of the 20 clusters (**Fig.S3d).**

Digital genomic footprinting

Digital genomic footprinting was performed using the *Wellington_footprints* function of the Wellington algorithm[95] on High-depth AML and CD34+ PBSC DHSs. DHS footprints probability and DNase forward and reverse cut coverages, were generated using the *dnase_wig_tracks* function of Wellington. AML-specific footprints compared to PBSC CD34+ cells were identified using *wellington_bootstrap* function of Wellington. Mutation-specific footprints of the groups were identified by using the *Wellington_footprints* function using the merged reads of the Mutation-specific individual DNaseI-Seq of each group.

Motif identification

De novo motif analysis was performed on peaks using HOMER[68]. Motif lengths of 6, 8, 10, and 12 bp were identified in within ± 200 bp from the peak summit. The annotatePeaks function in HOMER was used to find occurrences of motifs in peaks. In this case we used known motif position weight matrices (PWM).

Motif co-localisation clustering: Motif co-localisation clustering was performed as previously described[73]. A motif position search was done within DHSs that are group mutation-specifically footprinted. The distance between the centres of each motif pairs was calculated and the motif frequency was counted if the first motif was within 50bps distance from the second motif. Z-scores were calculated from the mean and standard deviation of motif frequencies observed in random sets using bootstrap analysis, peak sets with a population equal to that of the footprinted peaks were randomly obtained from the merged footprints of all AML and CD34+ footprints sets. Motif search and motif frequencies calculations were repeated 1000 times for each random set. A matrix was generated and Z-scores were displayed after hierarchical clustering as a heatmap with R.

Motif enrichment

To identify motifs that are relatively enriched in the distal footprinted DHSs of each of AML mutation groups (**Fig. S5**) and the AML DHSs clusters **(Fig. 2A).** For a given set *j* of footprints, we defined a motif enrichment score (ES_{ij}) for motif *i* in footprint set *j* as

$ES_{ij} = \frac{n_{ij}/M_j}{\sum_j n_{ij}/\sum_j M_j}$ where n_{ij} is the number of footprints in each subset *j (j=1,2,…,12)* containing motif *i (i=1, 2,….,I), I* is the total number of motifs used in the test, and M_j the total number of peaks in each subset *j (j=1,2,…,30).* A matrix was generated and the motif enrichment scores were displayed as a heatmap after hierarchical clustering with Euclidean distance and complete linkage. The heatmap was generated using R. The statistical significance for a ES_{ij} score of a given motif *i* in peak set *j* is computed as Z-scores using bootstrapping (N=1000), where a random set of peaks is extracted from a global set of footprinted regions and *ES* is calculated. After N iterations the mean (μ_{ij}) and the standard deviation (σ_{ij}) are computed and the z-scores are computed as $Z_{ij} = \frac{ES_{ij} - \mu_{ij}}{\sigma_{ij}}$. The global set of regions is a merged set of all the AML footprints. These Z-scores are provided in **Supplementary Dataset S3**. Similarly the motif enrichment score displayed in **Fig. 4a** was calculated using the ES_{ij} function. Initially the AML mutant specific footprints were overlapped with each differentiated stages of the progenitors ATAC-Seq,peaks. Motif searches were conducted within the coordinates of overlapping FPs. A matrix was generated with motif enrichment score ES_{ij} using FPs that overlapped with ATAC-seq peaks of the HSC, MPP, LMPP, MEP, CLP, CMP, GMP and Mono populations. Motifs with a number less than 20 were discharged. The statistical significance for an ES_{ij} score of a given motif *i* in FP set *j* is computed as Z-scores. The motif scores were displayed as a heatmap after hierarchical clustering with Euclidean distance and complete linkage.

Gene expression analysis

Differentially expressed genes were extracted using the limma R package[96]. Genes were said to be differentially expressed (DE) if there was a twofold change in expression between any each of the AML patient sample or each of the mutation-specific group and the PBSC CD34+ with a *p*-value less than or equal to 0.01 and with FPKM greater than 1 in at least one AML sample. For each value of a DE gene a pseudo-count $\gamma = 0.1$ was added to the FPKM values and the binary logarithm of this value was considered as the expression value of the gene in each sample (*j*), $e_{ij} = log_2(FPKM_{ij} + \gamma)$. These DE values were then clustered **(Supplementary Fig. 1a)** using hierarchical clustering with Euclidean distances (*stats* package in R). While Hierarchical clustering of transcription factors gene expression was carried out on fold-changes for genes associated with at least a 2-fold change compared to the CD34+.

Gene set enrichment analysis

A publically available RNA-seq data of hematopoietic cell types were downloaded from GEO with accession number GSE74246*.* The downloaded RNA-seq data were processed in similar way as described above. The GSEA software[97] was used to perform gene set enrichment analysis on group of genes. Module map[98] implemented by Genomic software was used to find which groups of genes are significantly up- or down-regulated using a statistical test based on the hyper-geometric distribution the fraction of up or down regulated is displayed as a heatmap (**Fig 2c** and **Supplementary Fig 4c**).

Gene ontology (GO) analysis: Gene ontology (GO) analysis was performed using clueGO tools[97] with Hypergeometric for overrepresentation and Benjamini and Hochberg (FDR) correction for multiple testing corrections. KEGG Pathway network analysis was performed using clueGO tools[97] with kappa score = 0.3. A right-sided enrichment (depletion) test based on the hypergeometric distribution was used for terms and groups. The size of the nodes reflects the number of genes within the term. The color of nodes reflects the enrichment significance of the terms. The network is laid out using Cytoscape. The KEGG pathway network figures for all DHS-cluster associated genes are shown in Table S6.

Expression profiles from larger patient cohort datasets

Microarray data from Verhaak et al.[99] were downloaded from GEO under the accession number GSE6891. Patients were split according to their mutational status; Boxplots showing the expression of the indicated genes in FLT3-ITD, NPM1, CEBPA, t(8;21), inv(16) and NRAS mutation groups. The statistical significance of the difference in expression between FLT3-ITD and other mutations was determined using an unpaired t-test.

4. SUPPLEMENTAL MATERIALS: REFERENCES

66. Jolma, A. *et al.* DNA-binding specificities of human transcription factors. *Cell* **152**, 327-39 (2013).

67. Corces, M.R. *et al.* Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nat Genet* **48**, 1193-203 (2016).

68. Heinz, S. *et al.* Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Molecular Cell* **38**, 576-589 (2010).

69. Wilson, N.K. *et al.* Combinatorial transcriptional control in blood stem/progenitor cells: genome-wide analysis of ten major transcriptional regulators. *Cell Stem Cell* **7**, 532-44 (2010).

70. Fujiwara, T. *et al.* Discovering hematopoietic mechanisms through genome-wide analysis of GATA factor chromatin occupancy. *Mol Cell* **36**, 667-81 (2009).

71. Illendula, A. *et al.* Small Molecule Inhibitor of CBFbeta-RUNX Binding for RUNX Transcription Factor Driven Cancers. *EBioMedicine* **8**, 117-131 (2016).

72. Loke, J. *et al.* RUNX1-ETO and RUNX1-EVI1 Differentially Reprogram the Chromatin Landscape in t(8;21) and t(3;21) AML. *Cell Rep* **19**, 1654-1668 (2017).

73. Ptasinska, A. *et al.* Identification of a Dynamic Core Transcriptional Network in t(8;21) AML that Regulates Differentiation Block and Self-Renewal. *Cell Reports* **8**, 1974-1988 (2014).

74. Sun, X.J. *et al.* A stable transcription factor complex nucleated by oligomeric AML1-ETO controls leukaemogenesis. *Nature* **500**, 93-7 (2013).

75. Mandoli, A. *et al.* CBFB-MYH11/RUNX1 together with a compendium of hematopoietic regulators, chromatin modifiers and basal transcription factors occupies self-renewal genes in inv(16) acute myeloid leukemia. *Leukemia* **28**, 770-8 (2014).

76. Sung, M.H., Guertin, M.J., Baek, S. & Hager, G.L. DNase footprint signatures are dictated by factor dynamics and DNA sequence. *Mol Cell* **56**, 275-85 (2014).

77. Cauchy, P. *et al.* Chronic FLT3-ITD Signaling in Acute Myeloid Leukemia Is Connected to a Specific Chromatin Signature. *Cell Rep* **12**, 821-36 (2015).

78. Behrens, K. *et al.* RUNX1 cooperates with FLT3-ITD to induce leukemia. *J Exp Med* **214**, 737-752 (2017).

79. Gentner, E. *et al.* VENTX induces expansion of primitive erythroid cells and contributes to the development of acute myeloid leukemia in mice. *Oncotarget* **7**, 86889-86901 (2016).

80. Walter, K. *et al.* Aberrant expression of CD19 in AML with t(8;21) involves a poised chromatin structure and PAX5. *Oncogene* **29**, 2927-37 (2010).

81. Dunne, J. *et al.* AML1/ETO and POU4F1 synergy drives B-lymphoid gene expression typical of t(8;21) acute myeloid leukemia. *Leukemia* **26**, 1131-5 (2012).

82. Ray, D. *et al.* Lineage-inappropriate PAX5 expression in t(8;21) acute myeloid leukemia requires signaling-mediated abrogation of polycomb repression. *Blood* **122**, 759-69 (2013).

83. Cobo, I. *et al.* Transcriptional regulation by NR5A2 links differentiation and inflammation in the pancreas. *Nature* **554**, 533-537 (2018).

84. Somerville, T.D.D. *et al.* Derepression of the Iroquois Homeodomain Transcription Factor Gene IRX3 Confers Differentiation Block in Acute Leukemia. *Cell Rep* **22**, 638-652 (2018).

85. Male, V., Nisoli, I., Gascoyne, D.M. & Brady, H.J. E4BP4: an unexpected player in the immune response. *Trends Immunol* **33**, 98-102 (2012).

86. Zhang, W. *et al.* Molecular cloning and characterization of NF-IL3A, a transcriptional activator of the human interleukin-3 promoter. *Mol Cell Biol* **15**, 6055-63 (1995).

87. Singh, K.P., Casado, F.L., Opanashuk, L.A. & Gasiewicz, T.A. The aryl hydrocarbon receptor has a normal function in the regulation of hematopoietic and other stem/progenitor cell populations. *Biochem Pharmacol* **77**, 577-87 (2009).

88. Boitano, A.E. *et al.* Aryl hydrocarbon receptor antagonists promote the expansion of human hematopoietic stem cells. *Science* **329**, 1345-8 (2010).

89. Love, M.I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550 (2014).

90. Ptasinska, A. *et al.* Depletion of RUNX1/ETO in t(8;21) AML cells leads to genome-wide changes in chromatin structure and transcription factor binding. *Leukemia* **26**, 1829-1841 (2012).

91. Langmead, B. & Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357-9 (2012).

92. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**, R137 (2008).

93. Kumar, V. *et al.* Uniform, optimal signal processing of mapped deep-sequencing data. *Nat Biotechnol* **31**, 615-22 (2013).

94. McKeown, M.R. *et al.* Superenhancer Analysis Defines Novel Epigenomic Subtypes of Non-APL AML, Including an RARalpha Dependency Targetable by SY-1425, a Potent and Selective RARalpha Agonist. *Cancer Discov* **7**, 1136-1153 (2017).

95. Piper, J. *et al.* Wellington: a novel method for the accurate identification of digital genomic footprints from DNase-seq data. *Nucleic Acids Res* **41**, e201 (2013).

96. Ritchie, M.E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* **43**, e47 (2015).

97. Bindea, G. *et al.* ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics* **25**, 1091-3 (2009).

98. Segal, E., Friedman, N., Koller, D. & Regev, A. A module map showing conditional activity of expression modules in cancer. *Nat Genet* **36**, 1090-8 (2004).

99. Verhaak, R.G.W. *et al.* Prediction of molecular subtypes in acute myeloid leukemia based on gene expression profiling. *Haematologica* **94**, 131-134 (2009).