Cell Reports

Supplemental Information

Chronic FLT3-ITD Signaling in Acute Myeloid Leukemia Is Connected to a Specific Chromatin Signature

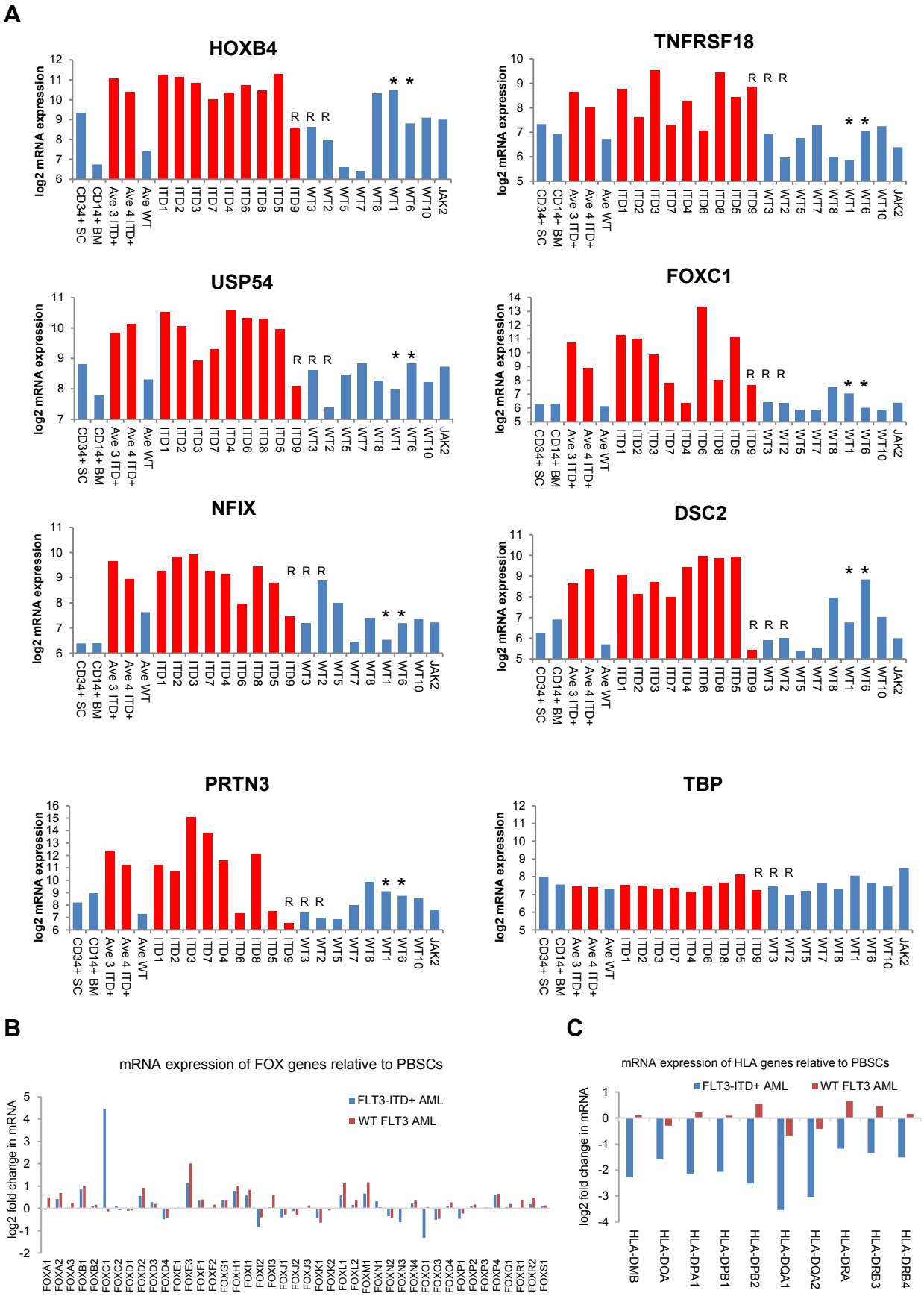
Pierre Cauchy, Sally R. James, Joaquin Zacarias-Cabeza, Anetta Ptasinska, Maria Rosaria Imperato, Salam A. Assi, Jason Piper, Martina Canestraro, Maarten Hoogenkamp, Manoj Raghavan, Justin Loke, Susanna Akiki, Samuel J. Clokie, Stephen J. Richards, David R. Westhead, Michael J. Griffiths, Sascha Ott, Constanze Bonifer, and Peter N. Cockerill

Table S1: List of genes screened for mutations

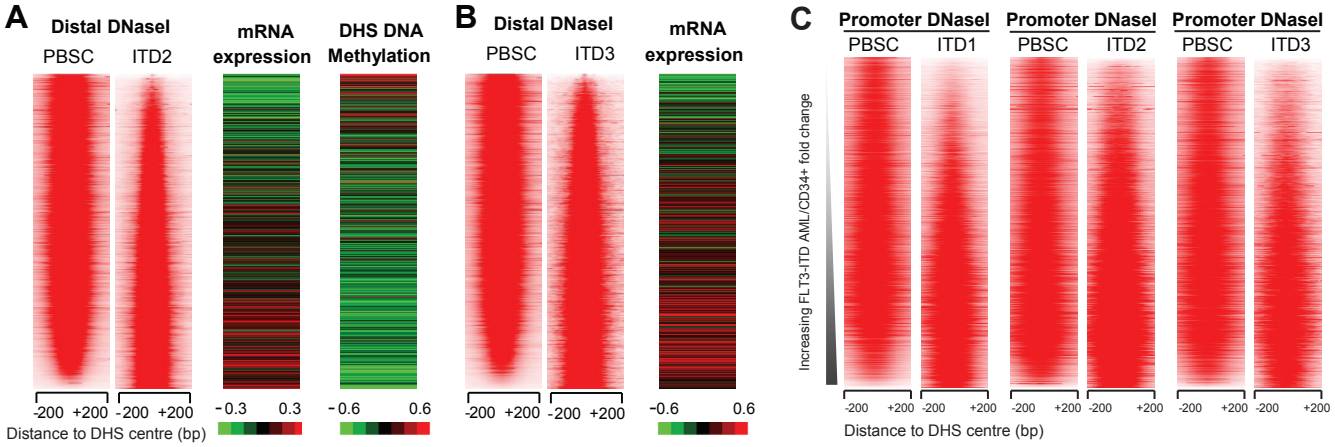
	Gene	mutations found	Region of Interest
Epigenetic	ASXL1	2	1/1, exon 13
	BCOR	3	8/14 exons, 4,5,8,9,11,12,14,15
	BCORL1	2	7/12 exons, 2,3,4,5,7,8,11
	CREBBP		25/31 exons 2,5,6,8,10-18,20-31
	DNMT3A	5	21/23 exons 3-23(NM_175629,NM_022552)
	EZH2		19/20 exons 2-20(NM_001203248)
	IDH1	3	exon 4 (R132)
	IDH2	2	exon 4 (R140)
	KDM6A (UTX)		19/29 exons 1,4,6,8,10,12,16-28
	MLL		13/36 exons 3,5,7,11,15,16,21,22,25,26,27,33,35
	SUZ12		4/16 exons 2,3,5,16
	TET2	7	9/9 exons 1-9
Chromatin Remodelling	ATRX		23/34 exons 1,4,7,9,11-15,17-23,25,26,28,29,31-33
	DAXX		7/8 exons 2,3,4,5,6,7,8
	H3F3A (H3.3)		exon 1 Codons K27M/G34R
Cell Cycle	CDKN2A (p16)		3/3 exons
Signalling	GNAS		amino acid R201/Q227
	HRAS		exons 2,3,4
	JAK2	2	exon 12, 14 (V617F)
	KRAS		exons 2,3,4
	MPL		exon 10 (W515K)
	NF1		58/58 exons 1-58
	NPM1	7	exons 11,12
	NRAS	1	exons 2,3,4
	PTEN		9/9 exons
	RASSF1		2/5 exons 4,5 (NM_170713)
	SH2B3 (LNK)		2/7 exons 5,6,(NM-005475)
	CTNNB1 (B-catenin)		codon Y654
	SOCS1	1	1/1 exon 1
	PTPN11	1	3,13,Y197(ex5)
Transcription Factor	CEBPa	4	1/1 exon 1
	ETV6		5/8 exons 2,3,5,6,7
	GATA1		3/6 exons 2,3,6
	GATA2	3	5/6 exons 2-6 (NM_032638)
	IKZF1		4/8 exons 2,4,6,8 (NM_006060)
	MAFB		1/1 exon 1
	PHF6	1	4/11 exons 2,7,8,9
	RUNX1	5	6/8 exons 3-8 (NM_001754)
	SPI1 (PU.1)		3/5 exons 3,4,5
	TP53	1	10/10 exons 1-10
	WT1	5	7,9,R430(ex8)
Receptor	KIT		exons 8,9,11,10,13, 17
	EGFR		exons 1-7, 18-22
	NOTCH-1		exon 26,27,34
	PDGFRA		exons 4,6,10
	FLT3	9	exon 14, 15, 20
Splisosome	PRPF40B		10/25 exons 1,2,4,7,12,13,14,17,20,22
	SF1		6/14 exons 2,4,6,7,8,9 (NM_001178030)
	SF3A1		8/16 exons 2-6,10,11,14
	SF3B1		12/25 exons 6,8,11,12,13,14,16,17,18,21,24
	SRSF2	3	amino acid P95
	U2AF1(35)		exons 2,6 (codons 34, 157)
	U2AF2(65)		3/12 exons 5,6,8
	ZRSR2		7/11 exons 1,3,4,7,8,9,11
Ubiquitination	CBL		2/16 exons 8,9

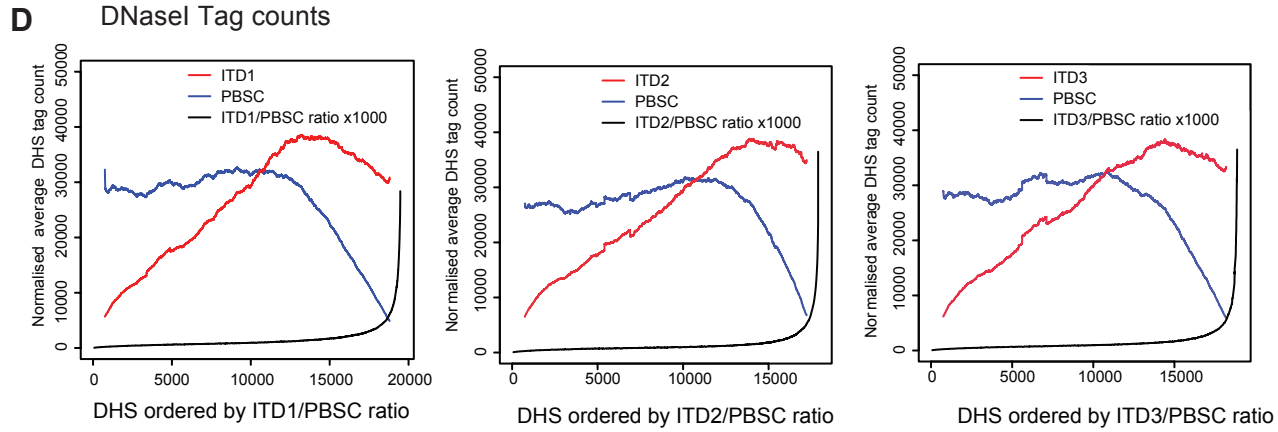
1212 amplicons

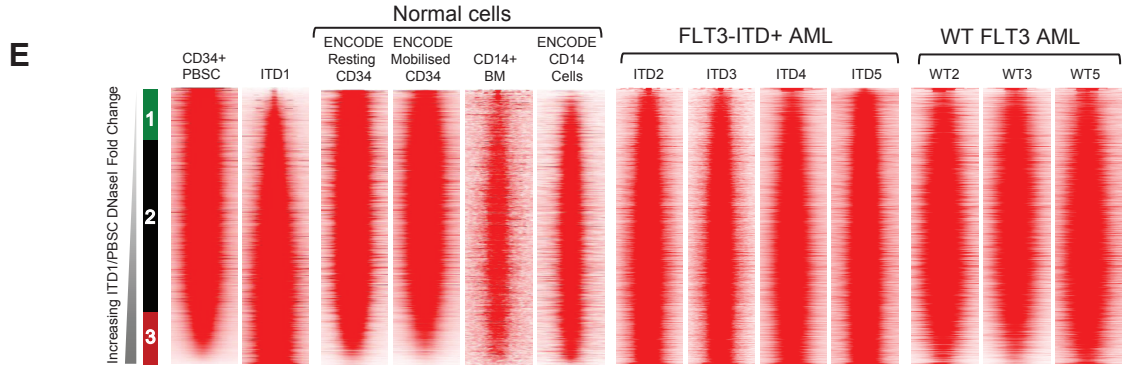
Supplemental Figure S1

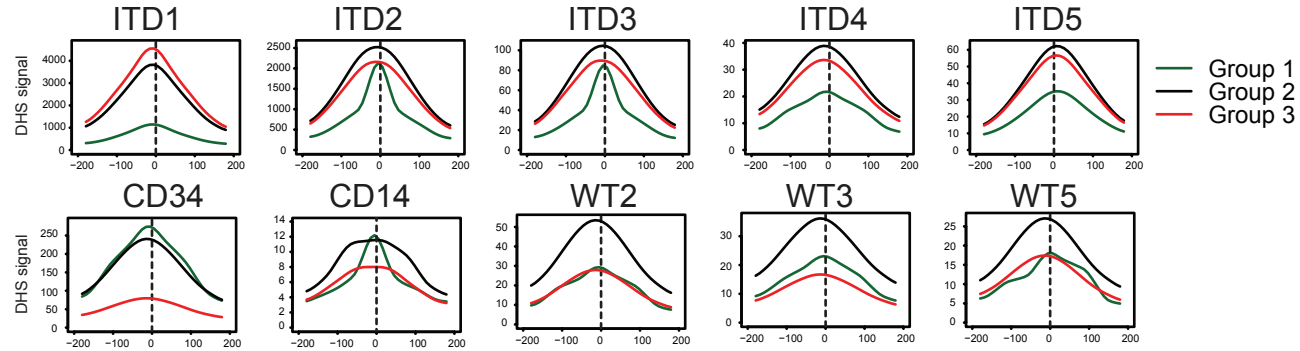


Supplemental Figure 2

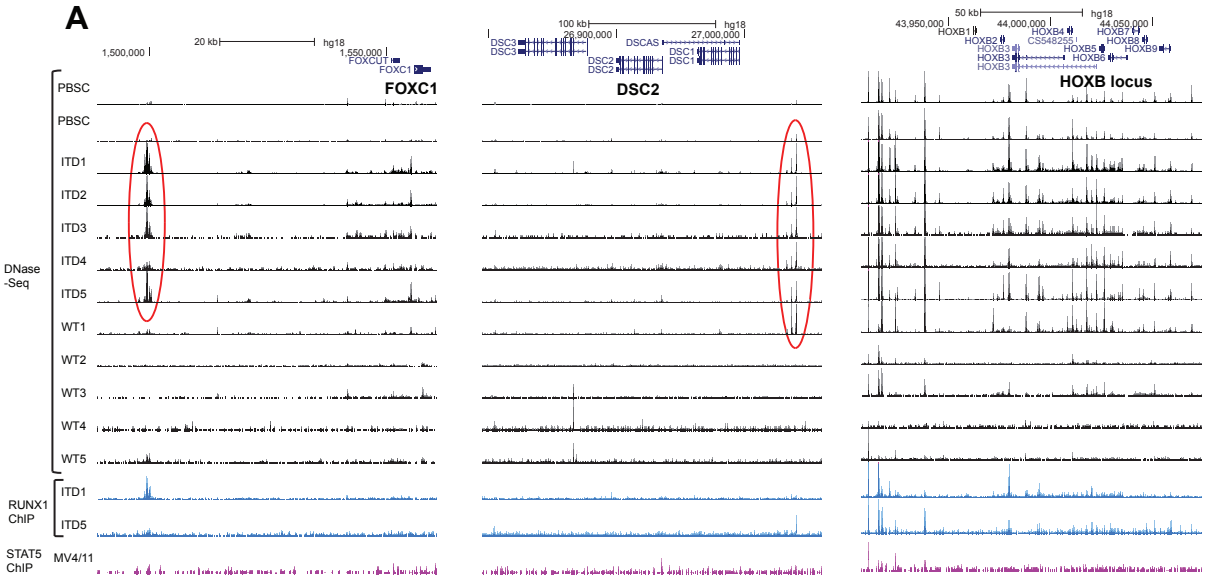








Supplemental Figure S3



FOXC1 DHS

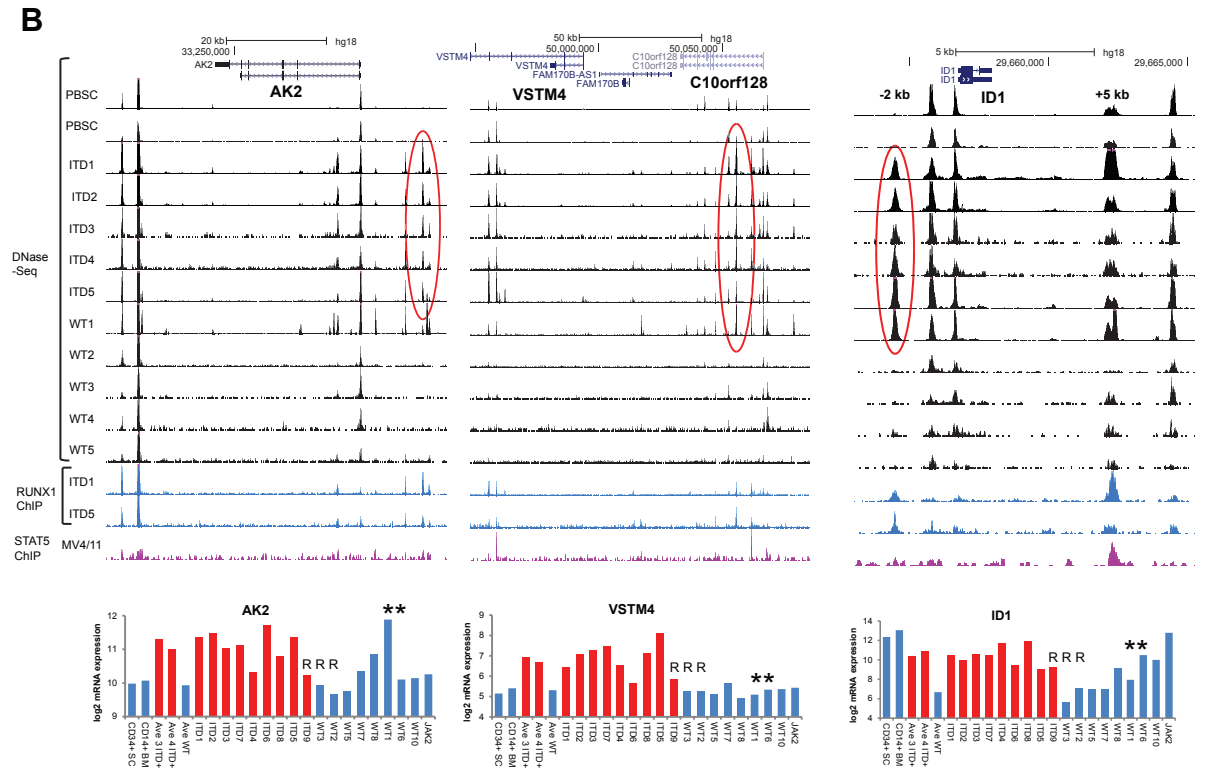
RUNX
TGTCTACTTACTGGCTT**TGTGGAGT**ATTAAAAAGCACTGCCCAGGGAACCTATTTG
RUNX
CTGG**TGTGGGTT**ATGCCAGGGGTGTCACGGTTTTAGAGTTGGCAGCACTAGCAGTA

GCTTTCGGCTTGGGGGAGTTTCTTGACTATATCAATTGCCTGTCATGGTGGTTTCC
RUNX
GACCCGGGCAGGTGGCATGGTTCTGGGGCTAGCAGC**TGTGGCCA**CCGTTTCCCTAC

DSC2 DHS

AP-1 AP-1 RUNX ETS
TCGG**TGAGTAA**TGCAG**TGACTAAAGCCCACA**CAA**AAAGGAAGC**AGTGCCC
RUNX
TTTTAATA**GAGCCACA**ATATAAAGAAAACACTCAAGCCAGTTCCAGCTAC

TGTGAATGATTTATTGGGTGAGCCCTGCCCGACTAGGCCCAGCAATTTCC
RUNX
CATTTGTTGCGTCCCAATCCTGCCC**CTCCCACA**TCTCCACCCTCCTGCTG



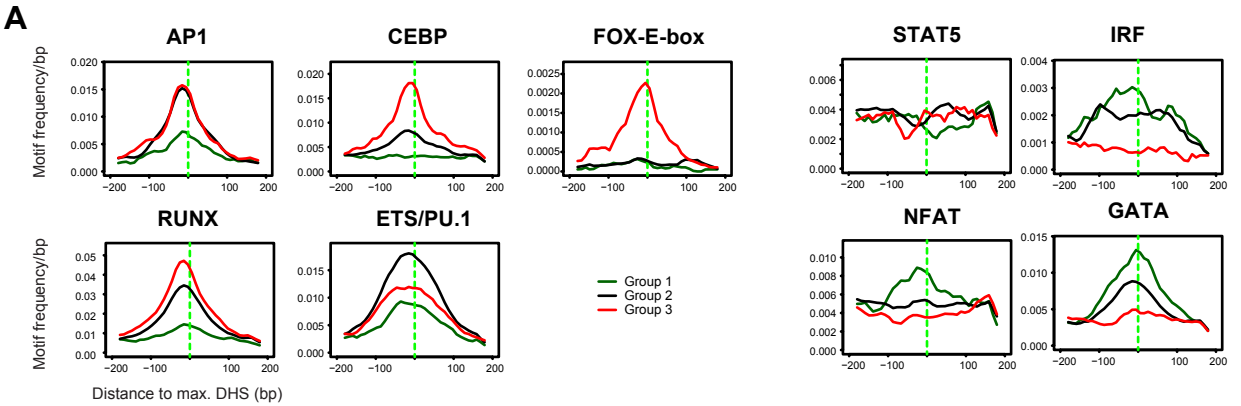
AK2 DHS

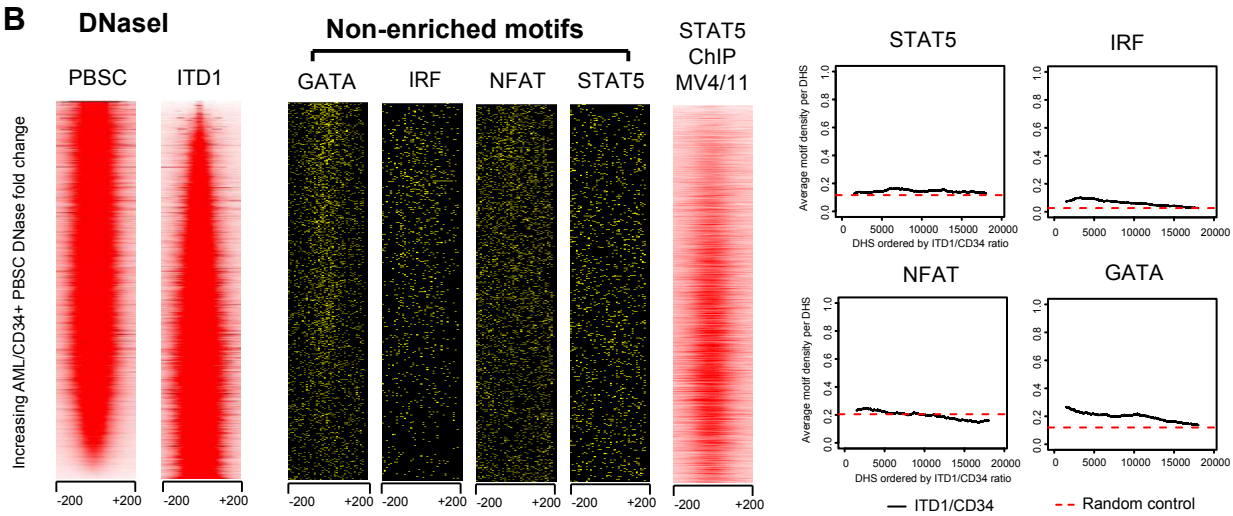
ETS
TGTCTTTTTTTGGGCTTCACAGCCCACTTGACCTAGGACC**GCTTCCCAA**T
ETS RUNX
TTTTGTAACAT**GCTTCCTGT**TCCTTTGTCAACC**AAACCACA**ACCATGAGA
RUNX
TGTGGTTTCTAAAGAACTGGCTAAGTCCAAGCCACCTGAACAAGCAAAG
CREB/ATF RUNX
CCTTTCTACTACTGTTT**TGAGGTCA**C**TGTGGCGA**AATTCCGATTTCCATA

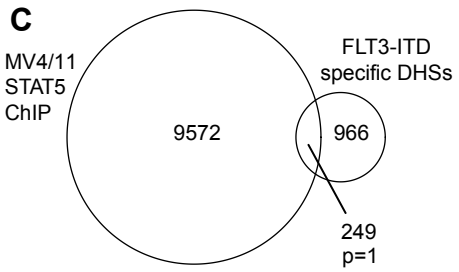
ID1 -2 kb DHS

RUNX AP-1
GCACAGATTTTTACCCCTGCTAATGTTG**TGTGG**CTTTGGG**TGACTAA**TCT
RUNX
GAGCTTCGGTTTCCCCACC**TGTGG**CATGTCCAGGACTCTGCAGTCAGACT
ETS C/EBP
GCCTGCTCTCA**GCAGGAAGC**G**GTTGCATTAC**AACACTGTACCCTCCTTAT
AP-1 RUNX
AATAACACAACTATGGTTA**TGACTTA**GGCGTCG**TGTGG**AGCCTCAGTTTC

Supplemental Figure 4

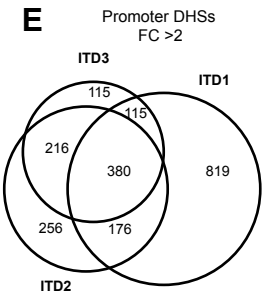






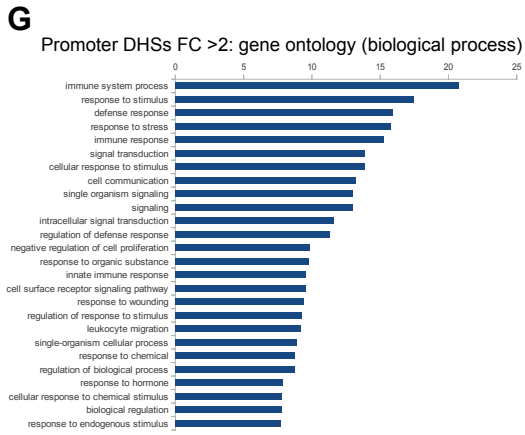
D

Motif	% targets	logo
ETS	43.7	
STAT	16.3	
RUNX	28.7	
C/EBP	19.8	
AP-1	20	

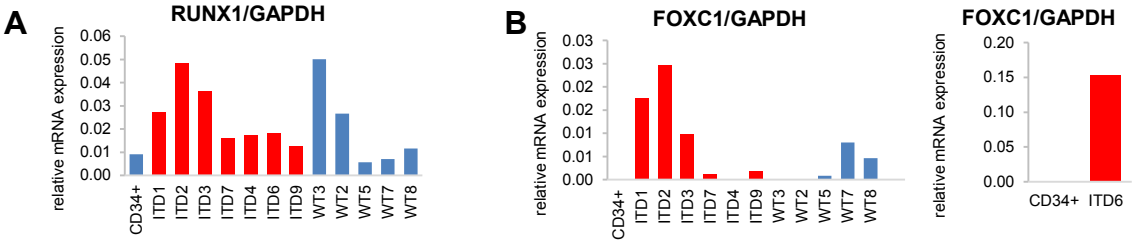


F

Motif	% targets	logo	logP
RUNX1	34.0		-37
ETS/PU.1	34.3		-30
AP-1	17.1		-22
NF-kB	8.6		-14
C/EBP	8.3		-14



Supplementary Figure 5



C **SCARA3 DHS**

TGCCGTCCCTCCAACTCGTCTCTAGCCAGGGCTGAGCTGACAACTTCTGT
TGCGCCAGCCCTGGATA**CACCTGTCAAAACA**GAGCCGCGGGAGAGGAGAG
CAGAGAGCTG**AAAAACA**TGGTC**TGACCACA**GTGAACTTTCCAGTGGGGCT
GGGCTAGGCTGGGCTGGCAAGCCTTGACC**TGTGGGTG**GTGTGCTGAGCAG

CTSG DHS

GTAGAAGATGGCA**TGACGTTA**TCCCCATGGCATCTCAGC**TGTGGTTT**TCC
TACTATAGCTTGCCAGCTTCT**CTGTAAAAACA**ACTCTGCCTGTCTCAGGT
ACTTCTAATGCTATTGTA**TCTTCCGGT**CATGAGGC**TGTGGTGG**GAGGACA
GTTTCCCTACAGCCTGGATGGTTGTAAAGC**TGTTTTTACT**CTGGACCCCA
CTCAAAACTGGTAGCCTTGCCGATTAC**TGAGTAA**ATGGGTAATAGGGGTA

MDGA1 DHS

CCTTCCTCCAGCCCAGCTGGTTGGATCAGGGTGGCATC**TGACTCA**AGGGC
AGCC**AACCCACA**GGTTGCCCAATGATGTATG**TGTGGCTA**GACCTGAAATG
GGGGCCTGTCCAATCAGCCCATTATCCTGAGCGGC**TGAGTCA**GTGAATGC
TGGCAGCTGCT**ACCGGAAGG**CCATGAAGGAGAGTTGAGCTGGAGAGGTCA
TGATGCCTGGAAAGCTGAAGTTA**TGAGGAAGC**AGACACCA**TGACTCA**GGA
GAGGAACTGCATGGACAAAAAGCAGCAGATGCTGGGGAGTGAC**TGAGTCA**

MED16 DHS

ACGACTTGGCTCTGGCTCTGCCGTGGCGGCACACAGG**CAGCTG**GTGCCAA
T**GAGCCACA**CTCACAAACCCAG**GCGGGGGCC**GTCAGGCCCTGGGCCGCAC
TGTGCAGACCTGAGCTATAGCACT**CAGGAAGT**GCTGTGTAAGCACCAGCG
GCTGCTGCGGGCACTATT**TGATTTCA**AATCAAATTTGTCAAGTCATTATT

GZMB DHS

TGAAGAAATGTCATAGAAAGATGAAAAATATTGAGGGAA**TGAGTTA**GGGC
TGTGAACGTACTGTGA**TTACTCA**A**TAACCACA**AACTTGCAATTAACTGCT
G**TAACCACA**AATCACCCTTCTAGGGTTCGCTGTGTAGAAAGAGCCGTTGA
TTTTTTATAGTCACTG**TAGCCACA**GCTGGAACCTCATAAAGAGTTCATGG

VSTM4/c10orf128 DHS

AGTGGCCTCTTTTCATGGGAAAGCC**CCAGGAAGT**GACGTGACCTCTTCAG
GCTCCATGCCTGAGCC**TGAGTCA**CTGGCTC**CGAGGAAGA**GTCTCACAGCA
GCATGGAGGGTGAACAGTCATTACAAACTGACCAAAGTATCTTAATGTGT
CACATG**TGACTCA**ACAATCAGGGCTCATC**TGCGGTTT**GGGAGGGTCTCTT

CCNA1 -14 kb DHS

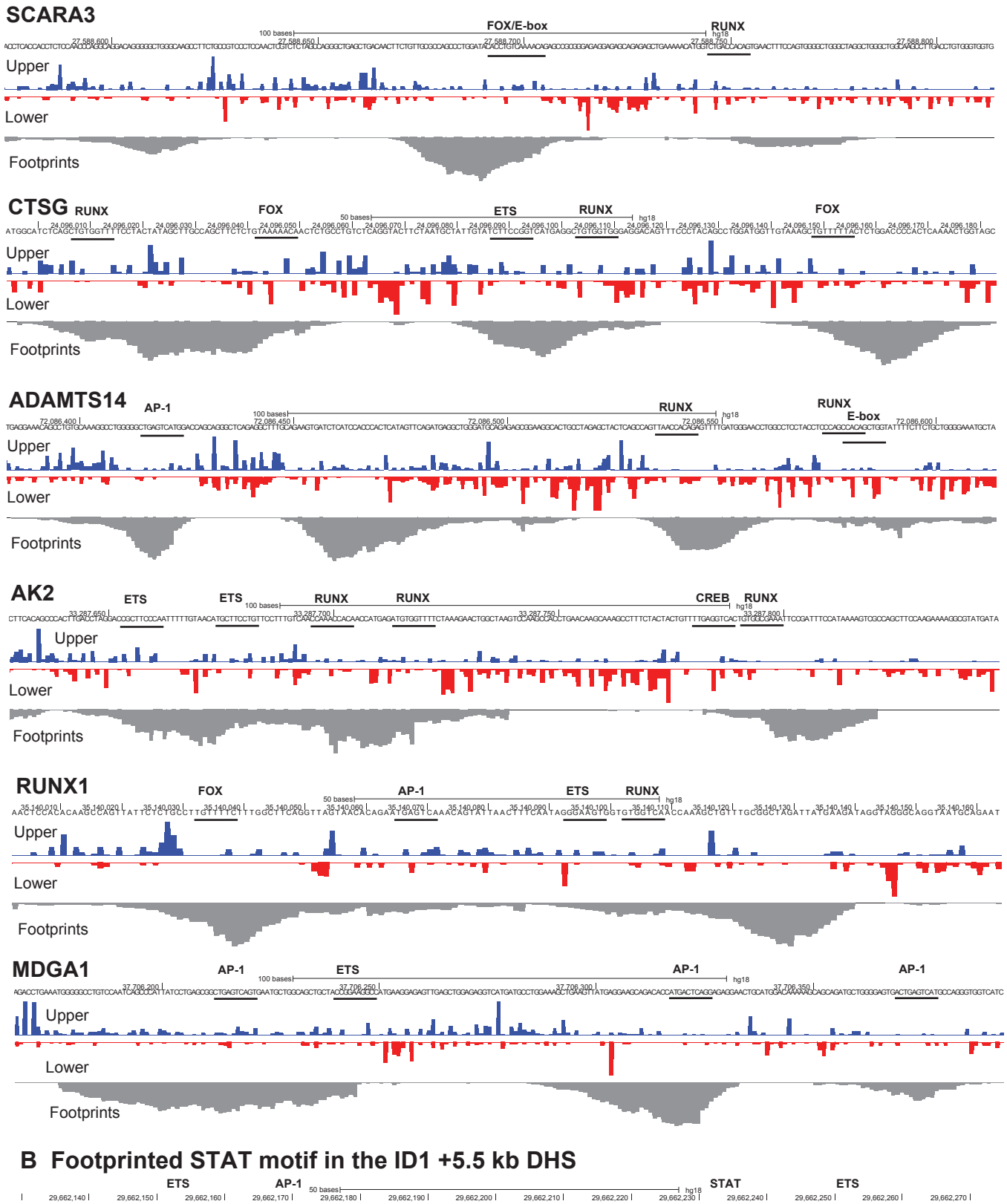
GTTTTTCTTAACTGCTCTAT**AGATAAC**AGAAGC**ATTGTGAAAT**ATTGAGT
TTTCAATTTGAGATATGATTTCAGTTCCCCTGTACCAGTAAAACTACTAA
CAC**CAGCTG**CTCTGAAGG**ATACCACA**AAGAGC**TGACTCA**CCAAAGAATGC

D **c8orf87/FAM92A1 DHS**

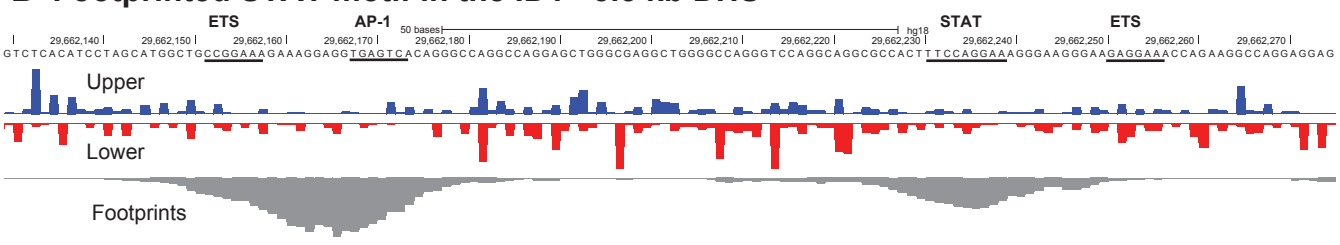
GTTTG**TGTTTACACAGGTGTGGCCA**TATTTAACAAAAAAATTTTTTATGC
ACTA**TGCAAACA**GCTGCCTGTC**TGTTTTTCT**TTGGCAGAGTTCTTTCCAAC
TGAAACAACTTATATCAAGCACATCGCTCTGTGATACTTA**TATAAACA**AC
ACAGAAA**AATTCCTG**TAC**AAGGAAGA**GGTTTCTCATAATGTGTCTGACAG

Supplemental Figure S6

A Footprints in ITD-specific DHSs



B Footprinted STAT motif in the ID1 +5.5 kb DHS



Supplemental table, figure, and data file legends

Table S1. Related to Figure 1.

List of genes and exons screened for mutations, including the total number of independent mutations found for each gene.

Figure S1. Related to main figure 1.

FLT3-ITD+ AML displays a characteristic mRNA expression profile.

(A) Log2 mRNA microarray values for 7 FLT3-ITD+ AML-specific target genes, plus *TBP* which serves as a control. Values are shown for the samples listed in **Table 1**, as well as for the average of 3 independent PBSC microarray analyses, the average of a central core of 3 ITD+ AML samples (ITD1, ITD2 and ITD3), the average of 4 additional ITD+ AML samples (ITD4, ITD6, ITD7 and ITD8), and the average of a core of 4 WT FLT3 AML samples (WT2, WT3, WT5 and WT7). Patients carrying RUNX1 mutations are labeled R, and patients with mutations in the signaling proteins NRAS or SOCS1 plus PTPN11 are labeled with an asterisk. **(B and C)** Expression of all annotated FOX genes **(B)** and 10 down-regulated MHC class II genes **(C)** in ITD+ AML (blue) and WT FLT3 AML (red) relative to PBSCs. Values shown represent the change in the average log2 values for the core group of 3 ITD+ AMLs (ITD1, ITD2 and ITD3) and for the core group of 4 WT FLT3 AMLs (WT2, WT3, WT5 and WT7) relative to the average log2 values for 3 PBSC data sets.

Figure S2. Related to main figure 3.

FLT3-ITD+ AML has a characteristic chromatin signature.

(A and B) Profiles of the DNase-Seq signals within each 400 bp window centered on each distal DHS peak for ITD2 compared to PBSCs **(A)** and ITD3 compared to PBSCs **(B)**, with peaks shown in the order of increasing DNase-Seq tag count signal for ITD2 or ITD3 relative to PBSCs. In each case these analyses include the union of all distal peaks present in either the AML sample or in PBSCs. Shown to the right of the DNase-Seq profiles are the relative mRNA expression values for genes with the nearest transcription start sites (TSS) in ITD2 or ITD3 relative to PBSCs, and the DNA methylation signals for the DHSs in ITD2 relative to PBSCs.

(C) Profiles of the DNase-Seq signals within each 400 bp window centered on each promoter-associated DHS peak for ITD1, ITD2 and ITD3 compared to PBSCs, with peaks shown in the order of increasing DNase-Seq tag count signal for the AML sample relative to PBSCs.

(D) Rolling averages of the DNase-Seq tag counts, plus the AML/PBSC ratios of these values for the DHSs depicted in **Figures 2E**, **S2A** and **S2B**.

(E) Side-by-side comparison of all the distal DHSs present in the ITD1/PBSC data set, ranked in order of increasing relative DNase-sensitivity as in **Figure 2E**, plus the DNase-Seq profiles across the same regions for the other samples depicted.
(F) Average DNase-Seq profiles across each of 3 groups of DHSs, grouped as indicated to the left, separated on the basis of a 2-fold difference between PBSCs and ITD1.

Figure S3. Related to main figure 4.

FLT3-ITD mutations are associated with a specific subset of DHSs.
(A and B) UCSC Genome browser views for DNase-Seq and RUNX1 ChIP-Seq data for ITD-specific DHSs (marked by red ovals) located near ITD-specific genes. Shown underneath these data are the profiles for a STAT5 ChIP assay of MV4-11 cells. ITD-specific DHSs are enclosed by red ovals. The log2 mRNA microarray values for some of the ITD-specific genes, displayed as in **Figures S1A** and **3F**, are shown in panel **B**. The mRNA values for *FOXC1*, *DSC2* and *HOXB4* are shown in **Figure S1A**. Note that the *C10orf128* locus is also a member of the group of 134 ITD-specific mRNAs, and that *ID1* is a known FLT3-ITD target gene (Tam et al., 2008) but is not ITD+ AML-specific. The DNA sequences are shown for 4 of the highlighted DHSs, with the regulatory motifs underlined in bold.

Figure S4. Related to main figure 5.

ITD-specific DHSs have a specific motif signature.
(A) Average densities of motifs across each of the 3 groups of distal DHSs subdivided as indicated at the left of **Figures 2E** and **4B** on the basis of relative DNase sensitivity.
(B) Alignment of STAT5, IRF, NFAT and GATA motifs with distal DHSs present in either ITD1 or PBSCs with the rolling averages of motif densities plotted on the right. Data is depicted in the same fashion as in **Figure 5B**, and with the MV4-11 STAT5 ChIP data presented alongside on the same coordinates.
(C) Venn diagram depicting the overlap between the MV4-11 STAT5 ChIP peaks and the 1216 ITD-specific DHSs.
(D) Motifs identified using HOMER to analyze the MV4-11 STAT5 Chip peaks.
(E) Venn diagram depicting the overlap between populations of promoter-associated DHSs that are 2-fold upregulated in AML samples ITD1, ITD2 and ITD3 compared to PBSCs (FC>2).
(F) Result of de novo motif search of 380 ITD-specific promoter-associated DHSs using HOMER.
(G) Gene ontology analysis of genes with ITD-specific promoter-associated DHSs.

Figure S5. Related to main figure 6.

FOXC1 and RUNX pathways are activated in ITD+ AML.

(A and B) Quantitative PCR/reverse transcriptase analysis of *RUNX1* **(A)** and *FOXC1* **(B)** mRNA levels in a subset of the AML samples listed in **Table 1**, and in CD34+ PBSCs. Values are expressed as mRNA levels relative to GAPDH mRNA.

(C and D) DNA sequences of ITD-specific DHSs associated within the *SCARA3, CTSG, MDGA1* and *c8orf87/FAM92A1, MED16, GZMB, c10orf128, and CCNA1* loci. Motifs representing the ITD-specific DHS signature are underlined in bold. The log2 mRNA microarray values for *FAM92A1* in panel D are displayed as in **Figure S1**.

Figure S6: Related to main figure 7.

ITD-specific DHS motifs are occupied in ITD+ AML.

(A and B) DNase I cleavage patterns spanning representative ITD1-specific footprints predicted by Wellington. Upper strand cut site frequencies are shown in blue and lower strand cut site frequencies are shown in red for ITD1. The relative probabilities for the presence of footprints predicted by Wellington are indicated by the grey histograms. ITD-specific DHS signature motifs that exist within predicted footprints are underlined in bold. Panel B displays a DHS that is 5 kb downstream of *ID1* that encompasses a footprinted STAT motif.

Supplemental data file 1. Related to Figure 1.

List of all the individual gene mutations identified from the mutation screen, including the DNA and amino acid changes, the proportion of sequences mutated (allele bias), the number of wild type and mutated sequences detected (allele depth), the method used for the mutation detection (Pindel or GATK), the nature of the 2 alleles (0=normal, 1= mutated), the lengths of insertions (INS) and deletions (DEL), and whether the mutation is previously recorded in the COSMIC database.

Supplemental data file 2. Related to Figure 1.

Alphabetical list of 134 upregulated genes and 77 downregulated genes in FLT3-ITD+ AML. Data is based on values obtained in **Figure 1** for the relative differences in the average log2 mRNA array values for the core group of ITD+ AML samples (ITD1, ITD2 and ITD3) relative to the other four populations defined in **Figure 1**. Also listed are the actual average log2 mRNA array values for each of the five populations.

Supplemental data file 3. Related to Figure 4.

Chromosomal coordinates of 1216 FLT-ITD+ AML-specific DHSs in the hg18 build of the human genome sequence. DHSs are listed alphabetically according the name of the nearest gene, and indicating the distance to the nearest transcription start site (TSS).

SUPPLEMENTAL EXPERIMENTAL PROCEDURES.

Patient sample, PBSC and CD14+ cell processing

Essentially all of the samples included in the collection described in Supplementary data File 1 are diagnostic samples from presentation cases before treatment. The two exceptions are a control sample with a JAK2 mutation from a patient who had progressed from MPD to AML, and sample ITD8 which came from a relapse of AML. All human tissue was obtained with the required ethical approval from the NHS National Research Ethics Committee. Most of the samples used in this study were surplus diagnostic samples obtained from the Haematological Malignancy Diagnostic Service (St James's Hospital, Leeds, UK), where cytogenetic abnormalities and sample immunophenotype were also determined at the time of disease diagnosis. Additional AML samples were obtained from the Centre for Clinical Haematology, Queen Elizabeth Hospital Birmingham, Birmingham, UK, and the West Midlands Regional Genetics Laboratory, Birmingham Women's NHS Foundation Trust, Birmingham, UK. AML samples were processed on the same day that they were received. Mobilized PBSCs were provided by NHS BT, Leeds, and NHS BT, Birmingham, in the UK.

For all samples used in this study, mononuclear cells were purified from bone marrow (BM), peripheral blood, or mobilized peripheral blood stem cells from patients or donors by differential centrifugation (20 mins, 2300RPM/881g, acceleration:4, no brake) using Lymphoprep (Axis-Shield UK, Cambridgeshire, UK). For most samples, undifferentiated blast cells were then isolated using antibody-coupled MACS Micro Beads (Miltenyi Biotec) and separation on magnetic columns (Miltenyi Biotec) according to the manufacturer's guidelines, using CD34 antibodies in most cases, and CD117 antibodies in 3 cases. For 6 samples with greater than 92 % blast cells (before the purification of mononuclear cells) the column purification was not performed. PBSCs were purified as for CD34+ AML cells, and CD14+ BM cells were obtained from orthopedic patients and were purified by the same protocol but using CD14 antibodies.

Cell lines

The cell lines MOLM14, MV4-11, THP1 and U937 were cultured in an incubator at 37^0C in GIBCO™ 1640 RPMI + Glutamax™ medium supplemented with 10% heat inactivated fetal calf serum (GIBCO), 100 U/ml Penicillin, 100 mg/ml Streptomycin.

DNAse-Seq library preparation

DNase-Seq libraries were prepared essentially as previously described (Ptasinska et al., 2014). To perform this global mapping of DHSs, DNase I digestions of permeabilized cells were performed as previously described (Bert et al., 2007). In this procedure, live cells were added directly to a solution of DNase I in dilute NP40, digested for 3 min at 22ºC, and the reactions then terminated by addition of SDS to 0.5%. This protocol maximizes the likelihood that transcription factors will remain bound during the digestion period, thereby increasing the probability of detecting DNase I footprints. DNase I (DPFF) was obtained from Worthington Biochemical Corporation and typically used in the range of 2-6 µg/ml using a final 1.5 x 10^7 cells/ml. The DNA digestion extent was comparable in all the generated samples as measured by RT-PCR (Ptasinska et al., 2012). The resulting cell lysates were then treated with 0.5 mg/ml Proteinase K overnight at 37 °C, then in 0.2 mg/ml RNAse A for 1 hour at 37ºC. DNA was isolated by phenol/chloroform extraction. Levels of DNase I digestion were assessed using quantitative real-time PCR, measuring the ratio of the presence of known DNase I hypersensitive regions compared to a more resistant inactive region. Sequences of the PCR primers used for this purpose were, for the active region, *TBP* promoter 5´-CTGGCGGAAGTGACATTATCAA and 5´- GCCAGCGGAAGCGAAGTTA; and for the inactive region, a region of chromosome 18: 5´- ACTCCCCTTTCATGCTTCTG and 5´-AGGTCCCAGGACATATCCATT. DNase-Seq samples were generated from a size selection of DNase I-digested DNA fragments comprised within a range of 100 to 250 bp (not including linkers) and subjected to library preparation as per manufacturer´s instruction (Illumina). Libraries were run on Illumina GAIIx, HiSeq 2000 and 2500 sequencers.

Chromatin immunoprecipitation

ChIP-Seq assays were performed essentially as previously described (Ptasinska et al., 2014). In this procedure, cells were resuspended in 10 ml of growing medium, and cross-linked with 1% formaldehyde (equivalent to ~0.33 M) for 10 min at RT. The cross-linking reaction was stopped by adding glycine in excess of the formaldehyde to a final concentration of 0.4 M, followed by two washes with ice–cold PBS. Cells were resuspended in 10 ml of ice-cold ChIP buffer A (10 mM HEPES pH 8.0, 10 mM EDTA, 0.5 mM EGTA, 0.25% Triton X-100, proteinase inhibitor cocktail (Roche UK, Burgess Hill, UK) and 0.1 mM PMSF), incubated for 10 min at 4°C with rotation, and centrifuged 5 min at 500 x g at 4 °C. The pellet was resuspended in 10 ml of ice–cold ChIP buffer B (10 mM HEPES pH 8.0, 200 mM NaCl, 1 mM EDTA, 0.5 mM

EGTA, 0.01% Triton X-100, protease inhibitor cocktail and 0.1 mM PMSF), incubated for 10 min at 4 °C with rotation and centrifuged for 5 min at 500 x g at 4 °C. Cells were resuspended in 600 µl of ice-cold ChIP lysis buffer (25 mM Tris-HCl pH 8.0, 150 mM NaCl, 2 mM EDTA, 1% Triton X-100, 0.25% SDS, protease inhibitor cocktail and 0.1 mM PMSF), incubated 10 min on ice and sonicated at 5 °C using a Bioruptor™ (Diagenode, Liege, Belgium) to generate fragments an average length of 400-500 bp (10 min with 30 s "ON" and "OFF" cycles, power setting high). The lysates were centrifuged for 5 min at 16,000 x g at 4 °C and the supernatants were diluted with two volumes of ice-cold ChIP dilution buffer (25 mM Tris-HCl pH 8.0, 150 mM NaCl, 2 mM EDTA, 1% Triton X-100, 7.5% glycerol, protease inhibitor cocktail and 0.1 mM PMSF). For each IP, 15 µl of Dynabeads® protein G were pre–incubated with 50 µg BSA and 2 µg antibody against RUNX1 (Abcam, ab23980) for 2 h at 4 °C with rotation. The blocked antibody-bound protein G mix was added to 20–25 µg chromatin in a total volume of 500 µl diluted ChIP lysis buffer and incubated for 2 h at 4°C with rotation. After magnetic separation the beads were washed once with 1 ml wash buffer 1 (20 mM Tris-HCl pH 8.0, 150 mM NaCl, 2 mM EDTA, 1% Triton X-100, 0.1% SDS), twice with 1 ml wash buffer 2 (20 mM Tris-HCl pH 8.0, 500 mM NaCl, 2 mM EDTA, 1% Triton X-100, 0.1% SDS), once with 1 ml LiCl buffer (10 mM Tris-HCl pH 8.0, 250 mM LiCl, 1 mM EDTA, 0.5% NP-40, 0.5% Na-deoxycholate) and twice with 1 ml TE/NaCl buffer (10 mM Tris-HCl pH 8.0, 50 mM NaCl, 1 mM EDTA). For each wash the beads were mixed with ice-cold washing buffers for 10 min at 4 °C. The immunoprecipitated DNA was eluted two times with 50 µl ChIP elution buffer (100 mM NaHCO3, 1% SDS) for 15 min at RT with shaking. At this step the input control (1% of the starting material) was included in the experimental procedure after first adjusting the final volume to 100 µl with ChIP elution buffer. The eluted DNA was incubated overnight at 65 °C in the presence of 50 µg proteinase K. The DNA was finally purified using Agencourt AMPure (Beckman Coulter) magnetic beads according to the manufacturer's instructions, eluted with 50 µl x TE.

ChIP assays of MV4-11 cells with control 2 µg of IgG (Millipore 12-370), 2 µg of Runx1 (Abcam, ab23980/GR2016781) 2 µg of STAT5 (Santa Cruz SC-835) and 5 µg of cFos (Santa Cruz SC-253) antibodies were performed by a modification of the above procedure, using 2 cross-linking agents. Cells were first washed with PBS and then cross-linked with Di(N-succinimidyl) glutarate (DSG) (Sigma, 8424-500MG-F). For each assay, 2 x10^7 cells were suspended in 30 mls PBS and incubated with 250 µl DSG (50mg/500µlDMSO) on a rotating wheel for 45 minutes at room temperature. After 45 minutes cells were washed 4 times with PBS and suspended in 10 mls of PBS. Cells were then cross-linked for a second time with 1% Formaldehyde for 10 minutes at room temperature. Cross-linking was terminated by adding 4 volumes of cold PBS+0.125 M of Glycine. Cells were washed with cold PBS, lysed and

sonicated (15 min with 30 s “ON” and “OFF” cycles, power setting high) and finally used for Chromatin Immunoprecipitation assays as above.

The quantitative PCR primers used for Fos and Runx1 ChIP assays had the following sequences.
MDGA1 DHS: GGGTGGCATCTGACTCAAG and ACTCTCCTTCATGGCCTTCC.
C10orf128 DHS1: CAAGGGCCTCTCTTGGGG and GACTGTTCACCCTCCATGCT.
CCNA1 (-14) DHS: ACTACTAACACCAGCTGCTCT and TGGTTAGGGTAAGGGGCATG.
GZMB DHS: GGAATGAGTTAGGGCTGTGAA and TCTTTATGAGGTTCCAGCTGTG.
CSF1R-FIRE DHS: GCCTGACGCCAACAATGTG and GGCAAAGGAGGGAAGTGAGAG.
IVL promoter: GCCGTGCTTTGGAGTTCTTA and CCTCTGCTGCTGCCACTT.

Mutation detection

Targeted exon sequencing of 55 cancer-associated genes was performed by the West Midlands Regional Genetics Laboratory using 1212 pairs of PCR primers, as summarized in **Supplemental Table 1** for amplification using a RainDance Technologies platform. The mutation sequence data summarized in Supplemental Data File 1 was analyzed using algorithms to detect either (i) nucleotide variants using the Genome Analysis Toolkit (GATK) (DePristo et al., 2011) or insertions and deletions using Pindel (Ye et al., 2009). Mutations were also screened against the COSMIC data base of previously observed mutations (http://cancer.sanger.ac.uk/cosmic/).

FLT3-ITD detection PCR

FLT3-ITD detection on genomic DNA from patient samples was adapted from a technique used to detect FLT3-ITD variants on cDNA (Kelly et al., 2002) using Taq polymerase (Life Technologies). Primers used to detect a wild-type amplicon size of 394 bp were:
FLT3det-FW: GGTGTTTGTCTCCTCTTCATTGT
FLT3det-RV AAAGCACCTGATCCTAGTACCTT
PCR products were separated on a 1.5% agarose gel.

EMSAs

Electrophoretic mobility shift assays were performed using an AP-1 probe as described previously (Cockerill et al., 1993).

Western blotting

Cells were lysed in RIPA buffer (Cell Signalling). After electrophoresis of protein extracts on polyacrylamide gels, proteins were transferred to nitrocellulose membranes (Thermo

scientific, Pierce) using a Mini-Trans blot cell (Bio-Rad). The membranes were blocked with 5% (w/v) milk powder in TBS-Tween 20 (0.1%) (TBST) at RT for 1 h and then incubated overnight at 4 °C with primary antibodies. Incubation with an anti-rabbit IgG, HRP-linked secondary antibody (TrueBlot, Rockland 18-8816-33) followed at RT for 1 h. Membranes were developed using the ECL Plus Western Blotting detection system (GE Healthcare) according to the manufacturer's protocol and signal was detected using autoradiography .
The antibodies used were supplied by Cell Signalling, with catalog numbers, as follows: FLT3 3462, Erk1/2 9102, Phospho Erk1/2 9101, Stat5 9363, Phospho Stat5 9359,
RSK2 9340, Phospho RSK2 3556, GAPDH 2118.

siRNA treatment in MV4-11 cells

$10x10^6$ MV4-11 cells were transfected with 300 nM of the control siRNA MMsiRNA (QIAGEN 1027286) and 300 nM of FLT3 siRNA with modification 5'-Cy5 (QIAGEN SI00059871) by using the Amaxa Cell Line Nucleofector kit L (VCA-1005). Cells were incubated in RPMI medium for 24 hours at 37°C. After this incubation time cells were washed with PBS and the efficiency of transfection was verified by FACs analysis. Then 1 $x10^6$ cells were used for RNA extraction (gene expression analysis), 2 $x10^6$ cells were used for protein extraction (Western blot) and the remaining of cells were used for Chromatin Immunoprecipitation assays.

The siRNA directed against FLT3 has the following sequences:
Sense strand: 5'-GGUUUAAAGCCUACCCACATT-3'
Antisense strand: 5'-UGUGGGUAGGCUUUAAACCTG -3'

MAPK inhibitors treatment in MV4-11 cells

$20x10^6$ MV4-11 cells were treated with 25 μM of 50 μM of PD98059 (Cell signalling No 9900) , SP600125 (Cell signalling No 8177) and 25 μM of SB202190 (Cell signalling No 8158) inhibitors, directed against MEK1/2, JNK and p38 respectively. Cells were incubated with DMSO (control) and inhibitors in RPMI medium for 5 hours at 37°C. After this incubation time cells were washed with 1X PBS and $1x10^6$ cells were used for RNA extraction (gene expression analysis)), $2x10^6$ cells were used for protein extraction (Western blot) and the remaining of cells were used for Chromatin Immunoprecipitation assays.

Gene expression microarray analysis

1-2 μg RNA was isolated from patient and donor samples via Trizol™ extraction. Sizes and quality of RNA preparations were checked using a RNA 6000 Pico Chip with a Bioanalyzer 2100 system (Agilent). 100 ng RNA was labelled with Cyanine 3-CTP according to the sample preparation protocol from Agilent: One-Color Microarray-Based Gene Expression Analysis (Low Input Quick Amp Labeling). Amplified labelled RNA samples were purified by using

Qiagen's RNeasy mini spin columns and cRNA quantified by using a Nanodrop spectrophotometer. Hybridisation samples were prepared for a 8-pack microarray using 600 ng cRNA each according to the hybridisation protocol from Agilent: One-Color Microarray-Based Gene Expression Analysis (Low Input Quick Amp Labeling), loaded onto as 8-pack SurePrint G3 Human GE 8x60K Microarray kit v1 design ID 028004 (Agilent) and hybridized at 65 °C overnight. After washing, microarrays were scanned on an Agilent G2565C Microarray Scanner using the Profile AgilentG3_GX_1Color for 8x60K microarrays (Dye channel: Green; Scan region: Scan Area (61 x 21.6 mm); Scan resolution (µm): 3; Tiff: 20 bit). Probe signals were extracted via the Agilent Feature Extraction software (version 10.7.1.1), protocol GE1_107_Sep09, using grid number 028004_D_F_20110325 (SurePrint G3 Human GE 8x60K). Arrays were normalized via quantile normalization in *R* using the *limma* package (Smyth et al., 2005). Transcript annotations were aggregated into unique gene names, whereby the mean of transcripts was computed for genes with more than one transcript. Genes with log2 intensities greater than 6.5 were considered expressed. Genes were considered enriched over CD34+ PBSC, WT AML and CD14+ PBSC using a two-fold change cutoff. The Pearson correlation matrix for all samples was computed in *R* and subsequently clustered via hierarchical clustering using *cluster 3.0* (de Hoon et al., 2004), with row, column Pearson correlation clustering and complete linkage. Heatmap images were generated via *Java TreeView* (Saldanha, 2004). For AML/ PBSC fold mRNA signal change heatmaps, expression fold change values were retrieved for each DHS, using values of the closest gene. Fold change values were sorted by increasing AML/ PBSC DHS fold change and plotted as heatmaps using *Java TreeView*.

Gene expression analysis by reverse transcriptase quantitative PCR analysis.

Gene expression patterns were confirmed by Real Time PCR analysis as previously described (Ptasinska et al., 2014). The PCR primers used were as follows:

GAPDH: CCTGGCCAAGGTCATCCAT and AGGGGCCATCCACAGTCTT.

FAM92A1: GGATGCTAGCCGAACAAGTC and ACCTCTAAAGCTTTGCCGTG.

CTSG: TCCTGGTGCGAGAAGACTTTG and GGTGTTTTCCCGTCTCTGGA.

NOV: CACGGCGGTAGAGGGAGATA and GGGTAAGGCCTCCCAGTGAA.

CCNA1: AGCACTTTTGGCCAGAAACC and GCTGAGGTCGATGGGGTATA.

PRTN3: CTCAATGTCACCGTGGTCAC and GGCCACCTGAGTCTCCGAA.

PT4A3: (PRL3): GCTTCCTCATCACCCACAAC and CGGCGTTTTGTCATAGGTCA.

IL2RA: CTGCCACTCGGAACACAAC and CTCGCTTGGTCCACTGGC.

FOXC1: CCCTCTCTTGCCTTCTTCCT and CGTCAGGTTTTGGGAACACT.

RUNX1: As used previously (Ptasinska et al., 2014).

DNA methylation array analysis

DNA methylation analysis was performed by Gen-Probe (UK) using an Illumina 450K methylation array kit. Methylation intensities were obtained via Illumina GenomeStudio to process array image files, probeset extraction and normalization. Per promoter intensities were retrieved via the *IMA R* package (https://www.rforge.net/IMA/index.html). Genome-wide coverage files (where probe data was available) were obtained using custom *R* and *Perl* scripts by generating *BED* files corresponding to intensities for all probes, followed by *WIG* format conversion.

Bioinformatic analyses

High-throughput sequencing alignment, peak detection and filtering

Sequencing reads were obtained as *FASTQ* files and uniquely aligned to the hg18 genome with *bowtie* using the following parameters: *--all --best --strata -v 2 -m 1*. For reads obtained from the HiSeq2500 sequencing platform, reads were retrieved in *BCL* format, converted to *FASTQ* via *bcl2fastq* and subsequently aligned using the same parameters. For samples run on more than one lane, the resulting *FASTQ* files were used as multiple inputs in *bowtie*, resulting in separate lanes being aligned together. Total and aligned read statistics are as follows:

Patient	total reads	aligned reads	peaks
PBSC	193,127,800	167,785,230	31,577
CD14+ BM	17,876,715	11,277,407	16,268
ITD1	496,164,295	342,121,088	25,156
ITD2	563,648,832	491,149,277	24,362
ITD3	31,191,334	22,512,783	27,899
ITD4	45,626,146	35,117,991	30,124
ITD5	232,690,946	172,049,755	26,127
WT1	173,684,899	155,121,178	25,091
WT2	46,523,235	34,921,707	23,592
WT3	27,668,387	22,368,270	26,510
WT4	33,683,955	28,775,417	20,707
WT5	38,565,964	33,615,251	25,868
WT8	56,390,352	40,690,809	37,406

To generate DHS coverage tracks, aligned reads were processed using a previously described bioinformatics pipeline (Koch et al., 2011). Briefly, fragment size was estimated using iterative extension, whereby the maximum overlap of reads corresponded to the mean

fragment size. Bases showing more than 5 read starts were considered as clonal artifacts and discarded. Reads were subsequently extended to the estimated size and depth coverage was derived by counting the number of reads in 10-bp windows genome-wide as fixed-length *WIG* files. Peak detection was subsequently carried out using *CoCAS* following *WIG* to *GFF* conversion. Peak detection parameters were set to the signal mean + 2 standard deviation ($p \leq 0.05$) for the both peak and extension thresholds. To account for further sequencing and/or repeat artifacts, peaks were cross-checked against an hg18-converted version of the ENCODE blacklist (ENCODE Project Consortium, 2012) via *bedtools* (Quinlan and Hall, 2010). Intersecting peaks were discarded as artifacts. Total numbers of peaks are indicated above.

Generation of peak summit unions

For AML vs CD34+ PBSC DHS comparisons, peak summit unions were performed as previously described (Kreher et al., 2014). Essentially, peaks were annotated to the nearest isoform for which a distance criterion of 5 kb upstream or downstream of the TSS was used to treat peaks as distal or proximal. Distal and proximal peaks were subsequently treated separately. Unions were computed as the concatenation of AML and CD34+ PBSC datasets, with peak summits closer than 400 bp being treated as one same peak. In this case, the average peak summit coordinate was used as the merge of both. DHS coverages were retrieved [-200bp; +200bp] around the union summits using custom Perl scripts. Distributions of coverages were plotted and tested positive for normality via Shapiro-Wilkes normality tests in *R*. Coverages were consequently normalized via log2 transformation, centre-scaling to the AML dataset and ranked by log2 AML/CD34+ PBSC fold change to account for read depth heterogeneity. To minimize redundancy, further overlaps of merged and other union summits (left as they were during the merging process) within 400 bp resulted in the first one in genomic coordinate order being retained. To avoid error values, values of 0 tags were replaced by 1 prior to log2 transformation. Heatmap images were generated via *Java TreeView*. Average signal profiles by increasing DHS fold change were computed using a rolling average with a window size of 1500.

DHS clustering

For total DHS clustering, the union of all distal, primary AML (FLT3-ITD, WT) and CD34+ PBSC DHS summits was computed. Tag coverages were recovered [-200 bp to +200bp] around the union of all summits, log2 transformed, center-scaled and expressed as log2 AML/CD34 fold change, with log2 signal intensities of 0 changed to 1. All values were collated into a single table. Hierarchical clustering was performed via *cluster 3.0*, using Pearson correlation clustering for row and column clustering, as well as single linkage due to the table size. For DHS correlation clustering, tags were recovered, log2 transformed and normalized

following the exact same procedure. A correlation matrix for all samples was computed in *R* and subsequently clustered hierarchically via *cluster 3.0*, using Pearson correlation clustering for rows and arrows, as well as complete linkage. Heatmap images were generated via *Java TreeView*.

Intersection of AML and CD34+ PBSC enriched DHSs, DHS group definition and average profiles

To define groups of DHSs with regards to their enrichment status vs CD34+ PBSC, we computed intersections of three representative FLT3-ITD AML/CD34+ PBSC unions (ITD1, ITD2 and ITD3) and selected DHSs that had a two-fold higher AML/PBSC signal ratio. Since for each union, the AML/CD34+ PBSC DHS fold change followed a normal distribution, we used the same two-fold threshold as the cutoff for AML/PBSC under-enrichment. All DHSs in between -1 and 1 log2 fold change were considered as shared between AMLs and CD34+ PBSCs. Per group average profiles for DNase I, ChIP-Seq, motif frequencies, CpG methylation were generated via a previously described *R* pipeline (Fenouil et al., 2012). Microarray gene expression and CpG methylation fold change boxplots were generated using *R*.

Significance of overlaps

2-way intersection p-values were computed in *R* using hypergeometric tests in the *ChIPpeakAnno* package (Zhu et al., 2010). 3-way intersection p-values were by deriving the distribution of probabilities for intersections of random samples with sizes corresponding to actual samples, via bootstrapping (10,000 iterations). The actual intersection p-value was retrieved by obtaining the p-value corresponding to the overlap in the simulated distribution of probabilities.

Motif discovery and heatmap generation

Motif discovery was performed via the *findMotifsGenome* function of *Homer* (Heinz et al., 2010) with a window of -200 bp to +200 bp around the summit. Solely de novo motif enrichment was considered to minimize motif redundancy. For composite motifs, motif length optimization was performed for full-length identification. Motif heatmaps were derived via the *annotatePeaks* function of *Homer* using all enriched motifs simultaneously, followed by custom *Perl* scripts to separate motif outputs. Individual motif presence tables were generated for 200 bp upstream and downstream of each summit for every consequtive10 bp windows and sorted by AML/ PBSC DHS fold change. Random occurrences of motifs were computed by performing motif discovery in similarly sized random sets of coordinates.

Motif co-occurrence clustering analysis

Motif co-occurrence clustering was essentially performed as previously described (Ptasinska et al., 2014). Briefly, digital genomic footprinting outputs from AML and CD34+ PBSC DHS samples were intersected and defined as AML-specific, shared or CD34+ PBSC –specific via the *bedtools intersect* function. Specific populations were scanned using motif matrices from *Homer*, via the *annotatePeaks* function of *Homer*. Motif mapping outputs were converted to the *BED* format. To obtain motif co-occurrence, motifs containing footprints were all intersected using the intersection_matrix function of the pybedtools package (Dale et al., 2011). We restricted motif selections to those corresponding to TFs that were actually expressed in any of our analyzed cell types. To assess significance with regards to random occurrence in footprints, we sought to estimate over-representation of occurrence as compared to background co-occurrence all footprints. We thus chose the union of AML plus CD34+ PBSC footprints. Background co-occurrence was estimated using bootstrapping (1000 repetitions) of motif mapping and co-association counts (within 50 bp) in randomly selected footprints within the background, using equally sized populations as the original number of specific footprints. Motif mapping was carried out via the *annotatePeaks* function of *Homer*, non-redundant, non-composite matrices

Corresponding ChIP-Seq heatmaps

ChIP-Seq tag counts were recovered for 200 bp upstream and downstream of merged DHS union summits via custom Perl scripts. Rows were reordered accordingly for each DHS union. Heatmap images were generated via *Java TreeView*. Average profiles by increasing AML/ PBSC DHS signal ratio were computed in *R* using a moving average with a window size of 1500.

Digital genomic footprinting

High-depth DHS datasets (ITD1, ITD2, and CD34+ PBSC) were aligned as described and subsequently converted to the sorted BAM format, whereby an index was generated via *samtools* (Li et al., 2009). DHSs were specifically identified using the *findPeaks* function of Homer. Digital genomic footprinting was performed using the *Wellington_footprints* function of the *Wellington* algorithm (Piper et al., 2013) on AML and CD34+ PBSC DHSs. DHS cut coverages, strand imbalance heatmaps and average profiles were generated using the *dnase_wig_tracks*, *dnase_to_javatreeview* and *dnase_average_profile* functions of Wellington. AML-specific footprints were identified by computing differences in footprinting occupancy scores at given genomic coordinates for both test and comparator datasets, then merging reads from each and estimating the footprinting score of merged reads against the randomized comparator dataset as a percentile. Heatmaps images were obtained via *Java TreeView*. For heatmaps showing the presence of footprinted motifs, sorted by AML/PBSC

footprint probability fold change, motif discovery results were converted to *BED* files and subsequently intersected with digital footprinting results using the *bedtools intersect* function. Footprinted motifs were mapped back to the AML/ PBSC DHS unions by calculating the distance of each footprinted motif to the merged DHS summit *BED* entry, whereby infinity was used if a DHS did not intersect with any footprinted motif. Resulting outputs were converted as distances to summit, and subsequently expressed as tables 200 bp upstream and downstream of each summit using custom Perl scripts, whereby motif frequencies were computed every 10 bp for all regions, ordered according to fold changes. Heatmaps were generated via *Java Treeview*. Motif densities were computed relative to each summit, where distances used were that between the start of each motif (regardless of the strand) and the summit.

Accession numbers

The DNA sequence files associated with this study have been deposited as a superseries in the Gene Expression Omnibus data base (accession number GSE64874). This superseries encompasses individually accessible DNase-Seq datasets GSE64864), ChIP-Seq datasets (GSE64862) and microarray expression datasets (GSE64873)

Public datasets

The previously published Runx1 ChIP-Seq dataset in CD34+ PBSC (Ptasinska et al., 2014) was retrieved from the Gene Expression Omnibus (GEO), accession GSM1466000. Following SRA to FASTQ conversion via the SRA toolkit (version 2.4.7), this sample was processed as other high throughput sequencing samples. ENCODE DNase I chromatin accessibility datasets for CD34+ stem cells, mobilized CD34+ PBSCs and CD14+ cells were retrieved from GEO accessions GSM595919, GSM530652 and GSM701541, respectively (Bernstein et al., 2010; Neph et al., 2012). These samples were processed similarly as above.

Expression profiles in larger patient cohort datasets

Gene expression results from large patient cohort datasets (Cancer Genome Atlas Research Network, 2013; Verhaak et al., 2009) were downloaded from the Leukemia Gene Atlas (Hebestreit et al., 2012) as text files. For data from Verhaak et al., patients were split according to their FLT3-ITD mutational status; for data from the Cancer Genome Atlas Research Network, since only the presence of mutations in the FLT3 gene was screened, patients were split according to their FLT3 mutational status. Probeset expression data was aggregated into per gene average expression data in the case of genes spanning more than one probe. Expression values for 134 and 77 FLT3-ITD up- and down- regulated genes identified in this study were subsequently retrieved using the *merge* function in *R*. Boxplots for individual genes and gene averages were plotted using *R*.

c-Fos ChIP-Seq analysis with and without knock-down in MV4-11 cells

c-Fos ChIP-Seq with mismatch and FLT3 siRNA datasets were processed as described above. To identify whether c-Fos signal changed genome-wide following knock-down, peak summits using the control mismatch siRNA c-Fos ChIP were isolated and average binding profiles were retrieved ±500bp around the summit via the *annotatePeaks* function of *Homer*, using -hist 10 -wig as parameters. Values were smoothed using a moving average of 5 bins (50bp). Average profiles were plotted using *LibreOffice Calc*.

SUPPLEMENTAL REFERENCES

Bernstein, B.E., Stamatoyannopoulos, J.A., Costello, J.F., Ren, B., Milosavljevic, A., Meissner, A., Kellis, M., Marra, M.A., Beaudet, A.L., Ecker, J.R*., et al.* (2010). The NIH Roadmap Epigenomics Mapping Consortium. Nat Biotechnol *28*, 1045-1048.

Bert, A.G., Johnson, B.V., Baxter, E.W., and Cockerill, P.N. (2007). A modular enhancer is differentially regulated by GATA and NFAT elements that direct different tissue-specific patterns of nucleosome positioning and inducible chromatin remodeling. Mol Cell Biol *27*, 2870-2885.

Cancer Genome Atlas Research Network (2013). Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. N Engl J Med *368*, 2059-2074.

Cockerill, P.N., Shannon, M.F., Bert, A.G., Ryan, G.R., and Vadas, M.A. (1993). The granulocyte-macrophage colony-stimulating factor/interleukin 3 locus is regulated by an inducible cyclosporin A-sensitive enhancer. Proc Natl Acad Sci U S A *90*, 2466-2470.

Dale, R.K., Pedersen, B.S., and Quinlan, A.R. (2011). Pybedtools: a flexible Python library for manipulating genomic datasets and annotations. Bioinformatics *27*, 3423-3424.

de Hoon, M.J., Imoto, S., Nolan, J., and Miyano, S. (2004). Open source clustering software. Bioinformatics *20*, 1453-1454.

DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M*., et al.* (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet *43*, 491-498.

ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57-74.

Fenouil, R., Cauchy, P., Koch, F., Descostes, N., Cabeza, J.Z., Innocenti, C., Ferrier, P., Spicuglia, S., Gut, M., Gut, I*., et al.* (2012). CpG islands and GC content dictate nucleosome depletion in a transcription-independent manner at mammalian promoters. Genome Res *22*, 2399-2408.

Hebestreit, K., Grottrup, S., Emden, D., Veerkamp, J., Ruckert, C., Klein, H.U., Muller-Tidow, C., and Dugas, M. (2012). Leukemia gene atlas--a public platform for integrative exploration of genome-wide molecular data. PLoS One *7*, e39148.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Molecular cell *38*, 576-589.

Kelly, L.M., Liu, Q., Kutok, J.L., Williams, I.R., Boulton, C.L., and Gilliland, D.G. (2002). FLT3 internal tandem duplication mutations associated with human acute myeloid leukemias induce myeloproliferative disease in a murine bone marrow transplant model. Blood *99*, 310-318.

Koch, F., Fenouil, R., Gut, M., Cauchy, P., Albert, T.K., Zacarias-Cabeza, J., Spicuglia, S., de la Chapelle, A.L., Heidemann, M., Hintermair, C*., et al.* (2011). Transcription initiation platforms and GTF recruitment at tissue-specific enhancers and promoters. Nat Struct Mol Biol *18*, 956-963.

Kreher, S., Bouhlel, M.A., Cauchy, P., Lamprecht, B., Li, S., Grau, M., Hummel, F., Kochert, K., Anagnostopoulos, I., Johrens, K*., et al.* (2014). Mapping of transcription factor motifs in active

chromatin identifies IRF5 as key regulator in classical Hodgkin lymphoma. Proc Natl Acad Sci U S A *111*, E4513-4522.
Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078-2079.
Neph, S., Vierstra, J., Stergachis, A.B., Reynolds, A.P., Haugen, E., Vernot, B., Thurman, R.E., John, S., Sandstrom, R., Johnson, A.K.*, et al.* (2012). An expansive human regulatory lexicon encoded in transcription factor footprints. Nature *489*, 83-90.
Piper, J., Elze, M.C., Cauchy, P., Cockerill, P.N., Bonifer, C., and Ott, S. (2013). Wellington: a novel method for the accurate identification of digital genomic footprints from DNase-seq data. Nucleic Acids Res *41*, e201.
Ptasinska, A., Assi, S.A., Mannari, D., James, S.R., Williamson, D., Dunne, J., Hoogenkamp, M., Wu, M., Care, M., McNeill, H.*, et al.* (2012). 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.
Ptasinska, A., Assi, S.A., Martinez-Soria, N., Imperato, M.R., Piper, J., Cauchy, P., Pickin, A., James, S.R., Hoogenkamp, M., Williamson, D.*, et al.* (2014). Identification of a Dynamic Core Transcriptional Network in t(8;21) AML that Regulates Differentiation Block and Self-Renewal. Cell Reports *8*, 1974-1988.
Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841-842.
Saldanha, A.J. (2004). Java Treeview--extensible visualization of microarray data. Bioinformatics *20*, 3246-3248.
Smyth, G.K., Michaud, J., and Scott, H.S. (2005). Use of within-array replicate spots for assessing differential expression in microarray experiments. Bioinformatics *21*, 2067-2075.
Tam, W.F., Gu, T.L., Chen, J., Lee, B.H., Bullinger, L., Frohling, S., Wang, A., Monti, S., Golub, T.R., and Gilliland, D.G. (2008). Id1 is a common downstream target of oncogenic tyrosine kinases in leukemic cells. Blood *112*, 1981-1992.
Verhaak, R.G., Wouters, B.J., Erpelinck, C.A., Abbas, S., Beverloo, H.B., Lugthart, S., Lowenberg, B., Delwel, R., and Valk, P.J. (2009). Prediction of molecular subtypes in acute myeloid leukemia based on gene expression profiling. Haematologica *94*, 131-134.
Ye, K., Schulz, M.H., Long, Q., Apweiler, R., and Ning, Z. (2009). Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics *25*, 2865-2871.
Zhu, L.J., Gazin, C., Lawson, N.D., Pages, H., Lin, S.M., Lapointe, D.S., and Green, M.R. (2010). ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data. BMC Bioinformatics *11*, 237.