Supplementary Information

Extended Experimental Procedures 1
Chromatin immunoprecipitation (ChIP) 1
rt-qPCR 2
Strand-specific RNA-seq 2
Illumina high-throughput sequencing 2
4C-Seq 3
Bioinformatic analyses 3
Alignment 3
Peak calling 4
MLL binding site classification 4
Tag counting 4
Peak distribution analysis 4
Motif analysis 4
Pathway Enrichment analysis 5
Expression analysis 5
4C-seq analysis 5
Antibodies 5
Primers 5
ChIP-qPCR 5
cDNA 6
4C-seq 6
Supplementary Figure Legends 6
Supplementary References 7

Extended Experimental Procedures

Chromatin immunoprecipitation (ChIP)

Cells were crosslinked with 1% formaldehyde for 20 min at room temperature, quenched with 0.125 M glycine and washed with three buffers: (i) PBS, (ii) 0.25% Triton X 100, 10 mM EDTA, 0.5 mM EGTA, 20 mM HEPES pH 7.6 and (iii) 0.15 M NaCl, 10mM EDTA, 0.5 mM EGTA, 20mM HEPES pH 7.6. Cells were then suspended in ChIP incubation buffer (0.15% SDS, 1% Triton X 100, 150 mM NaCl, 10 mM EDTA, 0.5 mM EGTA, 20mM HEPES pH 7.6) and sonicated using a Bioruptor sonicator (Diagenode) for 20 min at high power, 30 s ON, 30 s OFF. Sonicated chromatin was centrifuged at maximum speed for 10 min and then incubated overnight at 4°C in incubation buffer supplemented with 0.1% BSA with protein A Dynabeads (Thermo Fisher Scientific) and 2 µg of antibody. Beads were washed sequentially with four different wash buffers at 4˚C: two times with a solution of 0.1% SDS, 0.1%

DOC, 1% Triton, 150 mM NaCl, TEE (10mM Tris pH 8, 0.1mM EDTA and 0.5mM EGTA), one time with a similar buffer but now with 500 mM NaCl, one time with a solution of 0.25 M LiCl, 0.5% DOC, 0.5% NP-40, TEE and two times with TEE. Precipitated chromatin was eluted from the beads with 400 µl of elution buffer (1% SDS, 0.1 M $NaHCO_3$) at room temperature for 20 min. Protein-DNA crosslinks were reversed at 65°C for 4 h in the presence of 200 mM NaCl, after which DNA was isolated by Qiagen column. Antibodies and primers for qPCR can be found below. For qPCR, relative occupancy was calculated as fold over background, for which the promoter of the Myoglobin gene was used.

rt-qPCR

cDNA was synthesized using iScript (BioRad) and enrichment was calculated as fold over the 1st exon of GAPDH.

Strand-specific RNA-seq

Total RNA was extracted by TRIzol (Invitrogen), subjected to on-column DNase treatment (Qiagen) and the concentration was measured with a Qubit fluorometer (Invitrogen). 250 ng of total RNA was used with the Ribo-Zero rRNA Removal Kit (Illumina) to remove ribosomal RNAs according to manufacturer instructions. 16 µl of purified RNA was fragmented by addition of 4 µl 5x fragmentation buffer (200 mMTris acetate pH 8.2, 500 mM potassium acetate and 150 mM magnesium acetate) and incubated at 94°C for exactly 90 s. After ethanol precipitation, fragmented RNA was mixed with 5 µg random hexamers, followed by incubation at 70 °C for 10 min and chilling on ice. We synthesized first-strand cDNA with this RNA primer mix by adding 4 µl 5× first-strand buffer, 2 µl 100 mM DTT, 1 µl 10 mM dNTPs, 132 ng of actinomycin D, 200 U SuperScript III, followed by 2 h incubation at 48 °C. First strand cDNA was purified by Qiagen mini elute column to remove dNTPs and eluted in 34 µl elution buffer. Second-strand cDNA was synthesized by adding 91.8 µl, 5 µg random hexamers, 4 µl of 5× first-strand buffer, 2 µl of 100 mM DTT, 4 µl of 10 mM dNTPs with dTTP replaced by dUTP, 30 µl of 5× second-strand buffer, 40 U of *Escherichia coli* DNA polymerase, 10 U of *E. coli* DNA ligase and 2 U of *E. coli* RNase H, and incubated at 16 °C for 2 h followed by incubation with 10 U T4 polymerase at 16 °C for 10 min. Double stranded cDNA was purified by Qiagen mini elute column and used for library preparation as described in the KAPA Hyper Prep protocol. We incubated 1 U USER (NEB) with adaptor-ligated cDNA at 37 °C for 15 min followed by 5 min at 95 °C before PCR to ensure strand specificity.

Illumina high-throughput sequencing

ChIP-seq libraries were prepared from precipitated DNA of 5 million cells (5-8 pooled biological replicas) for MLL, AF9, AF4, and RUNX1, or from 1 million cells for the histone tail modifications. End repair was performed using Klenow and T4 PNK. A 3' protruding A base was generated using Taq polymerase and adaptors were ligated. The DNA was loaded on E-gel and fragments corresponding

to ~300 bp (ChIP fragment + adaptors) were excised. The DNA was isolated, amplified by PCR and used for cluster generation and sequencing on the Genome Analyzer (Illumina) or HiSeq 2000 (Illumina). For RNA-seq 250 ng of RNA was used for ribosomal RNA depletion with RiboZero (Illumina) and subsequent strand specific library preparation. The 42-50 bp tags were mapped to the reference human genome hg19 using the Burrows-Wheeler Alignment Tool (BWA) for ChIP-seq or TopHat2 (Bowtie2) for RNA-seq. For each base pair in the genome, the number of overlapping sequence reads was determined, averaged over a 10 bp window and visualized in the UCSC genome browser (http://genome.ucsc.edu).

4C-Seq

4C assays were performed as described previously[1] with minor modifications. Briefly, 10^7 cells were cross-linked for 10 minutes with 2% paraformaldehyde, quenched with glycine and lysed in 50 ml lysis buffer (50mM Tris pH 7.5, 150mM NaCl, 5mM EDTA, 0,5% NP-40, 1% TX-100, 1X protease inhibitors) for 30 min. Nuclei were then digested by DpnII enzyme followed by inactivation of restriction enzyme by incubating at 65˚ C for 20 min. The digested chromatin was subsequently ligated (circularized) overnight at 16˚ C with 50U T4 ligase. Ligated chromatin was then reverse cross-linked by incubating with proteinase K at 65˚ C and subsequently, the RNA was removed by additional incubation at 37˚ C with RNase A. The purified DNA was further digested with MseI as a second restriction enzyme, followed by circularization of the DNA. 4C product was subsequently amplified with bait-specific inverse primers (see Primers – 4C-seq). From each 4C library, about 3200 ng DNA was amplified in multiple parallel PCR reactions containing 200ng of DNA each, which were subsequently pooled and purified. Amplified bait-containing DNA fragments were ligated to NextFlex DNA barcoded adaptors (Bioo Scientific). Adaptor ligated DNA was purified by Agencourt AMPure XP purification system (Beckman Coulter), PCR amplified (8 cycles) and sequenced paired-end on the Illumina NextSeq 500 to obtain 50bp long reads.

Bioinformatic analyses

Alignment

Tags were mapped to the reference human genome hg19 using the Burrows-Wheeler Alignment Tool[2] (BWA) for ChIP-seq or TopHat2[3] (Bowtie2[4]) for RNA-seq samples. SamTools[5] was used for creation and manipulation of BAM files for each experiment. Before down stream use, duplicates reads and reads with a MAPQ<15 were discarded for ChIP-seq samples. For RNA-seq, only reads with MAPQ<15 were discarded. For ChIP-seq visualization, the number of overlapping sequence reads was determined per base pair, averaged over a 10 bp window and visualized in the UCSC genome browser (http://genome.ucsc.edu). For strand specific RNA-seq visualization, separate tracks were created for both strands, which were displayed in pairs using UCSC trackHubs.

Peak calling

Peak calling software MACS2[6] was used to detect 'sharp' mode binding sites of MLL and all binding sites of H3K4me3, H3K27ac, RUNX1, and CTCF with a q-value cut off of $1e^{-6}$ or $1e^{-9}$. For 'broad' mode MLL binding sites and all H3K79me2 binding sites, HOMER2 findPeaks software[7] was used with the following settings: -fdr $1e^{-6}$ -size 5000 -minDist 10000.

MLL binding site classification

MLL binding sites from MACS2 and HOMER2 were merged with bedTools[8], and split again in 'sharp' and 'broad' mode peak lists using a length cut-off of 7500bp, or split into 'fusion' and 'wild type' binding sites by means of fusion partner signal strength. For this, AF9 (or AF4) tags within MLL binding regions were counted and adjusted to represent the number of tags within a 1 kb region. Subsequently the percentage of these tags as a measure of the total number of sequenced tags of the sample was calculated. Finally, the AF9 (or AF4) normalized tag counts per MLL binding site obtained in this way were binned, and the MLL binding sites containing the upper quartile of AF9 (or AF4) normalized tag counts were identified as MLL-fusion binding sites, while the MLL binding sites containing the lower quartile of AF9 (or AF4) normalized tag counts were identified as MLL wild type binding sites. Overlaps between gene sets were tested for significance using a hypergeometric test with 23503 refseq hg19 genes as total universe.

Tag counting

Tags within a given region were counted and adjusted to represent the number of tags within a 1 kb region. Subsequently the percentage of these tags as a measure of the total number of sequenced tags of the sample was calculated and displayed as heat maps with the Fluff package[9] or as average profiles with ngs.plot software[10].

Peak distribution analysis

To determine genomic locations of binding sites, corresponding peak files were analyzed using a script that annotates binding sites according to all RefSeq hg19 genes. With this script every binding site is annotated either as promoter (2000 bp window around the Transcription Start Site), exon, intron or intergenic (everything else). MLL binding sites were identified as gene targets or distal binding sites based on this classification. Distal MLL binding sites were further characterized as active enhancers based on intersection with H3K27ac peaks using BED tools.

Motif analysis

Peaks were culled to 300 bp, and used with Gimme Motifs[11] software to determine underlying known motifs. Motifs were subsequently grouped into motif families and their relative occurrence as opposed to either motifs called from genomic annotation and length matched random regions or a second set of called peaks (e.g. fusion versus wild type).

Pathway Enrichment analysis

Specific sets of binding regions / genes were analyzed for molecular pathways enrichments using the Gene Set Enrichment Analyses (GSEA)[12,13] web tool with the 'canonical pathways' (CP) collection.

Expression analysis

Normalized (RPKM) values for all refSeq hg19 genes were calculated using HOMER2 analyzeRepeats software. Various subsets of these RPKM tables (e.g. only MLL-fusion bound genes) as well as the set of Ley et al. AML patients blast RPKMs[14] were loaded into an R 3.2.3 environment[15] for visualization as box and bar plots. Raw tag count tables were also generated using HOMER2 analyzeRepeats and loaded into R for distance and Pearson correlation clustering and PCA analyses using the DESeq2 package[16]. After PCA analysis, the top contributing genes per principal component (PC) were extracted by converting the rotation values per PC into z-scores and filtering out only those genes with sigma > 1.5.

4C-seq analysis

A reduced genome was generated by extracting the sequences flanking DpnII sites (30bp on each strand from the DpnII sites to downstream) based on build version hg19 of the human genome. Only uniquely mapped DpnII sites were considered for downstream analysis.

Reads from each library were parsed based on the bait-specific primer sequence and mapped to the reduced genome using bwa (version 0.6.2) with the default parameters. 4C signal was calculated using a sliding window of 10Kb (±5Kb of a given DpnII site) and normalized to the total number uniquely mapped reads. Interactions were called using the r3Cseq R package[17] using an aggregate window of 2 kb and an FDR q-value cut-off of 0.05.

Antibodies

Anti-	Epitope	Cat. No.	Manufacturer
MLL-1	QGQESDSSETSVRGP-C	AT71-Ab1542	Diagenode
MLL-1	PIFDNFRPPPLTPED-C	AT73-Ab1547	Diagenode
AF9	SSASSPLHHEPPPPL L-C	AT22-Ab1327	Diagenode
AF9	FTPSQTRQQGP RSI -C	AT23-Ab1474	Diagenode
AF4	-	Ab31812	Abcam
H3K4me3	-	pAb003-050	Diagenode
H3K27ac	-	pAb-174-050	Diagenode
H3K79me2	-	Ab3594	Abcam
RUNX1	-	Ab23980	Abcam

Primers

ChIP-qPCR

Target	Region	Forward	Reverse
HOXA7	Promoter	AGCTGGGAGACGTTGACTTT	GACAGGCCGGACTTAGACTC
HOXA9	Promoter	GGGAGACGGGAGAGTACAGA	GCTCTACGATGGGGTTTGTT

HOXA10	Promoter	ACCGCAGGATGAAACTGAAG	TTCCCCCAGAAAACAACAAA
MEIS1	Promoter	CGGGTTCTAGCATTCTGGTC	TCTCCCTCTTTGCAAGTGCT
MB	Promoter	GGATTGAGTCTGCCCAGG	GATGGAAGGGCAGAGGTG

cDNA

Name	Target Region	Forward	Reverse
MLL-N1	KMT2A N-terminus	TCAGCTGCAGGGAAGAAAGG	CTATAAACCGCCGAGGGGTC
MLL-N2	KMT2A N-terminus	TGTGAGAATCTTTCAGATGAGATGT	CTTTTTCAAGGGCCAGTCGC
MLL-C	KMT2A C-terminus	CCACAGAGTGTGGGAGGAAC	AGAGGAACTGGATGCCAAGC
AF9-N	MLLT3 N-terminus	TGGCTAGCTCGTGTGCCGTG	CCGGACCGCGTACGAACACC
AF9-C	MLLT3 C-terminus	GCAGCCGAAGTCGCAGAGTTA	AGGGTGGTGGAGGTTCGTGATGT
GAPDH	1st exon	GCCTCAAGATCATCAGCAATGC	TGGTCATGAGTCCTTCCACGA

4C-seq

Name	Reading	Non-Reading
PHLPP1 bait 1	GGAAGAGGTAGTCATTTTGGA	GGGTCCTGATGAGTCTTTGT
PHLPP1 bait 2	GTGGTTCTCGTAAGGGGATC	CTCCCGGCGCATGATTAA
BCL2 bait1	TTTCTCGAGCTCTTGAGATC	GGCAGGAATCCTCTTCTGAT
BCL2 bait 2	ATCAGGTCCTTGGAATGATC	TGCTAGCTCTCTTCTGAAGC

Supplementary Figure Legends

Supplemental Figure 1. **Genome wide binding of MLL and AF9.** (A) The expression of the C-terminal and N-terminal exons of AF9 (MLLT3) in THP-1 cells and monocytes (left). UCSC genome browser screenshot depicting the expression of the AF9 (MLLT3) locus in THP-1 cells. Arrows indicate the location of primers used for rt-qPCR and the fusion breakpoint. (B) UCSC genome browser screenshot depicting the MEIS1 locus in THP-1 and MV4-11 cells. (C) Average signal intensity at MLL-AF9 target genes for MLL (ab1547) ChIP-seq in THP-1 cells (left), and MLL (ab1542) and AF9 (ab1474) ChIP-seq in patient AML_MLLAF9_5. (D) Pathway enrichment for MLL-AF9 (top) and MLL wild type (bottom) target genes.

Supplemental Figure 2. **Genome wide binding patterns of MLL and AF4 in MV4-11 cells.** (A) Schematic representation of MLL, AF4 and MLL-AF4. Antibody binding locations are indicated with dotted lines. (B) ChIP-qPCR experiments using an anti-MLL-1 and an anti-AF4 antibody in MV4-11 cells with primers for HOXA7 -9, -10 and MEIS1. (C) Classification of MLL and AF4 binding events in 'broad' and 'sharp' modes. Left: boxplot showing dispersion of peak lengths. Right: barplot showing genomic distributions. (D) Classification of MLL-AF4 and MLL wild type binding events. Average

profiles showing ChIP-seq signal intensities for MLL-AF4 and MLL-wt binding events in MV4-11 cells. (E) Distribution of MLL-AF4 and MLL wild type binding events in 'broad' and 'sharp' mode. (F) Average signal of H3K27ac, H3K4me3 and H3K79me2 at MLL-AF4 (top) and MLL wild type (middle) target genes, as compared to a random pool of expressed genes (bottom). (G) Pathway enrichment for MLL-AF4 (top) and MLL wild type (bottom) target genes. (H) Motif families on MLL-AF4 target gene promoters enriched over MLL wild type target gene promoters. (I) Venn diagram illustrating the overlap between our human (MV4-111) MLL-AF4 AML targets, MLL-AF9 targets in a mouse LSC model (Bernt et al., 2011), and human MLL-AF4 ALL targets (Guenther et al., 2008).

Supplemental Figure 3: **RUNX1 and CTCF binding at MLL-fusion target genes.** (A) UCSC genome browser screenshots depicting binding of AF9, MLL, RUNX1, CTCF, and H3K79me2 at MLL-AF9 target genes (left) and MLL wt target genes (right). (B) Average signal of RUNX1 under MLL-AF4 (left) and MLL wt (right) target genes in MV4-11 cells.

Supplemental Figure 4. **Characterization of MLL-AF4 bound distal regulator elements.** (A) UCSC genome browser screenshots depicting binding of AF9, MLL, H3K79me2, H3K4me3, and H3K27ac at MLL-AF9 target enhancers. (B) Average signal on MLL-AF4 (left) and MLL wild type (right) bound enhancers for H3K4me3 and H3K27ac (top), MLL and AF4 (middle), and RUNX1 and H3K79me2 (bottom). (C) Rate of co-occupancy of MLL-AF4 and MLL wild type bound enhancers by H3K79me2 and RUNX1 (top). Expression levels of MLL-AF4 and MLL wild type intergenic enhancers (bottom). (D) Distance to nearest TSS for MLL-AF9 bound enhancers. (E) Distance to nearest TSS for MLL-AF4 bound enhancers. (F) Motif family enrichment for MLL-AF4 bound enhancers over MLL-wt bound enhancers.

Supplementary References

1. van de Werken HJG, Landan G, Holwerda SJB, Hoichman M, Klous P, Chachik R, et al. Robust 4C-seq data analysis to screen for regulatory DNA interactions. Nat Meth. 2012 Sep 9;9(10):969–72.

2. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009 Jul 15;25(14):1754–60.

3. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 2013;14(4):R36.

4. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Meth. Nature Publishing Group; 2012 Apr;9(4):357–9.

5. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. Oxford University Press; 2009 Aug 15;25(16):2078–9.

6. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based analysis

of ChIP-Seq (MACS). Genome Biol. 2008;9(9):R137.

7. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Molecular Cell. 2010 May 28;38(4):576–89.

8. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. Oxford University Press; 2010 Mar 15;26(6):841–2.

9. van Heeringen SJ, Georgiou G. fluff: 1.62. 2015 2015-11-25;(url).

10. Shen L, Shao N, Liu X, Nestler E. ngs.plot: Quick mining and visualization of next-generation sequencing data by integrating genomic databases. BMC Genomics. BioMed Central; 2014;15(1):284.

11. van Heeringen SJ, Veenstra GJC. GimmeMotifs: a de novo motif prediction pipeline for ChIP-sequencing experiments. Bioinformatics. Oxford University Press; 2011 Jan 15;27(2):270–1.

12. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA. National Acad Sciences; 2005 Oct 25;102(43):15545–50.

13. Mootha VK, Lindgren CM, Eriksson K-F, Subramanian A, Sihag S, Lehar J, et al. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat Genet. Nature Publishing Group; 2003 Jul;34(3):267–73.

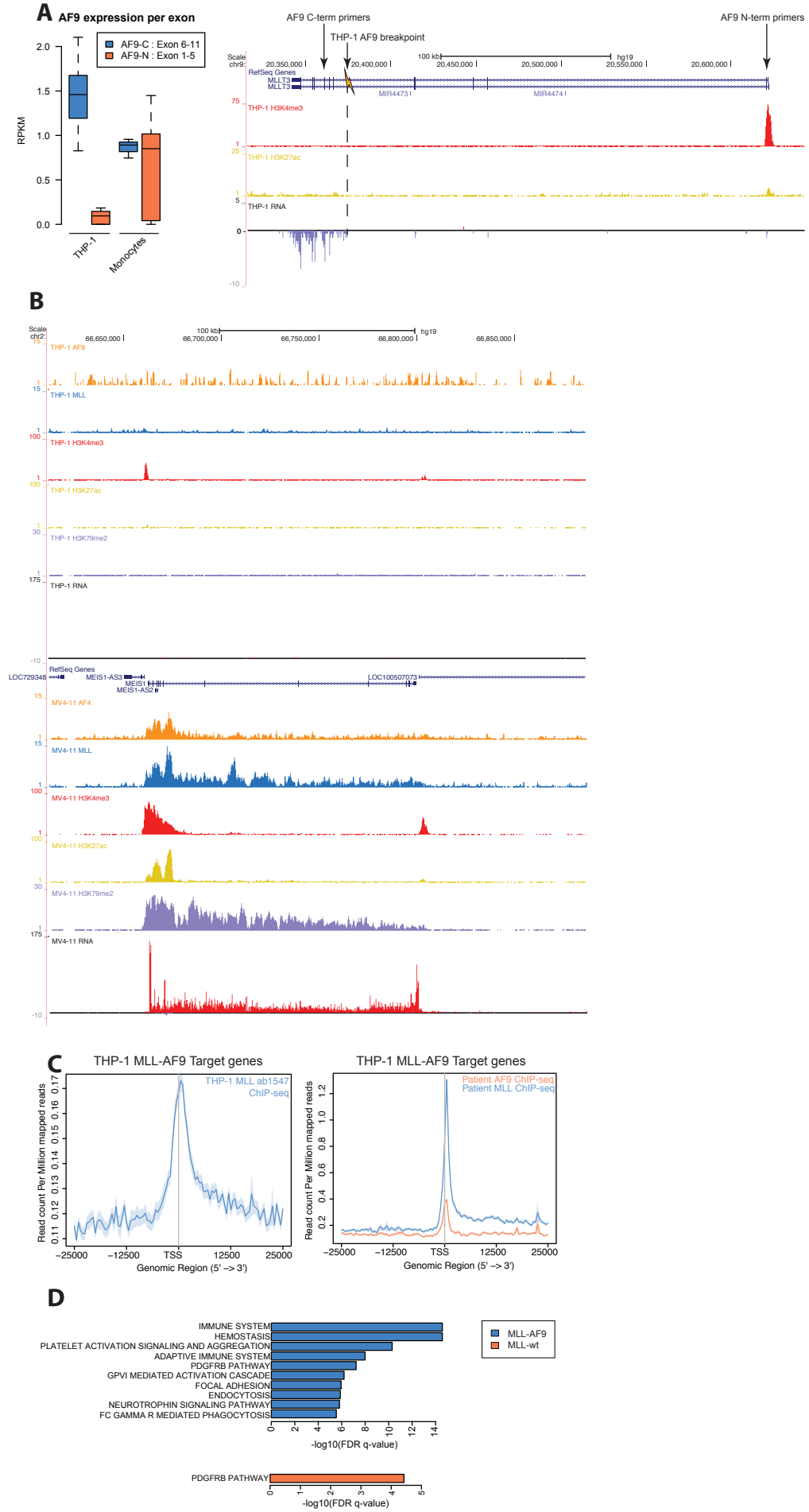
14. Cancer Genome Atlas Research Network. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. N Engl J Med. 2013 May 30;368(22):2059–74.

15. Team RC. R: A language and environment for statistical computing. [Internet]. 3rd ed. Vienna. Available from: https://www.R-project.org/.

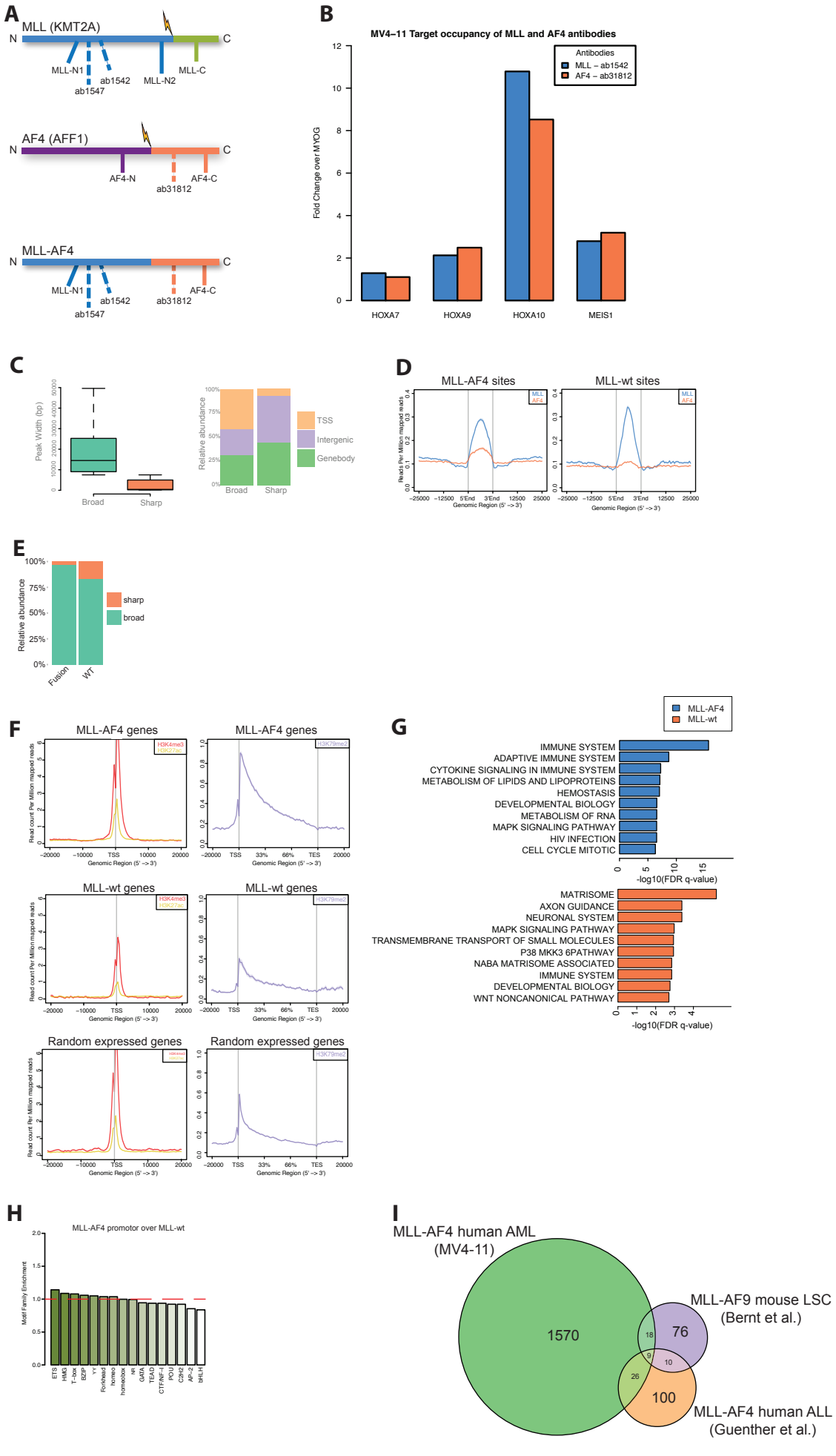
16. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. BioMed Central; 2014;15(12):550.

17. Thongjuea S, Stadhouders R, Grosveld FG, Soler E, Lenhard B. r3Cseq: an R/Bioconductor package for the discovery of long-range genomic interactions from chromosome conformation capture and next-generation sequencing data. Nucleic Acids Res. Oxford University Press; 2013 Jul;41(13):e132–2.

Supplemental Figure 1

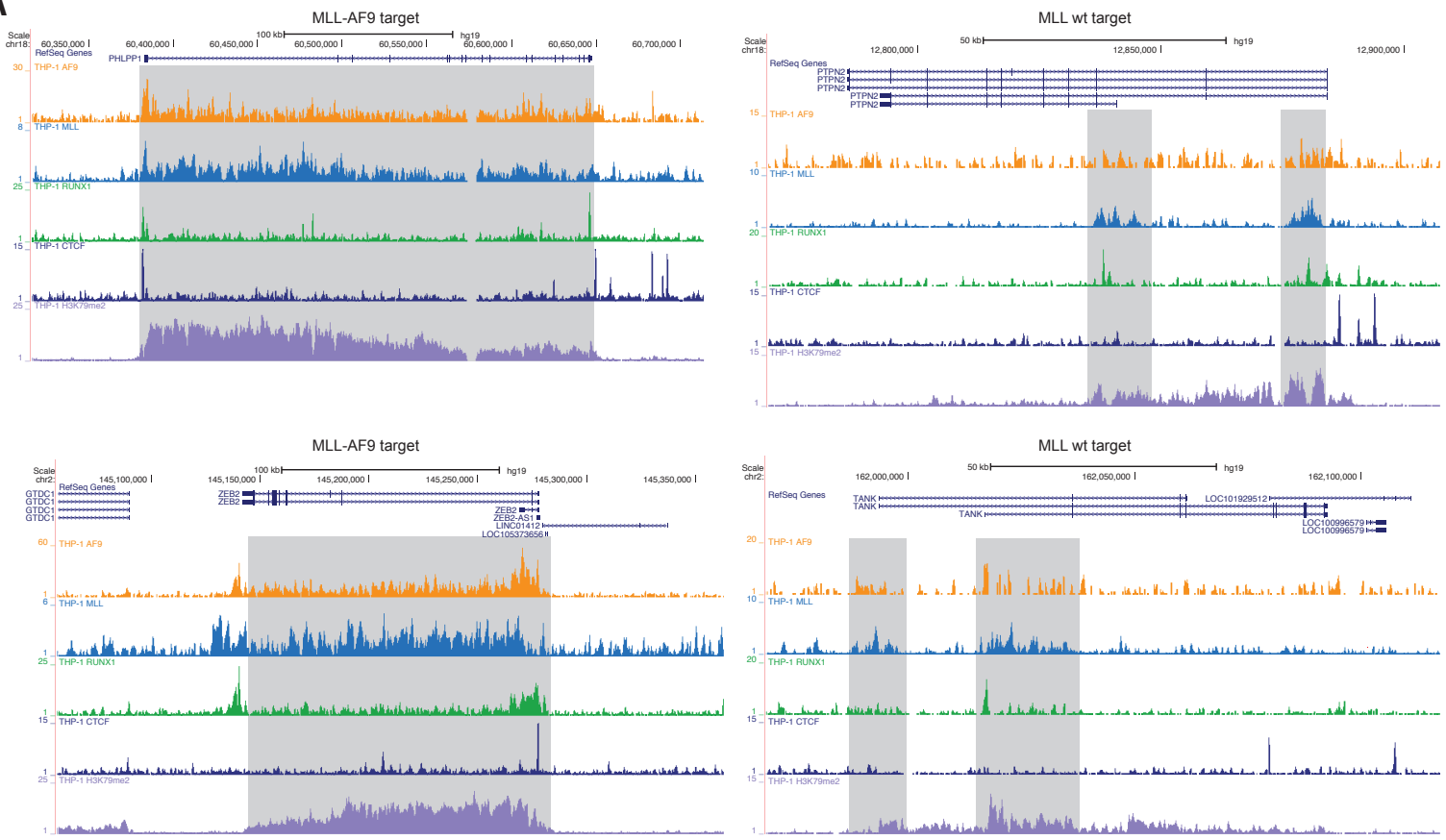


Supplemental Figure 2

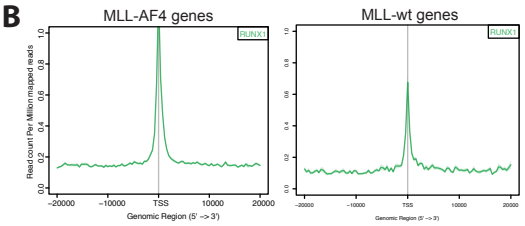


Supplemental Figure 3

A



B



Supplemental Figure 4