Supplemental Data

Integrative methylome-transcriptome analysis unravels cancer-cell vulnerabilities in infant MLL-rearranged B-cell acute lymphoblastic leukemia (Tejedor, Bueno et al 2021)

SUPPLEMENTAL METHODS

WGBSeq

Library construction

Genomic DNA libraries were constructed using the TrueSeq Sample preparation kit following the Illumina standard protocol. Genomic DNA was sheared by sonication followed by end-blunting, dA addition and ligation of solexa adapters. Then, adaptor-ligated molecules of 200 to 300 bp were isolated by agarose gel electrophoresis and subjected to sodium bisulfite conversion using the EZ DNA methylation-Gold kit (Zymo Research Corporation, Irvine, CA) following manufacturer's recommendations. PCR enriched libraries were purified and sequenced (2x150bp) in an Illumina HiSeqX sequencer at CD Genomics (CD Genomic Inc, Shirley, NY, USA).

Read mapping and calculation of cytosine methylation estimates from WGBSeq data

Quality control of the sequenced reads was performed with the FastQC software (v_0.11.7) and adapter removal was performed using Trim Galore (v_0.4.1). Reads were aligned to the reference human genome assembly GRCh37 using Bismark (v_0.19.1)(1) with the following parameters (Read aligner=bowtie2, N=1, L=20). DNA methylation calling was generated with the R/Bioconductor package Methylkit (v_1.2.4)(2), and the resulting CpGs were filtered out for extreme coverage (read depth >10 and < 99% of the maximum coverage) for downstream purposes. Statistics related to WGBS samples are included in **Supplemental Table 1.**

Identification of differentially methylated regions (DMRs) from WGBSeq data

To identify DMR between iB-ALL and healthy FL-BCP we used the R/Bioconductor package bsseq (v_1.18.0)(3) with the following two criteria: (i) a minimum number of CpGs to be included in a DMR of 3, and (ii) a minimum absolute difference between mean DNA methylation values between patients and controls of 0.2. Overlapping between DMRs obtained from different comparisons was calculated with R/Bioconductor package ChIPpeakAnno (v_3.16.1)(4) and Vennerable (v_3.1.0).

Detection of transcription factor binding sites (TFBS) from WGBSeq data or DNA methylation arrays

Transcription factor motif discovery from WGBSeq or high content microarray data was performed with Hypergeometric Optimization of Motif EnRichment (HOMER) software (v_4.10)(5) using the hg19 genome as reference. The genomic coordinates of the observed DMRs or DMPs were selected as target input, while the background was randomly selected by HOMER using sequences with similar features as the query dataset.

DNA pyrosequencing assays

DNA methylation patterns of human repetitive elements (LINE1) were analyzed by bisulfite pyrosequencing using previously validated primers(6). Genomic DNA was isolated using a standard phenol-chloroform extraction and then subjected to bisulfite conversion using the EZ DNA methylation-gold kit (Zymo Research Corporation). Next, converted DNA was PCR-amplified and the pyrosequencing reaction was performed using PyroMark Q24 system (Qiagen, Düsseldorf, Germany).

Microarray-based DNA methylation analysis

Microarray-based DNA methylation profiling was performed with Illumina's Infinium HumanMethylationEPIC 850K beadchip platform(7). Bisulfite conversion of DNA was performed as above. Processed DNA samples were then hybridized to the BeadChip, following Illumina Infinium HD methylation protocol. Genotyping services were provided by the Centro Nacional de Genotipado (CEGEN-ISCIII, Madrid, Spain).

HumanMethylationEPIC Beadchip data preprocessing

IDAT files from the HumanMethylationEPIC Beadchip platform were processed using the R/Bioconductor package minfi (v_1.22.1)(8). In order to adjust for the different probe design types present in the HumanMethylationBeadchip architecture, red and green signals from the IDAT files were corrected using the ssNOOB algorithm with the default parameters (offset=15, dyeCorr=TRUE and dyeMethod="single"). The following probes were discarded for downstream analyses: i) probes overlapping genetic variants (SNP137Common track from UCSC genome browser), ii) probes located in sexual chromosomes, iii) cross-reactive and multimapping probes, and, iv) probes with at least one sample with a detection p-value>0.01. In accordance with the method of Du and colleagues(9), both B-values and M-values were computed and employed across the analysis pipeline. M-values were used for all the statistical analysis assuming homoscedasticity, while B-values were mostly used for intuitive interpretation and visualization of the results, and for correlation analysis between DNA methylation and gene expression (see below).

Batch effect correction

Surrogate Variable Analysis (SVA)(10) was employed to capture the heterogeneity of the underlying methylation data and to account for possible batch effects or confounding variables that might be of interest. Coefficients of the detected surrogate variables (SVs) were later added to the

phenotypical data and included in the definition of the model in order to detect differentially methylated probes (DMPs). The R package svconfound was used to estimate the number of SVs and their coefficients, using group as covariate of interest, and an intercept-only model as the null background model.

Inference of CpG methylation levels from DNA methylation arrays at LINE1 repetitive elements

Genomic coordinates of LINE1 repetitive elements were obtained from the RepeatMasker database (hg19 – Feb2009 – RepeatMasker open-4.0.5 – Repeat Library 20140131). Since most of the microarray probes lack coverage within repetitive elements, we selected probes located in close proximity to DNA repeats (±20 bp) as a surrogate for their methylation status, as recently modelled by Zheng and colleagues(11).

Identification of differentially methylated probes (DMPs)

Significant methylation of a specific probe was determined by the moderated t-test implemented in the R/Bioconductor package limma (v_3.38.3)(12). A linear model, with methylation level as response and group as the main covariate of interest, was fitted to the methylation data. SVs generated using SVA were also included in the model definition. Contrasts were then defined as linear combinations of the different values the main covariate of interest could take, and each contrast generated a coefficient p-value for each probe. P values were corrected for multiple testing using the Benjamini-Hochberg method for controlling the false discovery rate (FDR). A FDR threshold of 0.05, and a minimum absolute difference between mean DNA methylation values of cases and controls of 0.25 was employed to determine DMPs. For SEM-WT versus SEM-FOSL2^{KO} and t(4:11)/*MLL-AF4*+ CD34+ versus non-manipulated CD34+ comparisons, DMPs were determined using a minimum absolute difference between cases and controls of 0.4 and 0.2, respectively.

CpG island status and genomic region analyses

Differentially methylated CpG sites (dmCpGs) were assigned to their corresponding genomic context or genomic location using the R/Bioconductor packages IlluminaHumanMethylationEPIC.anno.ilm10b2.hg19 (v_0.6.0) and ChIPseeker (v_1.18.0)(13) respectively. Odds ratio (OR) enrichment and statistical significance were calculated by means of two-sided Fisher's tests. For the different comparisons, appropriate background including all filtered CpG probes interrogated by the HumanMethylationEPIC Beadchip platform was used in order to calculate statistical significance.

Region set enrichment analysis

Chromatin and repeat enrichment analyses were performed with the R/Bioconductor package LOLA (v_1.4.0)(14) and region datasets were downloaded from the LOLA extended software environment (http://databio.org/regiondb), and the RepeatMasker database. DMR enrichments in DNA repetitive elements were calculated using data from hg19 DNA repeats obtained from RepeatMasker using as background all the CpG sites identified in any condition of the WGBS experiment.

dmCpG enrichments in six histone marks (H3K4me1, H3K4me3, H3K27me3, H3K36me3, H3K9me3 and H3K27ac) were calculated using ChIP-seq tracks from 6 epigenomes obtained from ENCODE and the NIH Roadmap Epigenome Consortia(15, 16). Chromatin state data from these same tissue/cell types were obtained from NIH Roadmap's ChromHMM expanded 18-state model (obtained from http://egg2.wustl.edu/roadmap/). dmCpG enrichments in TFBSs were performed using data from human meta-clusters obtained from the GTRD database(17). Clustered peaks corresponding to 476 human TFs across a panel of distinct cells and tissue types were used for

statistical purposes. For dmCpG analyses, enrichment significance was calculated using one-sided Fisher's tests (adjusted p-value <0.05), comparing the overlap of dmCpGs with the dataset of interest and using the set of filtered probes from the HumanMethylationEPIC as background.

DNA enhancers

The list of B-cell enhancers was obtained from Enhancer Atlas database (http://enhanceratlas.org/)(18). To analyze those enhancers with the strongest influence along the B-cell differentiation process, only enhancer regions with a peak score >2 and enhancer regions located at a maximum of 600 bp from a dmCpG observed in healthy cord blood (CB)-derived CD19+ B-cells were selected for downstream analyses. Correlation between DNA methylation (β-values) at enhancer elements on CD19+ B-cells and BCPs or CD19+ leukemic blasts from different iB-ALL subtypes was represented as two-dimensional kernel density estimation using the R/Bioconductor package MASS (v_7.3-47) with the probes overlapping the aforementioned enhancer elements (n=327).

RNA-seq

RNA-seq data was previously generated by Agraz-Doblas and colleagues for CD19+ leukemic blasts from iB-ALL patients (n=40) and healthy FL-BCP (n=5)(19). FASTQ files from iB-ALL and healthy FL-BCP were obtained from the European Nucleotide Archive under the accession number PRJEB23605. Raw data corresponding to healthy CD19+ B-cells was obtained from GEO dataset GSE74246 (GSM1915578, GSM1915582 and GSM1915592 respectively)(20). Raw data from 3 untransduced, 6 MLL-af4-transduced, and 3 MLL-AF9-transduced healthy CD34+ cells were obtained from ENA study PRJNA309171(21). RNA extraction from SEM cells was performed as in Agraz-Doblas and colleagues(19) and paired RNAseq data was generated using an Illumina Novaseq 2x150 bp platform at Genewiz (Leipzig, Germany). Adapter removal was performed using

Trim Galore (v_0.4.1) and reads were mapped to the reference human genome assembly GRCh37 using RSEM (v_1.3.1)(22). To identify differentially expressed genes (DEGs, q-value <0.001), we used the R/Bioconductor package EBSeq (v_1.22.1)(23) with the matrix of read counts obtained from RSEM. Correction by library size, as well as filtering out of low expressed genes was automatically performed by the EBSeq software. To control transcriptional noise, gene expression variability across the different conditions was measured as the ratio of the standard deviation δ to the mean μ (also known as coefficient of variation) using the FPKM matrix of DEGs identified in B-cells or iB-ALL subtypes as compared to healthy FL-BCPs. Gene set enrichment analyses (GSEA) comparing SEM-WT versus SEM-FOSL2^{KO} samples were performed using a basic fold-change approach in the GSEA pre-ranked mode.

Modular co-expression analyses

Gene co-expression network analysis was performed with the R/Bioconductor package CEMITool (v_1.6.11)(24). Gene filtering was done in accordance to gene variance with a p-value cutoff of 0.05, resulting in the selection of the 663 most variable genes. Module enrichment was performed using the gene set enrichment analysis (GSEA) function from the R/Bioconductor package fgsea (v_1.8.0). Over Representation Analyses of the biological functions associated with each of the modules was performed via the R/Bioconductor package clusterProfiler (v_3.10.1)(25) using the c5.all.v5.2.symbols.gmt and the h.all.v5.2.symbols.gmt gene sets from the MSigDb database(26, 27). Annotated module graphs were recreated with a matrix of *Homo sapiens* protein-protein interactions obtained from the HitPredict database(28).

Pathway enrichment analyses

Genes with consistent inverse correlation between expression and methylation levels were selected to interrogate the Reactome annotation database (R/Bioconductor package ReactomePA,

v_1.20.2)(29). Reactome pathway enrichment was performed with the R/Bioconductor package clusterProfiler (v_3.10.1). The total number of filtered genes (18,668) involved in such correlation analysis was used to set the background for appropriate ontology comparisons.

Network representation

Correlation pairs network representation between genes and their associated methylation loci were generated using the R/CRAN package igraph (v_1.2.4). Network nodes represent either correlated CpG sites or genes, while network edges indicate interactions between CpGs with strong correlation with gene expression and their corresponding target genes.

SUPPLEMENTAL REFERENCES

1. Krueger F, Andrews SR. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics 2011;27(11):1571–1572.

2. Akalin A et al. methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. Genome Biol. 2012;13(10):R87.

3. Hansen KD, Langmead B, Irizarry RA. BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. Genome Biol. 2012;13(10):R83.

4. Zhu LJ et al. ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data. BMC Bioinformatics 2010;11:237.

5. Heinz S et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol. Cell 2010;38(4):576–589.

6. Bollati V et al. Changes in DNA methylation patterns in subjects exposed to low-dose benzene. Cancer Res. 2007;67(3):876–880.

7. Moran S, Arribas C, Esteller M. Validation of a DNA methylation microarray for 850,000 CpG sites of the human genome enriched in enhancer sequences. Epigenomics 2016;8(3):389–399.

8. Fortin J-P, Triche TJ, Hansen KD. Preprocessing, normalization and integration of the Illumina HumanMethylationEPIC array with minfi. Bioinformatics 2017;33(4):558–560.

9. Du P et al. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. BMC Bioinformatics 2010;11:587.

10. Leek JT, Storey JD. Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS Genet. 2007;3(9):1724–1735.

11. Zheng Y et al. Prediction of genome-wide DNA methylation in repetitive elements. Nucleic Acids Res. 2017;45(15):8697–8711.

12. Ritchie ME et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43(7):e47.

13. Yu G, Wang L-G, He Q-Y. ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. Bioinformatics 2015;31(14):2382–2383.

14. Sheffield NC, Bock C. LOLA: enrichment analysis for genomic region sets and regulatory elements in R and Bioconductor. Bioinformatics 2016;32(4):587–589.

15. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature 2012;489(7414):57–74.

16. Roadmap Epigenomics Consortium et al. Integrative analysis of 111 reference human epigenomes. Nature 2015;518(7539):317–330.

17. Yevshin I, Sharipov R, Valeev T, Kel A, Kolpakov F. GTRD: a database of transcription factor binding sites identified by ChIP-seq experiments. Nucleic Acids Res. 2017;45(D1):D61–D67.

18. Gao T et al. EnhancerAtlas: a resource for enhancer annotation and analysis in 105 human cell/tissue types. Bioinformatics 2016;32(23):3543–3551.

19. Agraz-Doblas A et al. Unraveling the cellular origin and clinical prognostic markers of infant B-cell acute lymphoblastic leukemia using genome-wide analysis. Haematologica 2019;104(6):1176–1188.

20. Corces MR et al. Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. Nat. Genet. 2016;48(10):1193–1203.

21. Lin S et al. Instructive Role of MLL-Fusion Proteins Revealed by a Model of t(4;11) Pro-B Acute Lymphoblastic Leukemia. Cancer Cell 2016;30(5):737–749.

22. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 2011;12:323.

23. Leng N et al. EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments. Bioinformatics 2013;29(8):1035–1043.

24. Russo PST et al. CEMiTool: a Bioconductor package for performing comprehensive modular co-expression analyses. BMC Bioinformatics 2018;19(1):56.

25. Yu G, Wang L-G, Han Y, He Q-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS 2012;16(5):284–287.

26. Subramanian A et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. U.S.A. 2005;102(43):15545–15550.

27. Liberzon A et al. The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst 2015;1(6):417–425.

28. Patil A, Nakai K, Nakamura H. HitPredict: a database of quality assessed protein-protein interactions in nine species. Nucleic Acids Res. 2011;39(Database issue):D744-749.

29. Yu G, He Q-Y. ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization. Mol Biosyst 2016;12(2):477–479.

SUPPLEMENTAL TABLE TEXT

Supplemental Table 1 | WGBSeq statistics corresponding to iB-ALL samples and healthy BCPs.

Supplemental Table 2 | Detailed list of differentially methylated regions (DMR) between healthy BCPs and *MLL-AF4+, MLL-AF9+* and non-*MLL*r iB-ALL patients. Columns indicate genomic coordinates and related statistical information, including the number of CpGs contained in a given region, the average methylation status per condition and the direction of the change.

Supplemental Table 3 | Statistical information about DMR enriched in DNA repetitive elements (**Figure 1E**). Enrichments were calculated between the DMRs in each analyses and the full

collection of DNA repetitive regions from RepeatMasker (hg19) by means of the LOLA extended software.

Supplemental Table 4 | Clinico-biological features of the patients enrolled in this study and healthy BCPs. Abbreviations: FL: Fetal Liver, CB: Cord Blood, BM: Bone Marrow. *Months. **22 weeks-old fetal age. ***>98% purity after FACS sort of CD34+CD19+ BCP from normal FL.

Supplemental Table 5 | Statistics for HOMER enrichment analyses of WGBSeq data at TFBS (**Supplemental Figure 2**). Common or specific hyper- or hypomethylated DMRs for the different iB-ALL subtypes were used for these analyses.

Supplemental Table 6 | List of dmCpG sites in leukemic blasts from the three iB-ALL subtypes and healthy naïve B-cells. The accompanying "Annotations" sheets include information on the statistics, genomic coordinates and information related to the CpG island, associated gene name and region type.

Supplemental Table 7 | Statistical assessment of the iB-ALL enrichments at the CpG context and CpG locations (**Figure 3** and **Supplemental Figure 6**) as compared to the background distribution of the MethylationEPIC array platform. Data includes the adjusted p-value (Bonferroni correction) from all possible pairwise comparisons.

Supplemental Table 8 | Statistics related to dmCpG enrichment analyses (**Figure 3**) at TFBS from the GTRD database. Enrichments were calculated between the dmCpGs in each iB-ALL subtype (or healthy naïve B-cells) and the full collection of human TF motifs from the GTRD database (hg19) using the LOLA extended software.

Supplemental Table 9 | Histone mark enrichment analysis of dmCpGs in B-cells and iB-ALL. Enrichments were calculated between the dmCpGs of each experimental group and the full collection of Roadmap epigenomics (hg19) regions integrated in LOLA extended software. Corresponding array backgrounds were used for the different comparisons.

Supplemental Table 10 | Chromatin state enrichment analysis of dmCpGs in normal B-cells and iB-ALL. Enrichments were calculated between the dmCpGs of each experimental group and the chromatin segmentation regions (hg19, ChromHMM, 18 states) obtained from Roadmap and ENCODE consortia. A custom LOLA database including information related to the chromatin states in the different tissues/cell lines and the corresponding array background was used for proper enrichment calculation.

Supplemental Table 11 | List of CpG probes located within B-cell enhancer elements (related to **Supplemental Figure 4**). B-cell enhancer elements were obtained from enhancer atlas database. Data includes information corresponding to the genomic coordinates of these enhancer-related probes, their associated enhancer-gene interaction, and the relation to CpG Island and region type.

Supplemental Table 12 | List of DEG in iB-ALL and healthy naïve B-cells versus healthy BCPs. Information related to the statistical analyses performed with EBSeq is shown.

Supplemental Table 13 | List of CpG sites displaying robust correlation (>0.5 pearson corr) between DNA methylation and gene expression in iB-ALL. Columns indicate the relation between genes and particular CpG probes, the correlation score and the statistics for DNA methylation-gene expression integration, as calculated by the R/Bioconductor ELMER package.

Supplemental Table 14 | Histone mark enrichment analysis of CpG sites displaying robust correlation between DNA methylation and gene expression in iB-ALL. Enrichments were calculated between the highly correlated dmCpG sites obtained with ELMER in each of the analyses and the full collection of Roadmap epigenomics (hg19) regions integrated in LOLA extended software.

Supplemental Table 15 | Chromatin state enrichment analysis of CpG sites displaying robust correlation between DNA methylation and gene expression in iB-ALL. Enrichments were calculated between the highly correlated dmCpG sites calculated by the ELMER software and the chromatin segmentation regions (hg19, ChromHMM, 18 states) obtained from Roadmap and ENCODE consortia.

Supplemental Table 16 | List of dmCpG sites in SEM-FOSL2^{KO} as compared to SEM-WT cells. The accompanying columns include information on the genomic coordinates and information related to the CpG island, associated gene name and region type.

Supplemental Table 17 | Statistics for HOMER enrichment analyses of SEM-FOSL2^{KO} hyper- and hypo dmCpGs at TFBS.

Supplemental Table 18 | List of dmCpG sites in CD34$^{CRISPR\ t(4:11)}$ versus unedited CD34+ cells. The accompanying columns include information on the genomic coordinates and information related to the CpG island, associated gene name and region type.

Supplemental Table 19 | Statistics related to dmCpG enrichment analyses (**Figure 9F**) at TFBS from the GTRD database. Enrichments were calculated between the dmCpGs in CD34$^{CRISPR\ t(4:11)}$

vs unedited CD34 cells and the full collection of human TF motifs from the GTRD database (hg19) using the LOLA extended software.

Supplemental Table 20 | CRISPR/Cas9 crRNA sequences and primer sets used in this study.

SUPPLEMENTAL FIGURE LEGENDS

Supplemental Figure 1 | Validation of DNA methylation at LINE1 Repetitive Elements. A) Bisulfite pyrosequencing illustrating the DNA methylation status of two CpG sites located at LINE1 elements in an extended cohort of iB-ALL patients. Dashed line indicates the average methylation level of healthy BCPs. **B)** Violin plot representing β-value distributions of CpG sites obtained from the EPIC methylation array platform located at 20 bp from described human LINE1 repetitive elements as defined by RepeatMasker. The average value of CpG methylation for the indicated iB-ALL samples or healthy BCPs is represented. "n" denotes the number of CpGs included for the analysis.

Supplemental Figure 2 | HOMER enrichment analysis of known transcription factor motifs enriched in DMRs from iB-ALL patients. A) Heatmaps depicting the results from the *in silico* motif discovery pipeline to identify TFBS enriched in hyper- or hypomethylated DMRs. DMRs were classified as "common" when shared by at least two iB-ALL subgroups. **B-E)** Motif logo and chart representation of the observed p-values, the number of targets and the percentage of identified targets in the query and the background dataset for the transcription factors CEBP (**B**), FOSL2 (**C**), JUN (**D**) and RUNX1 (**E**), respectively.

Supplemental Figure 3 | Correlation between Human Methylation EPIC platform and WGBSeq. A) Density plots represent a paired β-value comparison for the same CpG sites measured with the Human Methylation EPIC platform (x-axis) and WGBSeq (y-axis) for the indicated samples/patients. Pearson correlation and the total number of coinciding CpG sites are indicated for each comparison. **B)** Violin plots depicting the overall β-value distribution of the interrogated CpG sites (indicated in the upper right corner) obtained with the Human Methylation EPIC platform (left panel) or WGBSeq (right panel). For panels **A** and **B**, statistical significance was calculated using a two-sided Wilcoxon rank sum test (***= $p<0.001$).

Supplemental Figure 4 | *MLL-AF4*+ shows an aberrant enhancer methylation associated to a less differentiated B-cell phenotype. A) Venn diagram representing the overlapping dmCpGs between robust B-cell enhancer elements identified in the Enhancer Atlas database (pink, enhancer score values>2) and dmCpG sites from the Human Methylation EPIC platform with substantial absolute differences (β-value>0.25) between healthy naïve B-cells and BCP. Only enhancers located within a 600 bp from a given dmCpG site and displaying consistent overlap with the probes from the Human Methylation EPIC array were selected for downstream analysis. **B)** Hierarchical clustering (Ward.D method) of iB-ALL patients (x-axis) using the naïve B-cells-specific 361 dmCpG probes from the Human MethylationEPIC platform (y-axis) located in close proximity to B-cell enhancer elements. **C)** Heat density scatterplots reflecting the correlation in DNA methylation levels (β-values) at B-cell enhancer elements between BCPs, each iB-ALL subtype and healthy naïve B-cells. Red and blue areas identify high or low probe density, respectively. Pearson correlation value for each comparison is shown (all pvalues < 0.001).

Supplemental Figure 5 | Co-expression network analysis identifies gene sets enriched in the differentially co-expressed modules. A) Line plot representing the gene expression profile of the

genes corresponding to a given module across all the samples analyzed. Each line corresponds to a given gene, and the vertical bold lines represent the average expression of the correlated genes from a given module. **B** and **C)** Gene set enrichment analysis of the pathways included in the GO gene set collection (**B**) or the Hallmark gene set collection (**C**) for genes shown in a. For **B** and **C**, the length and the color of the bar denotes the enrichment in a particular gene set category, by means of the –Log10 adjusted p-value. **D-F)** Network representation of the top gene hubs contributing to module 1 (**D**), 2 (**E**) and 3 (**F**), respectively. Dots represent network nodes (genes), and edges represent interaction links between those co-expressed genes obtained from HitPredict interactome database. Dot size indicates the number of interactions of a given node.

Supplemental Figure 6 | Genomic distribution and enrichment analyses of CpG sites showing robust correlation between DNA methylation and gene expression in iB-ALL. A and B) Stacked barplots representing the relative frequency of significant hyper- or hypomethylated CpGs in relation to their CpG context (**A**) or CpG location (**B**). Only gene expression-correlating dmCpGs (absolute cor>0.5) are included in the analysis. **C)** Heatmaps depicting histone mark enrichment analyses of such dmCpGs. Color scale represents odd ratio of significant dmCpGs obtained in previous analyses across six common histone modifications from the NIH Roadmap Epigenome consortium as compared with the background distribution of the Human Methylation EPIC platform. Bottom legend indicates the type of normal hematopoietic datasets used in these comparisons. **D)** Heatmaps displaying chromatin state enrichment analyses for such CpGs. Colour range indicates the odd ratio of the significant dmCpGs observed across 18 chromatin states obtained from the NIH Roadmap Epigenome consortium.

Supplemental Figure 7 | FOSL2 modulates the methylation status of downstream target motifs. A) Network representation of the top DNA methylation-gene expression interactions of

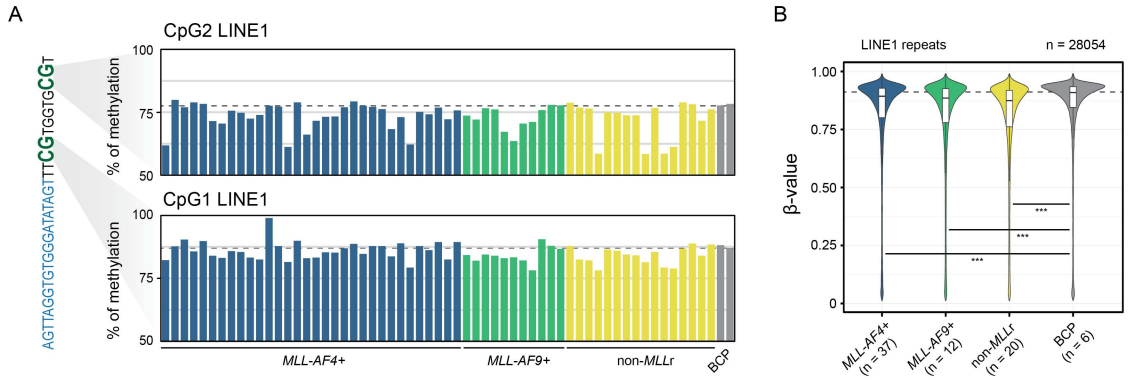
targets with FOSL2 binding motifs. Nodes represent genes and dmCpG sites, and edges represent interactions between dmCpGs and DEGs. Up- and downregulated genes are represented with red and blue colors, respectively. Hyper- and hypomethylated CpG sites are depicted in yellow or pale blue, respectively. Dot size is proportional to the degree (number of connections) of a given node. **B and C)** Boxplot depicting the average methylation (β-value) of the significant *DUSP10* or *CD44* expression-correlating CpG probes across healthy BCPs, iB-ALL subgroups and naïve B-cells. **D and E)** Boxplot reflecting the expression of *DUSP10 or CD44* in the aforementioned groups.

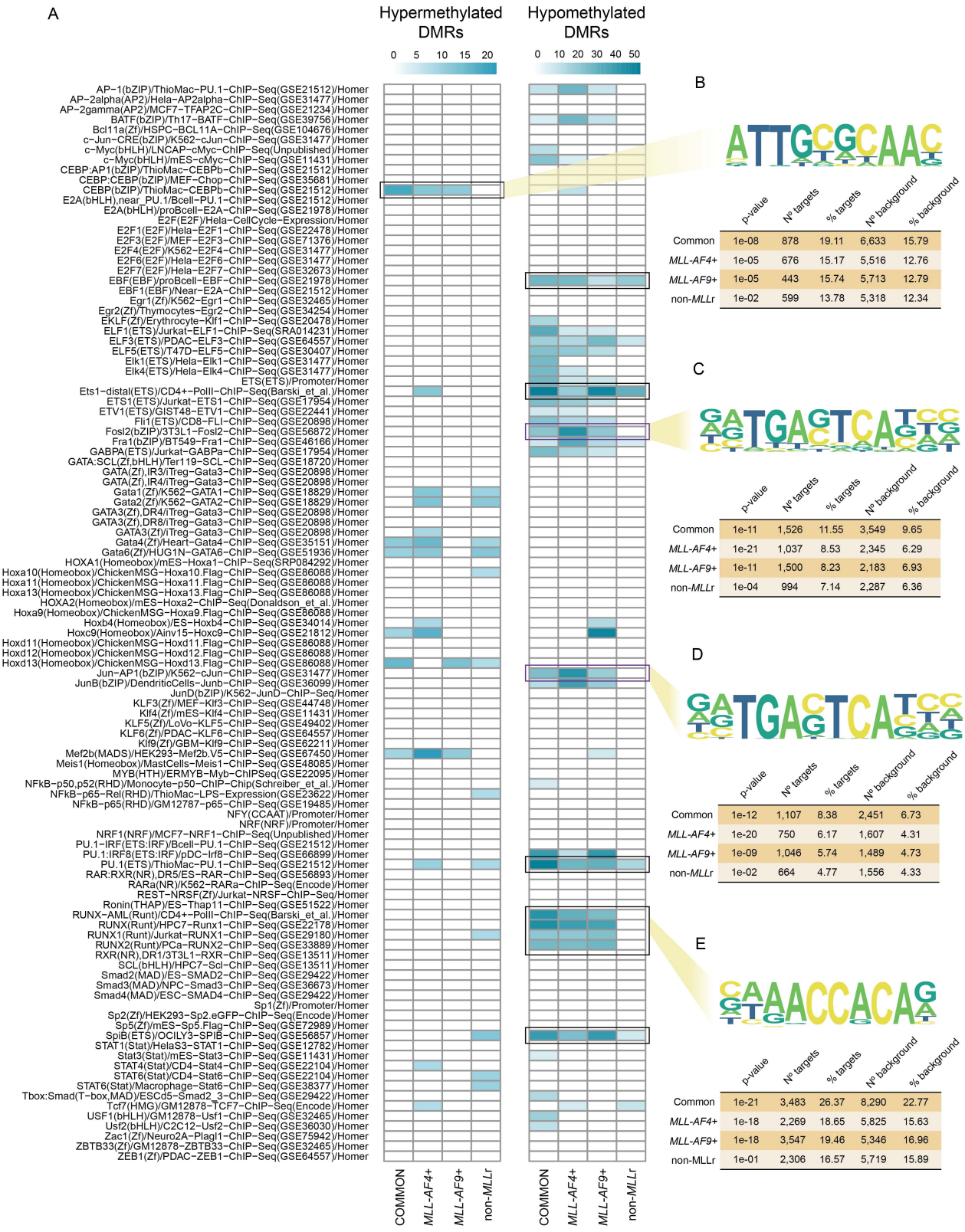
Supplemental Figure 8 | RUNX1 interacting factors control the methylation status of downstream target motifs. A) Ideogram representing the genomic location of RUNX1 expression-correlating dmCpG sites. n, denotes the number of significant correlating dmCpGs identified with ELMER algorithm. **B)** Boxplot depicting the average methylation (β-value) of the significant *RUNX1* expression-correlating CpG probes across healthy BCPs, iB-ALL subgroups and naïve B-cells. **C)** Boxplot reflecting the expression of *RUNX1* in the indicated groups. **D)** Scatter plot showing the correlation between average DNA methylation of RUNX1 motif targets (x-axis) with the expression of RUNX1 (y-axis). Colored dots: blue, BCP; red, *MLL-AF4*+; green, *MLL-AF9*+; yellow, non-*MLL*r. **E)** Violin plots indicating the distribution of gene expression changes (Log2 fold change of the indicated groups versus healthy BCPs) of target genes with RUNX1 motif obtained with ELMER algorithm (two-sided Wilcoxon rank sum test. *** $p<0.001$). All correlated genes with RUNX1 motif included in any of the iB-ALL subgroups were used for the representation of the B-cell gene expression distribution. The “random” group includes a random sampling of the same number of genes included in the B-cell group, but using the original gene expression matrix including all genes with detectable expression in the RNAseq dataset. **F)** Network representation of the top DNA methylation-gene expression interactions of targets with RUNX1 binding motifs. Nodes represent genes and dmCpG sites, and edges represent interactions between dmCpGs and DEGs. Up- and

downregulated genes are represented with red and blue colors, respectively. Hyper- and hypomethylated CpG sites are depicted in yellow or pale blue, respectively. Dot size is proportional to the degree (number of connections) of a given node. **G-H)** Same as b-c, but for the RUNX1 target, *RUNX2*.

Supplemental Figure 9 | MLLr governs the expression pattern of AP-1 members. A) Boxplot showing the average expression of *FOS, FOSL1, FOSB, JUN, JUNB, JUND* and *GUSB* (negative control) in BCP, *MLL-AF4*+, *MLL-AF9*+, non-*MLL*r, naïve B-cell as well as in healthy untransduced CD34+ cells or in CD34+ cells transduced with either human:murine chimeric MLL-af4 or human MLL-AF9. **B)** UCSC Genome Browser tracks representing the binding pattern of MLL-af4 (in CD34+ cells) or MLLN/AF4C (in SEM cells) in the vicinity of *FOS, FOSL1, FOSB, JUN, JUNB, JUND* and *GUSB* genes. Data represents ChIPseq signals obtained from GSE84116 and GSE74812, respectively. **C)** Barplots depicting RT-PCR relative fold-change of *FOS, FOSL1, FOSB, JUN, JUNB*, and *JUND* expression between non-edited CD34+ cells ($CD34^{control}$) and CRISPR-edited CD34+ cells carrying locus-specific t(4;11)/MLL-AF4+ ($CD34^{CRISPR\ t(4;11)}$). Barplots represent mean±SD (two-sided Welch's t-test. *p<0.05).

SUPPLEMENTAL FIGURES


A
CpG2 LINE1
CpG1 LINE1
% of methylation
100
75
50
AGTTAGGTGTGGGATATAGTTTCGTGGTGCGT
MLL-AF4+
MLL-AF9+
non-MLLr
BCP
B
LINE1 repeats
n = 28054
β-value
1.00
0.75
0.50
0.25
0
***
MLL-AF4+ (n = 37)
MLL-AF9+ (n = 12)
non-MLLr (n = 20)
BCP (n = 6)



B

	p-value	Nº targets	% targets	Nº background	% background
Common	1e-08	878	19.11	6,633	15.79
MLL-AF4+	1e-05	676	15.17	5,516	12.76
MLL-AF9+	1e-05	443	15.74	5,713	12.79
non-*MLLr*	1e-02	599	13.78	5,318	12.34

C

	p-value	Nº targets	% targets	Nº background	% background
Common	1e-11	1,526	11.55	3,549	9.65
MLL-AF4+	1e-21	1,037	8.53	2,345	6.29
MLL-AF9+	1e-11	1,500	8.23	2,183	6.93
non-*MLLr*	1e-04	994	7.14	2,287	6.36

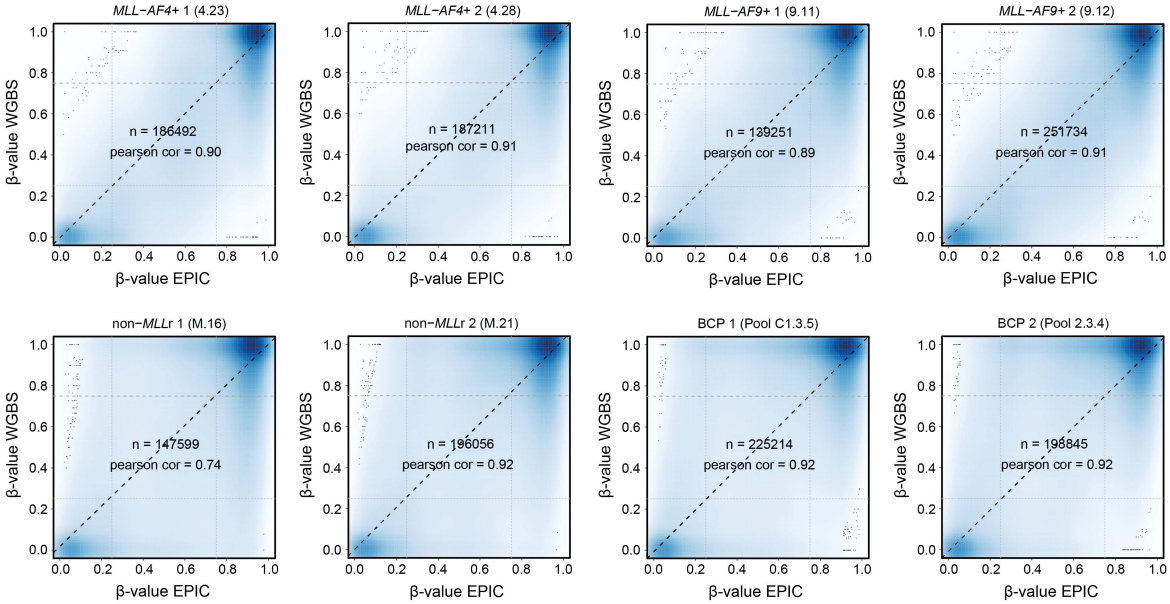
D

	p-value	Nº targets	% targets	Nº background	% background
Common	1e-12	1,107	8.38	2,451	6.73
MLL-AF4+	1e-20	750	6.17	1,607	4.31
MLL-AF9+	1e-09	1,046	5.74	1,489	4.73
non-*MLLr*	1e-02	664	4.77	1,556	4.33

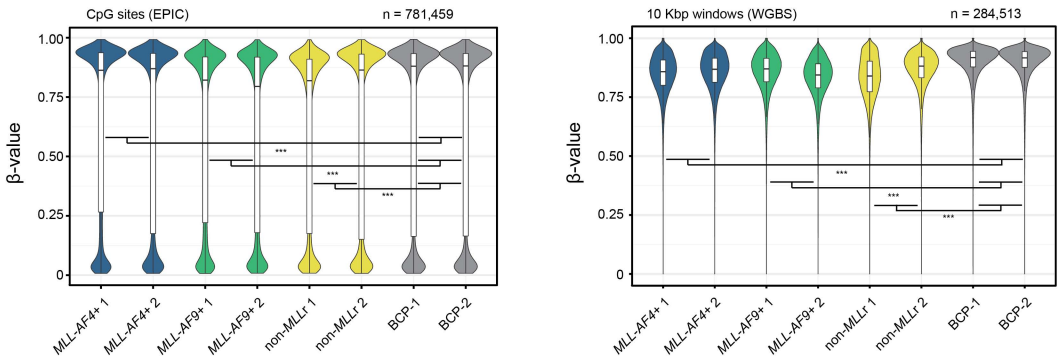
E

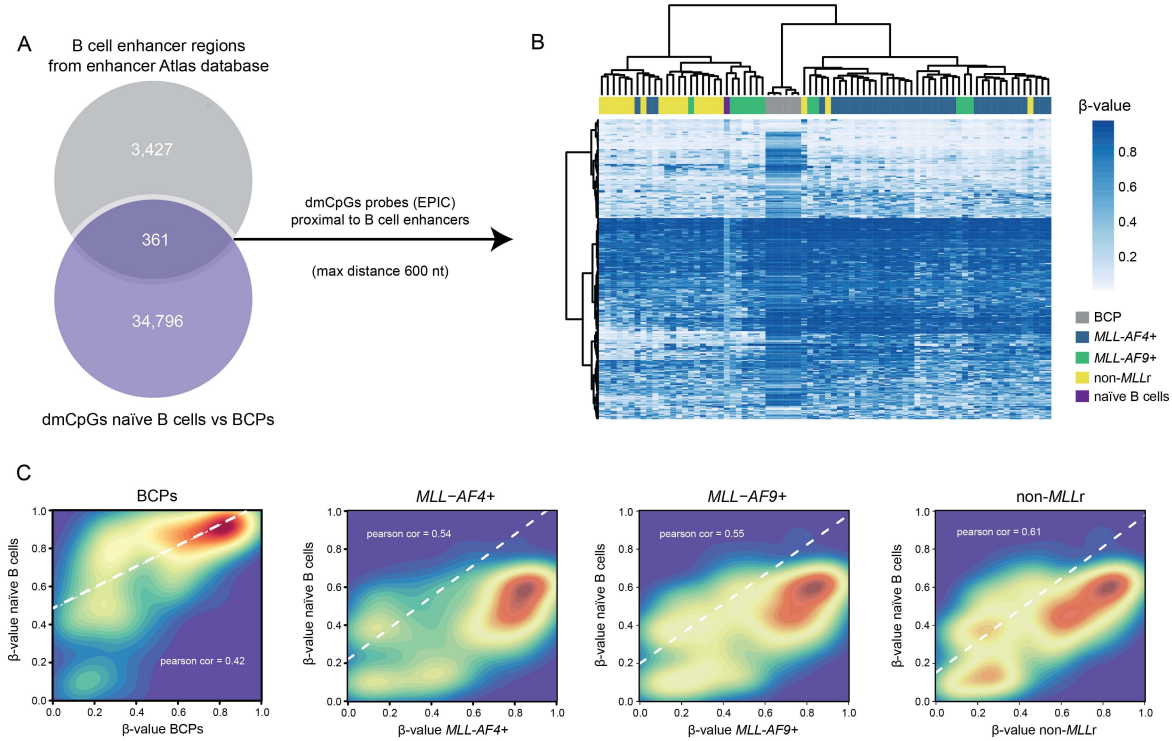
	p-value	Nº targets	% targets	Nº background	% background
Common	1e-21	3,483	26.37	8,290	22.77
MLL-AF4+	1e-18	2,269	18.65	5,825	15.63
MLL-AF9+	1e-18	3,547	19.46	5,346	16.96
non-MLLr	1e-01	2,306	16.57	5,719	15.89

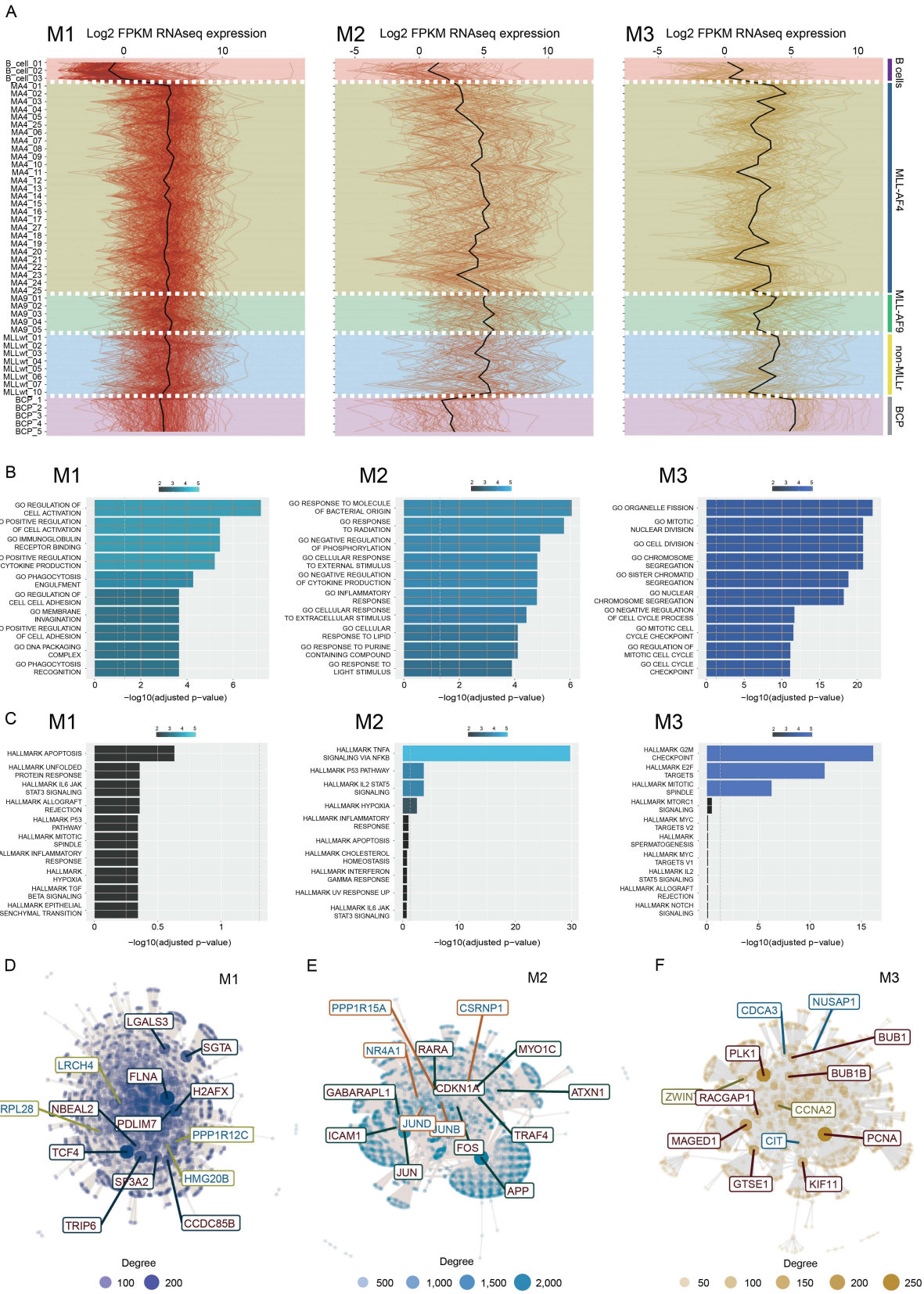
A

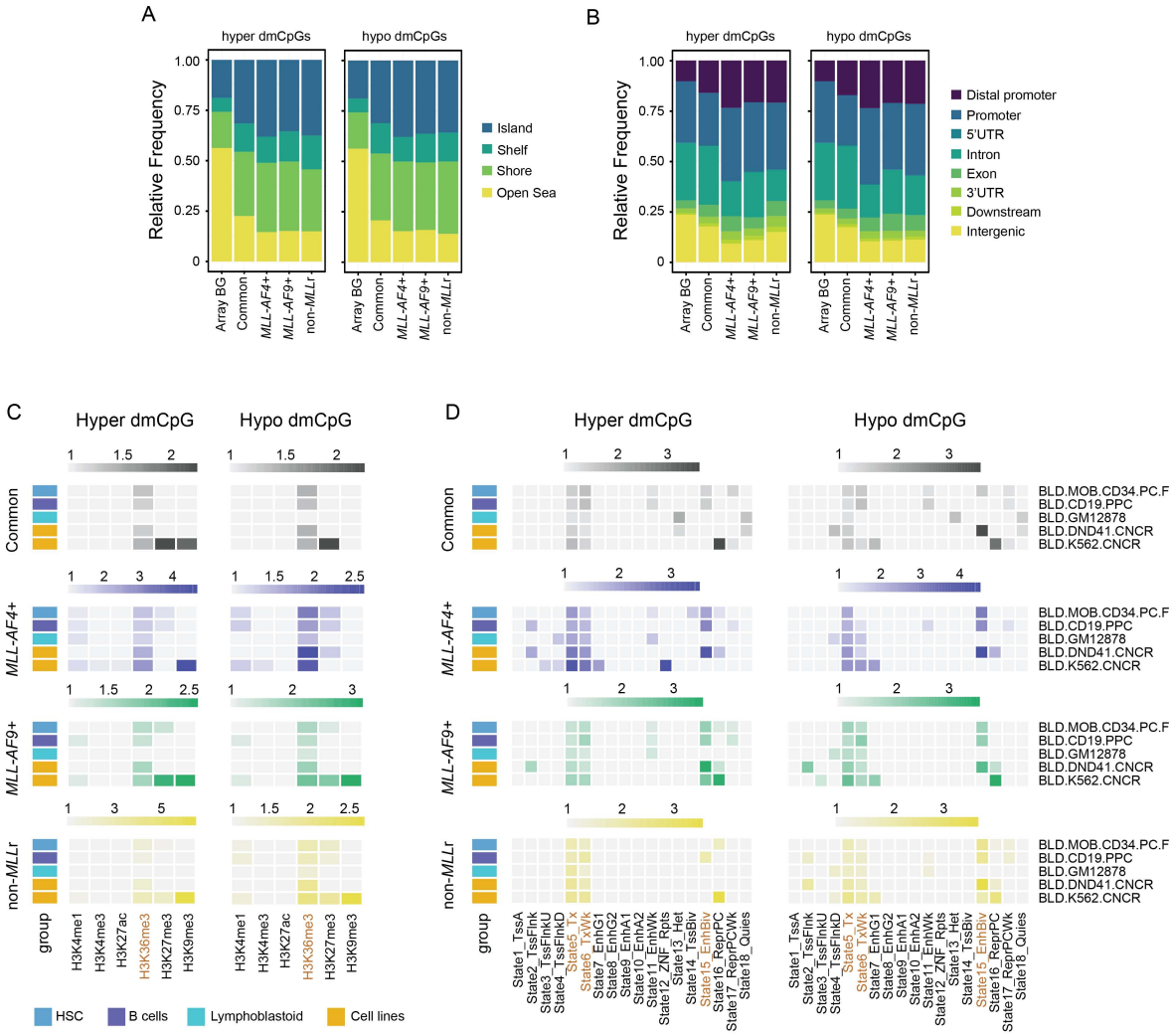

MLL−AF4+ 1 (4.23)
n = 186492
pearson cor = 0.90
MLL−AF4+ 2 (4.28)
n = 187211
pearson cor = 0.91
MLL−AF9+ 1 (9.11)
n = 139251
pearson cor = 0.89
MLL−AF9+ 2 (9.12)
n = 251734
pearson cor = 0.91
non−MLLr 1 (M.16)
n = 147599
pearson cor = 0.74
non−MLLr 2 (M.21)
n = 196056
pearson cor = 0.92
BCP 1 (Pool C1.3.5)
n = 225214
pearson cor = 0.92
BCP 2 (Pool 2.3.4)
n = 198845
pearson cor = 0.92
β-value WGBS
β-value EPIC

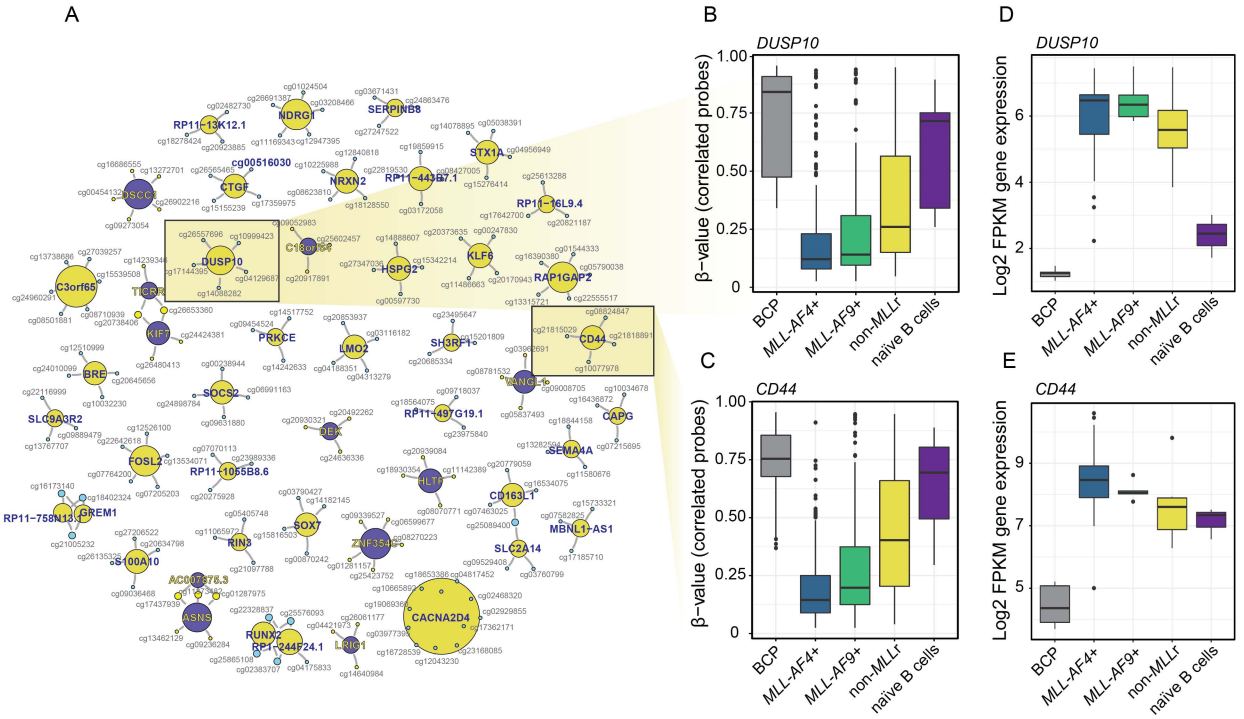
B

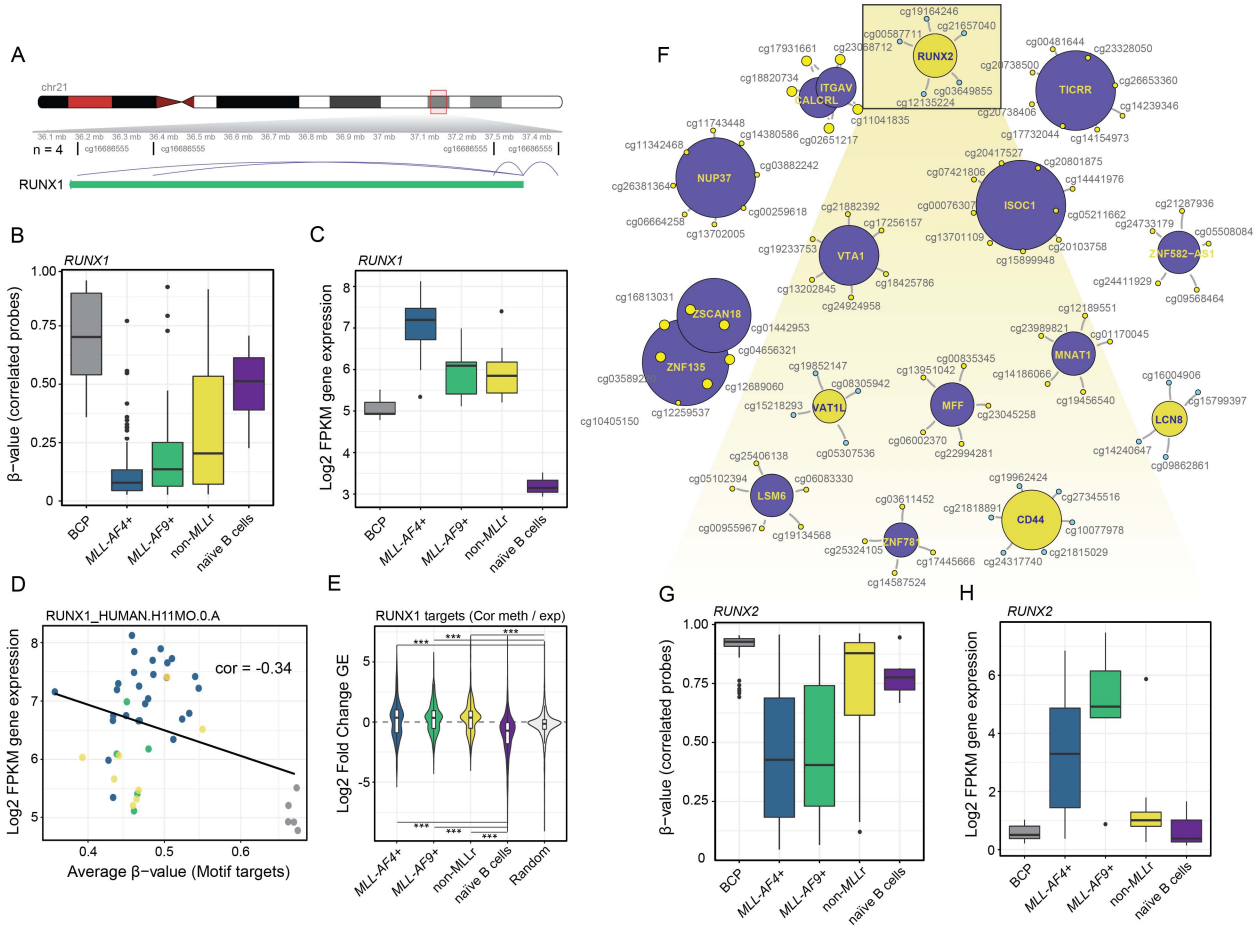

CpG sites (EPIC)
n = 781,459
10 Kbp windows (WGBS)
n = 284,513
β-value
MLL-AF4+ 1
MLL-AF4+ 2
MLL-AF9+ 1
MLL-AF9+ 2
non-MLLr 1
non-MLLr 2
BCP-1
BCP-2

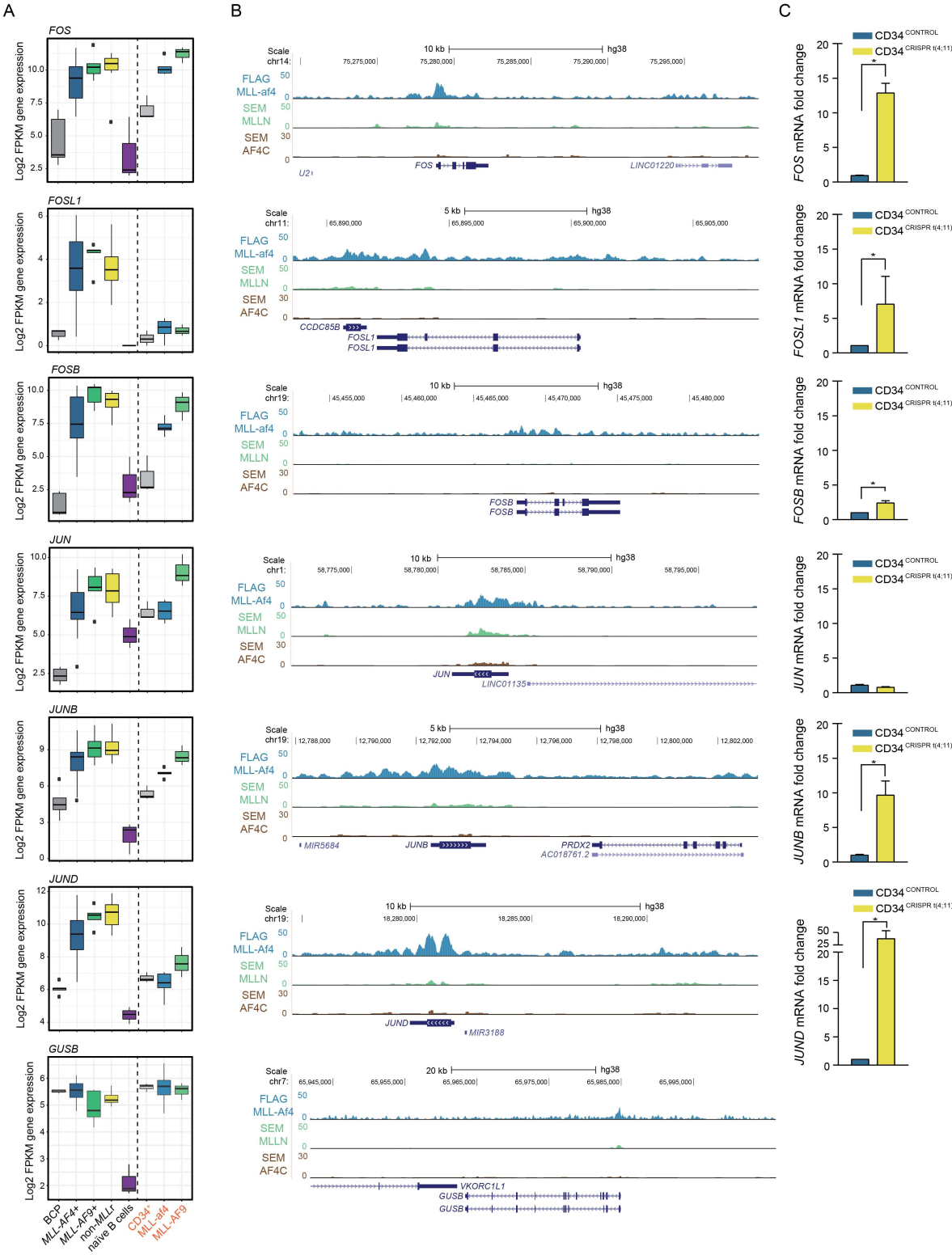

A
B cell enhancer regions
from enhancer Atlas database
3,427
361
34,796
dmCpGs naïve B cells vs BCPs
dmCpGs probes (EPIC)
proximal to B cell enhancers
(max distance 600 nt)
B
β-value
0.8
0.6
0.4
0.2
BCP
MLL-AF4+
MLL-AF9+
non-MLLr
naïve B cells
C
BCPs
MLL−AF4+
MLL−AF9+
non-MLLr
β-value naïve B cells
β-value BCPs
β-value MLL-AF4+
β-value MLL-AF9+
β-value non-MLLr
pearson cor = 0.42
pearson cor = 0.54
pearson cor = 0.55
pearson cor = 0.61


A
M1
Log2 FPKM RNAseq expression
M2
Log2 FPKM RNAseq expression
M3
Log2 FPKM RNAseq expression
B_cell_01
B_cell_02
B_cell_03
MA4_01
MA4_02
MA4_03
MA4_04
MA4_05
MA4_25
MA4_06
MA4_07
MA4_08
MA4_09
MA4_10
MA4_11
MA4_12
MA4_13
MA4_14
MA4_15
MA4_16
MA4_17
MA4_27
MA4_18
MA4_19
MA4_20
MA4_21
MA4_22
MA4_23
MA4_24
MA4_26
MA9_01
MA9_02
MA9_03
MA9_04
MA9_05
MLLwt_01
MLLwt_02
MLLwt_03
MLLwt_04
MLLwt_05
MLLwt_06
MLLwt_07
MLLwt_10
BCP_1
BCP_2
BCP_3
BCP_4
BCP_5
B cells
MLL-AF4
MLL-AF9
non-MLLr
BCP
B
M1
GO REGULATION OF CELL ACTIVATION
GO POSITIVE REGULATION OF CELL ACTIVATION
GO IMMUNOGLOBULIN RECEPTOR BINDING
GO POSITIVE REGULATION OF CYTOKINE PRODUCTION
GO PHAGOCYTOSIS ENGULFMENT
GO REGULATION OF CELL CELL ADHESION
GO MEMBRANE INVAGINATION
GO POSITIVE REGULATION OF CELL ADHESION
GO DNA PACKAGING COMPLEX
GO PHAGOCYTOSIS RECOGNITION
−log10(adjusted p−value)
M2
GO RESPONSE TO MOLECULE OF BACTERIAL ORIGIN
GO RESPONSE TO RADIATION
GO NEGATIVE REGULATION OF PHOSPHORYLATION
GO CELLULAR RESPONSE TO EXTERNAL STIMULUS
GO NEGATIVE REGULATION OF CYTOKINE PRODUCTION
GO INFLAMMATORY RESPONSE
GO CELLULAR RESPONSE TO EXTRACELLULAR STIMULUS
GO CELLULAR RESPONSE TO LIPID
GO RESPONSE TO PURINE CONTAINING COMPOUND
GO RESPONSE TO LIGHT STIMULUS
−log10(adjusted p−value)
M3
GO ORGANELLE FISSION
GO MITOTIC NUCLEAR DIVISION
GO CELL DIVISION
GO CHROMOSOME SEGREGATION
GO SISTER CHROMATID SEGREGATION
GO NUCLEAR CHROMOSOME SEGREGATION
GO NEGATIVE REGULATION OF CELL CYCLE PROCESS
GO MITOTIC CELL CYCLE CHECKPOINT
GO REGULATION OF MITOTIC CELL CYCLE
GO CELL CYCLE CHECKPOINT
−log10(adjusted p−value)
C
M1
HALLMARK APOPTOSIS
HALLMARK UNFOLDED PROTEIN RESPONSE
HALLMARK IL6 JAK STAT3 SIGNALING
HALLMARK ALLOGRAFT REJECTION
HALLMARK P53 PATHWAY
HALLMARK MITOTIC SPINDLE
HALLMARK INFLAMMATORY RESPONSE
HALLMARK HYPOXIA
HALLMARK TGF BETA SIGNALING
HALLMARK EPITHELIAL MESENCHYMAL TRANSITION
−log10(adjusted p−value)
M2
HALLMARK TNFA SIGNALING VIA NFKB
HALLMARK P53 PATHWAY
HALLMARK IL2 STAT5 SIGNALING
HALLMARK HYPOXIA
HALLMARK INFLAMMATORY RESPONSE
HALLMARK APOPTOSIS
HALLMARK CHOLESTEROL HOMEOSTASIS
HALLMARK INTERFERON GAMMA RESPONSE
HALLMARK UV RESPONSE UP
HALLMARK IL6 JAK STAT3 SIGNALING
−log10(adjusted p−value)
M3
HALLMARK G2M CHECKPOINT
HALLMARK E2F TARGETS
HALLMARK MITOTIC SPINDLE
HALLMARK MTORC1 SIGNALING
HALLMARK MYC TARGETS V2
HALLMARK SPERMATOGENESIS
HALLMARK MYC TARGETS V1
HALLMARK IL2 STAT5 SIGNALING
HALLMARK ALLOGRAFT REJECTION
HALLMARK NOTCH SIGNALING
−log10(adjusted p−value)
D
M1
LGALS3
SGTA
LRCH4
FLNA
H2AFX
MRPL28
NBEAL2
PDLIM7
PPP1R12C
TCF4
SF3A2
HMG20B
TRIP6
CCDC85B
Degree
100
200
E
M2
PPP1R15A
CSRNP1
NR4A1
RARA
MYO1C
GABARAPL1
CDKN1A
ATXN1
ICAM1
JUND
JUNB
TRAF4
FOS
JUN
APP
Degree
500
1,000
1,500
2,000
F
M3
NUSAP1
CDCA3
BUB1
PLK1
BUB1B
ZWINT
RACGAP1
CCNA2
MAGED1
CIT
PCNA
GTSE1
KIF11
Degree
50
100
150
200
250


A
hyper dmCpGs
hypo dmCpGs
Relative Frequency
1.00
0.75
0.50
0.25
0
Array BG
Common
MLL-AF4+
MLL-AF9+
non-MLLr
Island
Shelf
Shore
Open Sea
B
Distal promoter
Promoter
5'UTR
Intron
Exon
3'UTR
Downstream
Intergenic
C
Hyper dmCpG
Hypo dmCpG
Common
MLL-AF4+
MLL-AF9+
non-MLLr
group
H3K4me1
H3K4me3
H3K27ac
H3K36me3
H3K27me3
H3K9me3
HSC
B cells
Lymphoblastoid
Cell lines
D
BLD.MOB.CD34.PC.F
BLD.CD19.PPC
BLD.GM12878
BLD.DND41.CNCR
BLD.K562.CNCR
State1_TssA
State2_TssFlnk
State3_TssFlnkU
State4_TssFlnkD
State5_Tx
State6_TxWk
State7_EnhG1
State8_EnhG2
State9_EnhA1
State10_EnhA2
State11_EnhWk
State12_ZNF_Rpts
State13_Het
State14_TssBiv
State15_EnhBiv
State16_ReprPC
State17_ReprPCWk
State18_Quies


A
B
C
D
E
DUSP10
CD44
β−value (correlated probes)
Log2 FPKM gene expression
BCP
MLL-AF4+
MLL-AF9+
non-MLLr
naïve B cells


A
chr21
n = 4
RUNX1
B
RUNX1
β−value (correlated probes)
BCP
MLL-AF4+
MLL-AF9+
non-MLLr
naïve B cells
C
RUNX1
Log2 FPKM gene expression
D
RUNX1_HUMAN.H11MO.0.A
cor = -0.34
Log2 FPKM gene expression
Average β−value (Motif targets)
E
RUNX1 targets (Cor meth / exp)
Log2 Fold Change GE
Random
F
RUNX2
ITGAV
ALCRL
NUP37
ISOC1
TICRR
VTA1
ZSCAN18
ZNF135
MNAT1
VAT1L
MFF
LCN8
LSM6
CD44
ZNF78
G
RUNX2
β−value (correlated probes)
H
RUNX2
Log2 FPKM gene expression


A
B
C
FOS
FOSL1
FOSB
JUN
JUNB
JUND
GUSB
Log2 FPKM gene expression
BCP
MLL-AF4+
MLL-AF9+
non-MLLr
naive B cells
CD34+
MLL-af4
MLL-AF9
Scale
FLAG
MLL-af4
MLL-Af4
SEM
MLLN
AF4C
hg38
FOS mRNA fold change
FOSL1 mRNA fold change
FOSB mRNA fold change
JUN mRNA fold change
JUNB mRNA fold change
JUND mRNA fold change
CD34 CONTROL
CD34 CRISPR t(4;11)