Epigenetic regulator genes direct lineage switching in *MLL/AF4* leukaemia

Ricky Tirtakusuma[1]*, Katarzyna Szoltysek[1,2,3]*, Paul Milne[4], Vasily V Grinev[5], Anetta Ptasinska[6], Paulynn S Chin[6], Claus Meyer[7], Sirintra Nakjang[1], Jayne Y Hehir-Kwa[2], Daniel Williamson[1], Pierre Cauchy[6], Salam A Assi[6], Minoo Ashtiani[2], Peter Keane[6], Sophie G Kellaway[6], Maria R Imperato[6], Fotini Vogiatzi[9], Elizabeth K Schweighart[2], Shan Lin[10], Mark Wunderlich[10], Janine Stutterheim[2], Alexander Komkov[8], Elena Zerkalenkova[8], Paul Evans[11], Hesta McNeill[1], Alex Elder[1], Natalia Martinez-Soria[1], Sarah E Fordham[1], Yuzhe Shi[1], Lisa J Russell[1], Deepali Pal[1], Alex Smith[12], Zoya Kingsbury[13], Jennifer Becq[13], Cornelia Eckert[14], Oskar A Haas[15], Peter Carey[16], Simon Bailey[1,16], Roderick Skinner[1,16], Natalia Miakova[8], Matthew Collin[4], Venetia Bigley[4], Muzlifah Haniffa[17,18,19], Rolf Marschalek[7], Christine J Harrison[1], Catherine A Cargo[11], Denis Schewe[9], Yulia Olshanskaya[8], Michael J Thirman[20], Peter N Cockerill[6], James C Mulloy[10], Helen J Blair[1], Josef Vormoor[1,2], James M Allan[1], Constanze Bonifer[6*], Olaf Heidenreich[1,2*†], Simon Bomken[1,16*†]

Author Affiliations

[1]Wolfson Childhood Cancer Research Centre, Translational and Clinical Research Institute, Newcastle University, Newcastle upon Tyne, UK

[2]Princess Maxima Center for Pediatric Oncology, Utrecht, The Netherlands

[3]Maria Sklodowska-Curie Institute - Oncology Center, Gliwice Branch, Gliwice, Poland

[4]Translational and Clinical Research Institute, Newcastle University, Framlington Place, Newcastle upon Tyne, UK

[5]Department of Genetics, the Faculty of Biology, Belarusian State University, 220030 Minsk, Republic of Belarus.

[6]Institute of Cancer and Genomic Sciences, University of Birmingham, Birmingham, UK

[7]Institute of Pharmaceutical Biology/DCAL, Goethe-University, Frankfurt/Main, Germany

[8]Dmitry Rogachev National Research Center of Pediatric Hematology, Oncology, and Immunology; Moscow, Russia

[9]Pediatric Hematology/Oncology, ALL-BFM Study Group, Christian Albrechts University Kiel and University Hospital Schleswig-Holstein, Campus Kiel, Germany

[10]Experimental Hematology and Cancer Biology, Cancer and Blood Disease Institute, Cincinnati Children's Hospital Medical Center, Cincinnati, USA

[11]Haematological Malignancy Diagnostic Service, St James's University Hospital, Leeds, UK

[12]Epidemiology and Cancer Statistics Group, University of York, York, United Kingdom

[13]Illumina Cambridge Ltd., Great Abington, UK

[14]Department of Pediatric Oncology/Hematology, Charité Universitätsmedizin Berlin, Berlin, Germany

[15]Children's Cancer Research Institute, St. Anna Kinderkrebsforschung, Vienna, Austria

[16]Department of Paediatric Haematology and Oncology, The Great North Children's Hospital, Newcastle upon Tyne, UK

[17]Biosciences Institute, Newcastle University, Framlington Place, Newcastle upon Tyne, UK

[18]Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton UK

[19]Department of Dermatology and Newcastle NIHR Newcastle Biomedical Research Centre, Newcastle Hospitals NHS Foundation Trust, Newcastle upon Tyne

[20]Department of Medicine, Section of Hematology/Oncology, University of Chicago, Chicago, USA

***Authors contributed equally**
****Co-senior authors**
†Co-corresponding Authors

Running title: Lineage switching in *MLL/AF4* leukaemias

Keywords: *MLL/AF4*; *KMT2A-AFF1*; acute lymphoblastic leukaemia; nucleosome remodelling and deactylation complex (NuRD); chromatin remodelling

Corresponding authors

Dr Simon Bomken
Wolfson Childhood Cancer Research Centre
Translational and Clinical Research Institute
Level 6 Herschel Building
Brewery Lane
Newcastle University
Newcastle upon Tyne
NE1 7RU, UK

Tel: +44 (0)191 2082231
E mail: s.n.bomken@ncl.ac.uk

Professor Olaf Heidenreich
Princess Maxima Center for Pediatric Oncology
Heidelberglaan 25
3584 CS Utrecht
The Netherlands

Tel: +31 (0)88 972 7272
E mail: O.T.Heidenreich@prinsesmaximacentrum.nl

Supplements – Table of contents

Supplementary methods

Cell Lines

SEM, 697, REH and RS4;11 cells were grown in RPMI-1640 with 10% fetal bovine serum (FBS, Sigma) at 37°C in a humidified 5% CO_2 incubator. All cell lines contain no PHF3, CHD4 (with exception of the REH cell line - missense mutation at p.R61Q), PCGF6, AUTS2 mutation according to the CCLE (https://portals.broadinstitute.org/ccle) database.

MLL/Af4 transduced cord blood cells were cultured in IMDM with 10% FBS, supplemented with recombinant human SCF, IL-3, IL-6, FLT-3L, and TPO (10 ng/ml each) at 37°C in a humidified 5% CO_2 incubator [1]. To prime towards lymphoid differentiation, the cells were co-cultured with MS-5 (RRID:CVCL_2128, murine) stroma cells in α-MEM with 10% FBS, 2 mM glutamine supplemented with SCF, FLT-3L, and IL-7 (10 ng/ml each). The cells were demi-populated weekly.

Lymphoid differentiation of transduced MLL/Af4 cord blood cells

MLL/Af4 cord blood cells[1] were transduced with short hairpin constructs targeting CHD4, PHF3 or NTC control (as described in supplementary methods) and co-cultured with MS-5 stroma cells in lymphoid culture conditions. Single and triple transduced populations were identified using the construct specific fluorophores and lineage specific surface markers assessed as a proportion of the total transduced leukocyte population.

MLL/Af4 stem cell expression analysis

Following myeloid or lymphoid culture of *MLL/Af4* transduced CD34+ cord blood cells, CD19+CD33-, CD19-CD33+ and CD19-CD33- populations were flow sorted (based on the fluorescent protein expression), followed by RNA extraction using RNeasy Micro Kit (Qiagen). Input RNA was equilibrated to a starting input cell number of 300 cells per population before cDNA and sequencing library generation, performed using SMARTSeqv4 (Clontech) and NexteraXT (Illumina) kits. The resultant libraries were submitted for paired-end 150 bp sequencing on a NEXTSeq500 (Illumina). For each sample, transcript abundance was quantified from raw reads with Salmon (version 0.8.2) using the reference human transcriptome (hg38) defined by GENCODE release 27. An R package Tximport (version 1.4.0) was used to estimate gene-level abundance from Salmon's transcript-level counts. Gene-level differential expression analysis was performed using DESeq2 (version 1.16.1).

Flow cytometry and cell sorting

One million cells (of all tested cell lines, cord blood or LS03ALL PDX cells) were used for flow cytometry analysis of CD19 (BV421) and CD33 (APCCy7). The antibody details are listed in the Key Resources Table. Cells were collected and washed in a 3 ml wash buffer containing 0.2% BSA in PBS, re-suspended in 100 µl wash buffer and 5 µl each antibody was added, followed by incubation at RT for 20 min in the dark. Cells were washed and re-suspended in a 0.5 ml wash buffer. Flow cytometry was performed on FACSCanto II from Becton Dickinson (BD) and data analysed with FlowJo (Treestar).

Haematopoietic hierarchy analysis was performed from 10^7 cells of primary and PDX samples. The cells were collected and incubated for 30 min in a wash buffer containing 0.5% FBS, 2 mM EDTA in PBS. Following the incubation, the cells were washed and resuspended in a wash buffer. Fluorescence activated cell sorting (FACS) was performed using a FACS Aria Fusion (BD) or ASTRIOS EQ (Beckman Coulter). To separate early progenitors or lymphoid and myeloid fractions, MPAL patient samples were sorted from 10^7 or $3x10^6$ cells for MPAL1

and MPAL2 respectively. Sorted cells were collected into 1.5 ml microfuge tubes containing 500 µl RPMI with 10% fetal calf serum or SFEM II media in case of cell lines or primary samples, respectively. To analyse single cell derived clones, sorted early progenitor cells derived from MPAL samples were deposited onto 384-well plates containing 75µl of SFEM II media (supplemented with SCF, FLT3L, IL6, IL3 and TPO). Single cells were grown into 60-90% confluency colonies followed by RNA/DNA extraction and MLL/AF4 detection by PCR.

Single cell sorting and whole genome amplification (WGA)

Single cells were labelled (see flow cytometry and cell sorting section above) and sorted into 96 well plate (Eppendorf twin.tec PCR Plate 96 full skirt) using FACSAria Fusion (BD). The plates were filled with 3.5 µl PBS/well prior to the sorting. Each plate included forty-five single cell samples, two 0 cells as negative control, and one bulk (300 – 1,000 cells) as positive control sample. The cells were prepared from as many as 10^5 cells in 500 µl buffer (0.5% FBS, 2 mM EDTA in PBS).

Cell sorting was performed using the “single cell” setting on FACSAria Fusion. Following the sorting, the plate was centrifuged at 900 g for 1 min, snap-frozen on dry ice, and then kept at -20°C until the whole genome amplification (WGA) procedure.

WGA was performed using REPLI-g Single Cell Kit (Qiagen, Cat#150345). The cells were lysed by adding 1.5 µl Buffer D2 and incubation at 65°C for 10 min in a thermal cycler (HYBAID PCRExpress). Subsequently, the reaction was terminated by adding 1.5 µl Stop Solution and stored on the ice. The mixture of 20 µl REPLI-g sc Reaction Buffer and REPLI-g sc DNA Polymerase was added and incubated at 30°C for 8 h. The reaction was terminated by heat inactivation at 65°C for 3 min. The products were diluted 100-fold with TE Buffer ahead of PCR amplification.

DHS library generation, sequencing, and mapping

DHS analysis was performed as described previously [2]. DNase I (Worthington, Cat# LS006328) digestion was performed using ~5 million patient sample cells using 8 units (presentation) or 14 units (relapse) for 3 min at 22°C in a 1 mM $CaCl_2$ supplemented buffer. Nuclear proteins were digested with 1 mg/ml Proteinase K overnight at 37°C. DNase I digestion products were size-selected on an agarose gel, cutting below 150 bp. High-throughput sequencing libraries were prepared from 10 ng of size-selected material, using the Kapa Hyperprep kit. Libraries were sequenced with 50 bp single end reads on an Illumina HiSeq 2500.

Fastq files were generated using bcl2fastq (1.8.4), aligned to the hg19 assembly (NCBI Build 37) with the use of bowtie2 (2.1.0), with –very-sensitive-local as a parameter. Read coverage generation and peak detection were carried out using MACS 1.4.1 using --keep-dup=all –g hs -w -S. Pairwise comparisons were performed as previously described [2]. Digital footprinting was carried out using the Wellington package [3]. Differential footprinting analysis was carried out on footprints using the Wellington-bootstrap package with default parameters [4]. Average profiles and heatmaps were obtained using the functions dnase_average_profile and dnase_to_javatreeview from the Wellington package. Heatmaps were plotted using Java TreeView.

ATAC-seq

ATAC-seq was performed accordingly to Omni-ATAC protocol as previously described [5]. Briefly, $5x10^4$ cells were treated with Tagment DNA TDE1 Enzyme (Illumina) and subjected to library preparation procedure as described [6]. Libraries were sequenced by paired-end 50 bp sequencing on NovaSeq6000 (Illumina).

Chromatin immunoprecipitation

Briefly, 10^7 cells were double-crosslinked (Di(N-succinimidyl) glutarate for 45 minutes, followed by 1% formaldehyde for 10 minutes), lysed for 10 minutes in cell lysis buffer (10mM

HEPES pH 8.0, 10 mM EDTA, 0.5 mM MEGTA, 0.25% Triton-X 100, 1:1000 Protease Inhibitory Cocktail [PIC]) followed by nuclear lysis for 10 minutes in nuclear lysis buffer (10 mM HEPES pH 8.0, 1 mM EDTA, 200mM NaCl, 0.5 mM MEGTA, 0.01% Triton-X 100, 1:1000 PIC). Chromatin was then resuspended (25 mM Tris-HCl pH 8.0, 150 mM NaCl, 2 mM EDTA pH 8.0, 1% Triton-X 100, 0.25% SDS) and sonicated in a Bioruptor Pico (Diagenode) to generate 200-600 bp fragment size. Chromatin was resuspended (25 mM Tris pH 8.0, 150 mM NaCl, 2 mM EDTA pH 8.0, 1% Triton X-100, 7.5% glycerol and 1:100 PIC), diluted and incubated overnight at 4^0C with anti-CHD4 (AbCam) ChIP grade antibody-coupled Dynabeads G (ThremoFisher Scientific). Following incubation, beads were washed in sonication wash buffer I (50 mM Tris-HCl pH 8.0, 150 mM NaCl, 2 mM EDTA pH 8.0, 1% Triton-X 100, 0.1% SDS), twice in wash buffer II (20 mM Tris-HCl, pH 8.0, 2 mM EDTA pH 8.0, 500 mM NaCl, 0.1% SDS, 1% Triton X-100), once with Lithium buffer (10 mM Tris-HCl pH 8.0, 250 mM LiCl, 1 mM EDTA pH 8.0, 0.5% NP-40 and 0.5% sodium deoxycholate) and 1xTE. Chromatin was eluted in 100mM NaHCO3 and 1% SDS and reverse crosslinked for 4-16 hours at 65^0C with 100 µg Proteinase K. DNA was then cleaned up with 1.8x SPRI beads.

ChIP-seq libraries were prepared using the KAPA Hyperprep kit and sequenced by single-end 75 bp sequencing on NovaSeq6000 (Illumina).

ATAC-seq and ChIP-seq data analysis

Raw reads were processed to remove low-quality sequences using trimmomatic v0.39. Reads were then aligned to the human genome (version hg38) using bowtie2 v2.2.5 with the option --very-sensitive-local. Potential PCR duplicated reads were removed from alignments using the MarkDuplicates function in picard v2.26.10.

Peaks were called using macs2 v2.2.7.1 with default parameters or -f BAMPE --keep-dup all --nomodel -q 0.001 -B –trackline for ChIP-seq and ATAC-seq respectively. The resulting peaks were then filtered by height, whereby only peaks with a summit height of at least 60 in both replicates of an experimental condition were retained for further analysis. Peaks were then

further filtered against the hg38 blacklist to remove any potential artefacts from the data [7]. In order to define a single set of high-confidence peak positions which could be used to compare across different samples, we first combined all alignments into a single BAM file using the merge function in samtools v1.12 [8]. Peak calling was then repeated on this merged alignment. The resulting peak coordinates were then used as the reference peak positions for all further downstream analyses. ChIP-seq peaks were annotated to their closest, potential target genes using the annotatePeaks.pl function in homer v4.9.1.

Differential peak accessibility analysis was carried out by first counting the number of aligned fragments per peak with featureCounts v2.0.1 using the paired-end option [9]. Read counts were normalized and processed using the DESeq2 v1.34.0 package in R v4.1.2. A peak was considered to be differentially accessible if it had a fold-difference of at least 2 between condition, and a Benjamin-Hochberg adjusted p-value < 0.05. De-Novo motif searches were carried out in the sets of gained and lost peaks using the findMotifsGenome.pl function in homer v4.9.1[10]. Read density plots were created by first ranking peaks according to fold-difference. The average read density in peaks was then calculated using the annotatePeaks.pl function in homer, with the options -size 2000 -hist 10 -ghist -bedGraph, using the bedGraph files created by macs2. The resulting matrix was then plotted a heatmap using Java TreeView v1.1.6r4.

Exome sequencing

Germline DNA from cases LS08 and LS09 were extracted from formalin fixed paraffin embedded remission bone marrow using QIAamp DNA FFPE Tissue Kit (Qiagen, Cat#56404). Other DNA samples were extracted from either bone marrow or peripheral blood using AllPrep DNA/RNA Mini Kit (Qiagen, Cat#80204), QIAamp DNA Mini Kit (Qiagen, Cat#51306), or innuPREP DNA/RNA Mini Kit (Analytik Jena, Cat#845-KS-2080050). The exons were captured using SureSelect XT2 Human All Exon V6 (Agilent) and sequenced by paired-end 75 bp sequencing on HiSeq4000 (Illumina). DNA from the myeloid and lymphoid

cellular compartments derived from MPAL patient samples, were pre-processed with KAPA HyperPlus Kit (Roche) followed by exons enrichment with KAPA HyperCapture Kit (Roche), and sequenced by paired-end 300 bp sequencing on NovaSeq6000 (Illumina).

Raw reads were aligned to human reference genome (hg19 or hg38 for lineage switch or MPAL patients, respectively) using Burrows-Wheeler Aligner (BWA) 0.7.12 and were processed using the Genome Analysis Toolkit (GATK, v3.8 or 4.1). MuTect (v1.1.7) and MuTect2 (4.1) were used to identify somatic variants for each matched sample pair. Variants were annotated using Ensembl Variant Effect Predictor (VEP, version 90).

RNA sequencing

Total RNA was extracted with AllPrep DNA/RNA Mini Kit (Qiagen, Cat#80204), innuPREP DNA/RNA Mini Kit (Analytik Jena, Cat#845-KS-2080050), or TRIzol (Thermo Fisher Scientific, Cat# 15596026) followed by RNeasy Mini Kit (Qiagen, Cat#74106) from either bone marrow or peripheral blood. Messenger RNA was captured using NEBNext Ultra Directional RNA Kit in combination with NEBNext poly(A) mRNA Magnetic Isolation Module or KAPA RNA HyperPrep Kit with RiboErase (HMR) in case of lineage switch or MPAL patients respectively, and submitted for paired-end 150 bp sequencing on HiSeq4000 (Illumina) or paired-end 300 bp sequencing on NovaSeq6000 (Illumina) depending on the analysed patients group. For each sample, transcript abundance was quantified from raw reads with Salmon (version 0.8.2) using the reference human transcriptome (hg38) defined by GENCODE release 27. An R package Tximport (version 1.4.0) was used to estimate gene-level abundance from Salmon's transcript-level counts. Gene-level differential expression analysis was performed using DESeq2 (version 1.16.1). Differential splicing events were identified in both presentation/relapse pairs or lymphoid/myeloid cellular fractions, using the pipelines described previously [11].

For the fusion gene identification, RNA-seq data has been analysed as previously described [12]. Briefly, data was processed as per the GATK 4.0 best practices workflow for variant calling,

using a wdl and cromwell based workflow (https://gatk.broadinstitute.org/hc/enus/sections/360007226651-Best-Practices-Workflows). This included performing quality control with Fastqc (version 0.11.5) to calculate the number of sequencing reads and the insert size. Picard (version 2.20.1) for RNA metrics output and MarkDuplicates. The raw sequencing reads were aligned using Star (version 2.7.0f) to GRCh38 and gencode version 29. Gene fusion detection was performed using Star fusion (version 1.6.0) based on accuracy comparisons previously performed[13,14]. Finally, expression counts were determined at exon and gene level using Subread Counts[15].

Nested multiplex PCR and targeted sequencing

The 12 mutation candidate driver genes and *MLL/AF4* in LS01RAML were amplified by nested multiplex PCR method using gDNA as template.

PCR amplifications were carried out using Phusion® High-Fidelity PCR Master Mix with HF Buffer (NEB, Cat#M0531L) with 80 – 200 nM of each primer set (listed in the supplements). Products of the first multiplex reaction (98°C/2 min, thirty cycles of 98°C/10 s, 63°C/30 s, and 72°C/30 s, followed by 72°C/10 min) were diluted 500-fold, and 1 µl used as the template for the second PCR reactions (98°C/2 min, twenty cycles of 98°C/10 s, 65°C/30 s, and 72°C/30 s, followed by 72°C/ 10 min). The amplicons taken forward for next-generation targeted sequencing had additional CS1 (ACACTGACGACATGGTTCTACA) and CS2 (TACGGTAGCAGAGACTTGGTCT) Fluidigm tag sequences on the nested PCR primer forward and reverse, respectively. The amplicons were barcoded using Fluidigm Access Array Barcode Library for Illumina Sequencers (Cat#100-4876) by taking 0.8 µl multiplex PCR products (multiplex group A-C), 4 µl Fluidigm barcode primer (400 nM final concentration), 10 µl of 2X Phusion Master Mix (NEB, Cat#M0531L), and 3.6 µl H_2O (98°C/2 min, six cycles of 98°C/10 s, 60°C/30 s, and 72°C/1 min, followed by 72°C/10 min). Products were separated with 2% agarose gel and extracted using the QIAquick Gel Extraction Kit (Qiagen, Cat#28706).

The purified products were submitted for paired-end 300 bp sequencing on MiSeq (Illumina), resulting in >1,000 coverage per gene.

Long-distance inverse PCR experiments

MLL rearrangement sequences were identified using long-distance inverse PCR (LDI-PCR) as previously described [16,17].

PCR-based clonality immunoglobulin gene targets

Clonality analysis based on immunoglobulin gene rearrangements was performed according to the BIOMED-2 Concerted Action BMH-CT98-3936 using standardised protocols and primer sets [18]. Duplicate samples were analysed by GeneScan analysis on a Life Technologies 3100 platform.

Reverse Engineering of Transcriptional Networks

A mutual information network was developed such that the connections between genes within the network represented an inferred likelihood of a causal relationship within ALL/AML. The network was further refined by retaining only those nodes (genes) and edges (causal connections) which are significantly associated, in expression terms, to the distinction between normal myeloid and lymphoid cells. Thus, the centrality of the mutations within the network reflects the estimated extent of their causal influence upon a lymphoid/myeloid distinction within the framework of primary ALL/AML.

Mutual Information networks were reverse engineered from 216 published AML/ALL Affymetrix HGU133p2 expression profiles (GSE11877 & GSE24006) processed with the Rma (affy package R/Bioconductor) using the ARACNe2 algorithm [19] set to adaptive partitioning with a DPI tolerance of 0.1 and a p-value threshold of 0.01 and 1000 bootstraps. The probes/genes were filtered to remove those that were insufficiently variable as per the Aracne method. Probes had to be greater than 3-fold expression and 300 delta between the min and max excluding the 5% most extreme values to be included. Probes without >20 intensity in

greater than 20% of samples were also removed. All probes which fulfilled these requirements were included in the construction of this network (n=9780).

Nodes within the network were annotated with a metagene value calculated using Non-Negative Matrix Factorisation (k=2) and reflecting the differences between myeloid and lymphoid cells (GSE24759) and the other. A metagene cutoff of 0.13 (based on a calculated alpha of 0.1) was applied to trim the network. Cytoscape [20] was used to merge nodes where probe sets represented common genes and to calculate centrality statistics for each of the mutated genes of interest. Nodes of interest (i.e. mutated nodes) were ranked according to their centrality (e.g. degree). Nodes were sized and coloured according to their differential expression between AML and ALL types; size = significance, colour = log2 fold change.

PHF3, CHD4, PCGF6, and AUTS2 shRNAs

shRNAs against PHF3 (TRCN0000019118, TRCN0000019114, TRCN0000274376), CHD4 (TRCN0000380981, TRCN0000021363, TRCN0000021360), PCGF6 (TRCN0000229804, TRCN0000073109, TRCN0000073109), AUTS2 (TRCN0000119058, TRCN0000304019, TRCN0000304081), or non-targeting control (ATCTCGCTTGGGCGAGAGTAAG) were cloned into pLKO5d.SFFV.miR30n [21]. Each target gene was linked with different fluorescent protein, including dTomato (PHF3 and AUTS2), eGFP (CHD4 and PCGF6), and RFP657 (non-targeting control).

The vector contained BsmBI at the cloning sites. The oligonucleotides were designed with the appropriate complementary overhang sequences of BsmBI-cleaved vector. They were ligated and transformed into STBL3 chemically competent cells (Invitrogen) according to manufacturer's instructions and plated on agar plates containing 100 µg/ml ampicillin incubated for 16 h at 37°C. Single colonies were inoculated for DNA preparation, and extracted

using EndoFree Plasmid Maxi Kit (Qiagen, Cat#12362). All clones were verified by sequencing prior to lentivirus production.

Lentivirus Production and Transduction

Viral particles were produced using calcium phosphate precipitation method on 293T cells. The cells were grown in 100 mm tissue culture dishes at a concentration of 1-2 x 10^6 cells in 10 ml medium the day prior to co-transfection. Equimolar amounts of envelope plasmid pMD2.G, packaging plasmid pCMVΔR8.91, and the shRNA vector pLKO5d.SFFV.miR30n were mixed and the volume adjusted with HEPES buffer solution (2.5 mM HEPES containing deionized water at pH 7.3) to 250 µl. A volume of 250 µl of 0.5 M $CaCl_2$ was added to the mixture. This solution was added to 500 µl of 2X HeBS (0.28 M NaCl, 0.05 M HEPES and 1.5 mM Na_2HPO_4 in deionized water at pH 7.00) and mixed by vortexing. It was incubated at RT for 30-40 min to allow the formation of the calcium phosphate precipitate, before adding dropwise on the 293T cells. After 16 h, the cells were washed with 10 ml PBS and added with 10 ml culture media. They were incubated at 37°C in a humidified atmosphere with 5% CO_2 for the next two days. Lentivirus particles were collected by centrifuging the supernatant at 400 g for 10 min at 4°C and filtered through Acrodisc Syringe 0.45 µm filters. They were stored in aliquots at -80°C. Cell lines and PDX samples were transduced with lentivirus as previously described [22].

qRT-PCR

One million cells were collected, and RNA was extracted using RNeasy Mini Kit (Qiagen, Cat#74106). cDNA was synthesised from 1 µg RNA using RevertAid H Minus First Strand cDNA Synthesis Kit (Thermo Fisher Scientific, Cat#K1632). The product was diluted by adding 80 µl H_2O [23]. Sequences of primers used in this study were designed using Primer Express (Applied Biosystems) software.

List of antibodies and oligonucleotides used in the study.

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
CD3-APC-H7	BD Biosciences	Cat#641397; RRID:AB_1645731
CD3-FITC	BD Biosciences	Cat#345763
CD3-BV650	Biolegend	Cat#300468
CD10-BV650	BD Biosciences	Cat#563734; RRID:AB_2738393
CD11b-APC/Fire750	Biolegend	Cat#301351
CD14-FITC	BD Biosciences	Cat#555397; RRID:AB_395798
CD14-BV605	Biolegend	Cat#367125
CD16-FITC	BD Biosciences	Cat#335035
CD19-PE-CF594	BD Biosciences	Cat#562294; RRID:AB_11154408
CD19-APCCy7	Biolegend	Cat#363010; RRID:AB_2564193
CD19-BV421	Biolegend	Cat#302234; RRID:AB_11142678
CD19-PECy7	Biolegend	Cat#363011
CD20-PE	BD Biosciences	Cat#345793
CD20-FITC	Biolegend	Cat#302303
CD33-APC	BD Biosciences	Cat#345800
CD33-APCCy7	Biolegend	Cat#366614; RRID:AB_2566416
CD33-BV421	Biolegend	Cat#303416; RRID:AB_2561690
CD34-PerCPCy5.5	BD Biosciences	Cat#347222
CD34-APCCy7	Biolegend	Cat#343514; RRID:AB_1877168
CD34-APC	Biolegend	Cat#343607
CD38-PeCy7	BD Biosciences	Cat#335825
CD38-PE	Biolegend	Cat#303506
CD45RA-BV510	Biolegend	Cat#304142; RRID:AB_2561947
CD56-FITC	BD Biosciences	Cat#345811
CD64-BV711	Biolegend	Cat#305041
CD90-A700	Biolegend	Cat#328120; RRID:AB_2203302
CD90-PerCPCy5.5	Biolegend	Cat#328118; RRID:AB_2303335
CD90-BV421	Biolegend	Cat#328122
CD117-PE	BD Biosciences	Cat#332785
CD117-BV605	BD Biosciences	Cat#562687; RRID:AB_2737721
CD123-BV421	BD Biosciences	Cat#306018; RRID:AB_10962571

HLA-DR-BV786	Biolegend	Cat#307642; RRID:AB_2563461
HLA-DR-A700	Biolegend	Cat#560743; RRID:AB_1727526
HLA-DR-PerCPCy5.5	Biolegend	Cat#307629
CHD4 [3F2/4] ChIP grade	AbCam	Cat#Ab264521
Oligonucleotides		
shPHF3-1	GPP, Broad Institute	TRCN0000019118
shPHF3-2	GPP, Broad Institute	TRCN0000019114
shPHF3-3	GPP, Broad Institute	TRCN0000274376
shCHD4-1	GPP, Broad Institute	TRCN0000380981
shCHD4-2	GPP, Broad Institute	TRCN0000021363
shCHD4-3	GPP, Broad Institute	TRCN0000021360
shPCGF6-1	GPP, Broad Institute	TRCN0000229804
shPCGF6-2	GPP, Broad Institute	TRCN0000073109
shPCGF6-3	GPP, Broad Institute	TRCN0000073109
shAUTS2-1	GPP, Broad Institute	TRCN0000119058
shAUTS2-2	GPP, Broad Institute	TRCN0000304019
shAUTS2-3	GPP, Broad Institute	TRCN0000304081
shNTC (ATCTCGCTTGGGCGAGAGTAAG)	(16)	N/A
shNCOA2 Fw (agcgcATCCGTTCTCAGACTACTAATtagtgaagccacagatgtaATTAGTAGTCTGAGAACGGATt)	This paper	N/A
shNCOA2 Rev (ggcaaATCCGTTCTCAGACTACTAATtacatctgtggcttcactaATTAGTAGTCTGAGAACGGATg)	This paper	N/A
shDHX36 Fw (agcgcTCCGCTGAGTGGGTTAGTAAAtagtgaagccacagatgtaTTTACTAACCCACTCAGCGGAt)	This paper	N/A
shDHX36 Rev (ggcaaTCCGCTGAGTGGGTTAGTAAAtacatctgtggcttcactaTTTACTAACCCACTCAGCGGAg)	This paper	N/A
shCEP164 Fw (agcgcGCAGTGAAAGTTCTGAATCTTtagtgaagccacagatgtaAAGATTCAGAACTTTCACTGCt)	This paper	N/A
shCEP164 Rev (ggcaaGCAGTGAAAGTTCTGAATCTTtacatctgtggcttcactaAAGATTCAGAACTTTCACTGCg)	This paper	N/A
shPPP1R7 Fw (agcgcGCAACTTACATCAACTACAGAtagtgaagccacagatgtaTCTGTAGTTGATGTAAGTTGCt)	This paper	N/A

shACAP1 Fw (agcgcGCAGGAGATGAGACGTATCTTtagtgaagccacagatgtaAAGATACGTCTCATCTCCTGCt)	This paper	N/A
shACAP1 Rev (ggcaaGCAGGAGATGAGACGTATCTTtacatctgtggcttcactaAAGATACGTCTCATCTCCTGCg)	This paper	N/A
PHF3 Fw (ATGGACCTGGGCTTGAACTG)	This paper	N/A
PHF3 Rev (TGGTGGTGCACTTTCAGGAG)	This paper	N/A
CHD4 Fw (TGCTGACACAGTTATTATCTATGACTCTGA)	This paper	N/A
CHD4 Rev (ACGCACGGGTCACAAACC)	This paper	N/A
PCGF6 Fw (GGGAAATCCGACGTGCAAT)	This paper	N/A
PCGF6 Rev (GGAGAAACCACAAGACCATAATGA)	This paper	N/A
AUTS2 Fw (AAAAGGACCCGAGGTTGACA)	This paper	N/A
AUTS2 Rev (GCGATGTGAACATGCATAGCA)	This paper	N/A
GAPDH Fw (GAAGGTGAAGGTCGGAGTC)	This paper	N/A
GAPDH Rev (GAAGATGGTGATGGGATTTC)	This paper	N/A
CD19 Fw (TGACCCCACCAGGAGATTCTT)	This paper	N/A
CD19 Rev (CACGTTCCCGTACTGGTTCTG)	This paper	N/A
CD33 Fw (CTCGTGCCCTGCACTTTCTT)	This paper	N/A
CD33 Rev (CCCGGAACCAGTAACCATGA)	This paper	N/A
CSF3R Fw (CCCAGGCGATCTGCATACTT)	This paper	N/A
CSF3R Rev (AACAAGCACAAAAGGCCATTG)	This paper	N/A
KIT Fw (GGACCAGGAGGGCAAGTCA)	This paper	N/A
KIT Rev (GATAGCTTGCTTTGGACACAGACA)	This paper	N/A
NCOA2 fw (TGCGAATTTCACAGAGCACTTTT)	This paper	N/A
NCOA2 rev (GGAAAGGTCCAGCACCAGTT)	This paper	N/A
PPP1R7 fw (CAGGAGATGATGGAGGTTGACA)	This paper	N/A
PPP1R7 rev (CGATGCCACTGCTGTGTTTC)	This paper	N/A
CEP164 fw (GCCTGGACTTCGGTT)	This paper	N/A
CEP164 rev (TGTCTTCTATTCCCAGTGGTTGCT)	This paper	N/A
ACAP1 fw (CTTCGTTGTCGGCATTTGTG)	This paper	N/A

ACAP1 rev (GGCTCACGGTGAATTTTTCC)	This paper	N/A
DHX36 fw (ATGCCTACAGTTAACCAGACACA)	This paper	N/A
DHX36 rev (ATACAGATGATAGCAATGACCAGG)	This paper	N/A

Supplementary figures

Figure S1

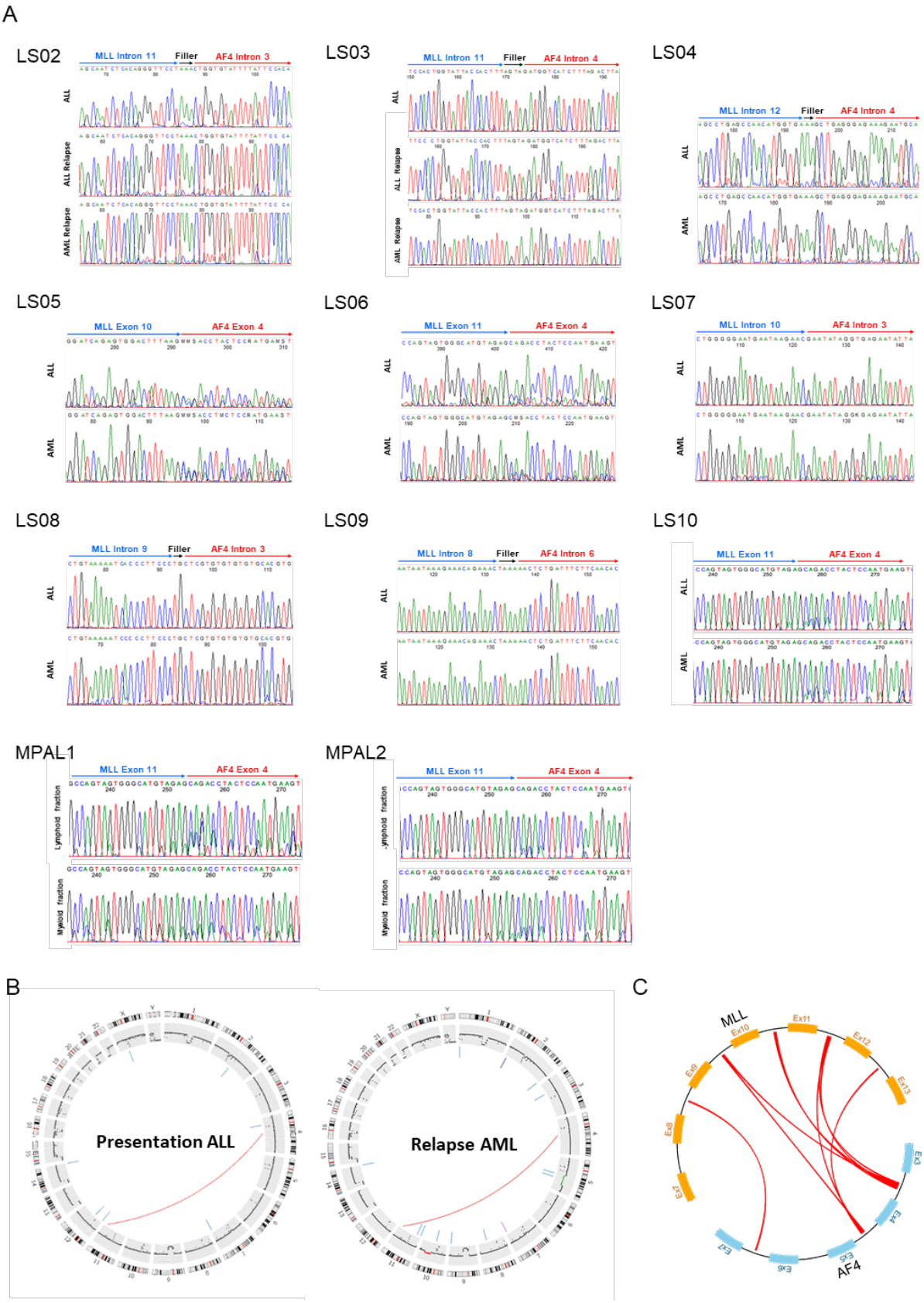

Figure S1. Genomic characterization of the *MLL/AF4* lineage switch cases. (A) Sequencing of *MLL/AF4* fusion breakpoints from DNA (LS02, LS03, LS04, LS07, LS08, LS09) or RNA (LS05, LS06, LS10, MPAL1, MPAL2). (B) Whole genome sequencing data of LS01 showing karyotype (outer circle), copy number changes (log2 depth ratio in 1Mb windows, loss <2 green dots, gain >2 red dots) and structural variants (translocations - red connecting lines, deletions – blue lines, inversions, purple lines). (C) Schematic representation of identified fusion variants, located within the major *MLL* breakpoint region, present in analysed t(4;11) cases (detailed breakpoint description presented also at Table S1). Red connecting line indicates MLL/AF4 translocation positions of each gene partner.

Figure S2

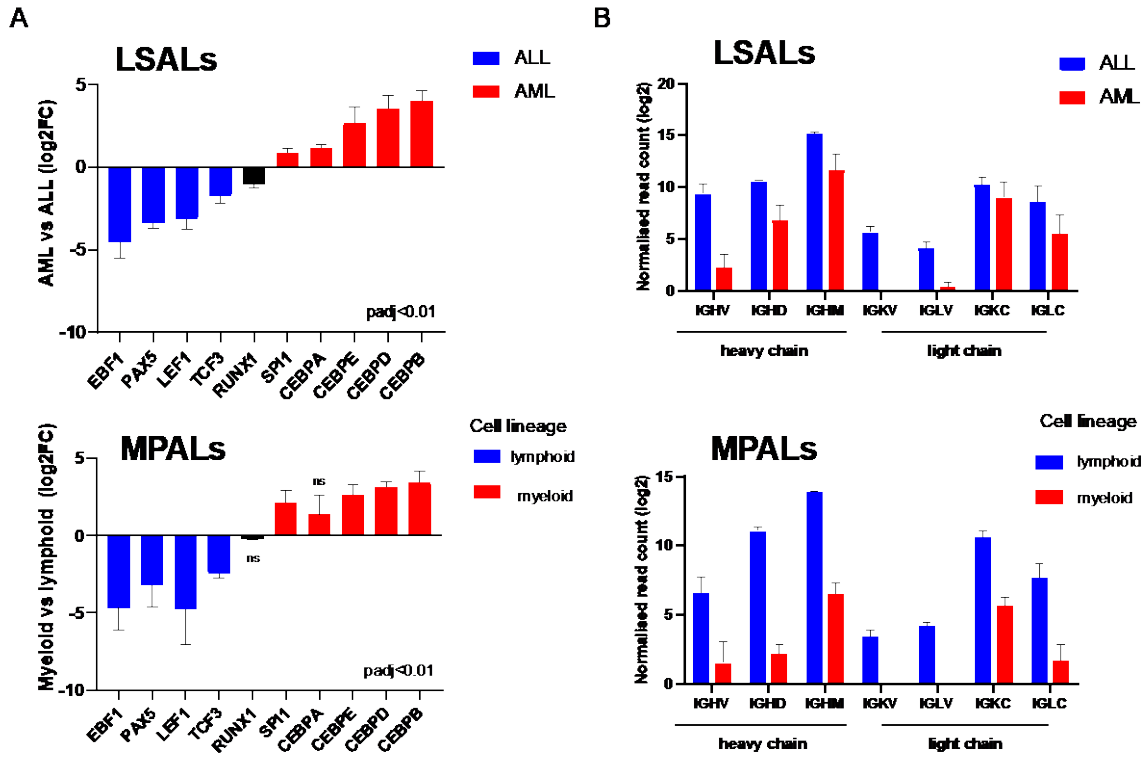

Figure S2. Transcriptional changes associated with lineage switching. (A) Differential expression of TFs regulating lymphoid (in blue) and myeloid phenotype (in red). RUNX1 in black reflects enriched accessibility of different RUNX1 binding sites in ALL and AML. (B) Differential expression of genes encoding immunoglobulin heavy and light chains in lineage switch and MPAL cases. Error bars show standard error of the mean (SEM) for lineage switch cases and ranges for two MPAL cases.

Figure S3

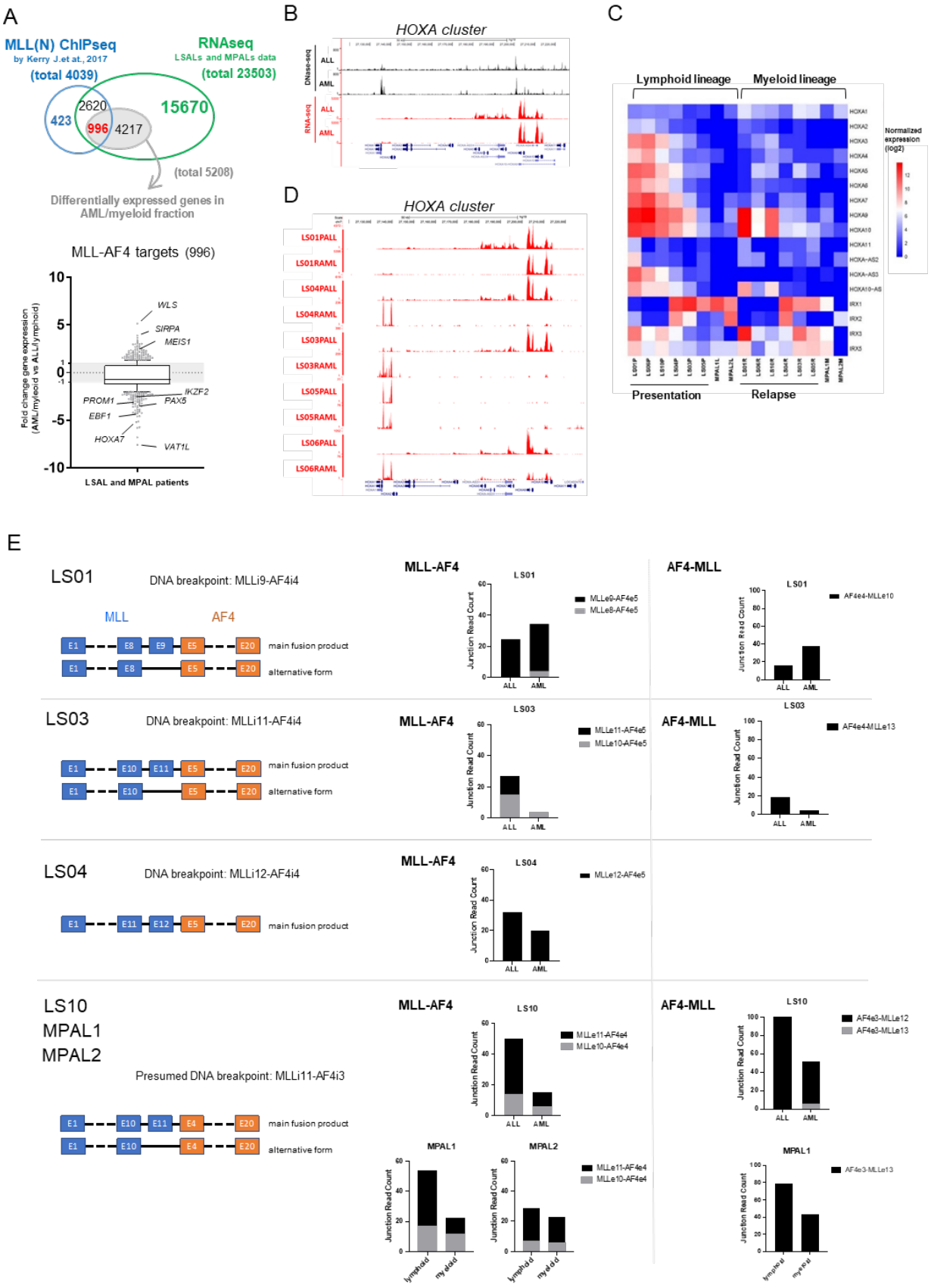
A
MLL(N) ChIPseq
by Kerry J.et at., 2017
(total 4039)
RNAseq
LSALs and MPALs data
(total 23503)
2620
15670
423
996
4217
(total 5208)
Differentially expressed genes in AML/myeloid fraction
MLL-AF4 targets (996)
Fold change gene expression (AML/myeloid vs ALL/lymphoid)
WLS
SIRPA
MEIS1
IKZF2
PAX5
PROM1
EBF1
HOXA7
VAT1L
LSAL and MPAL patients
B
HOXA cluster
DNase-seq
RNA-seq
ALL
AML
C
Lymphoid lineage
Myeloid lineage
Normalized expression (log2)
Presentation
Relapse
D
HOXA cluster
LS01PALL
LS01RAML
LS04PALL
LS04RAML
LS03PALL
LS03RAML
LS05PALL
LS05RAML
LS06PALL
LS06RAML
E
LS01
DNA breakpoint: MLLi9-AF4i4
MLL
AF4
main fusion product
alternative form
MLL-AF4
AF4-MLL
Junction Read Count
MLLe9-AF4e5
MLLe8-AF4e5
AF4e4-MLLe10
LS03
DNA breakpoint: MLLi11-AF4i4
MLLe11-AF4e5
MLLe10-AF4e5
AF4e4-MLLe13
LS04
DNA breakpoint: MLLi12-AF4i4
MLLe12-AF4e5
LS10
MPAL1
MPAL2
Presumed DNA breakpoint: MLLi11-AF4i3
MLLe11-AF4e4
MLLe10-AF4e4
AF4e3-MLLe12
AF4e3-MLLe13
lymphoid
myeloid

Figure S3. Impact of leukaemia lineage switch on *MLL/AF4*-regulated genes. (A) Venn diagram showing the number of direct *MLL/AF4* target genes affected by lineage switch. The diagram overlays ChIPseq data from SEM cells and t(4;11) patient cells (Kerry *et al.*, 2017) and differentially expressed genes in our lineage switch and MPAL cases (Venn diagram, upper panel). Highly enriched genes are further listed on the lower panel. (B) Differential chromatin accessibility and expression across the HOXA cluster at ALL presentation and AML relapse of case LS01, demonstrating differential DNase hypersensitivity of HOXA cluster. (C) HOXA cluster analysis includes a general reduction on HOXA3-10 in relapse AML cases. (D) Differential expression across the HOXA cluster in five cases of lineage switch from ALL (upper panel for each case) and AML (lower panel). (E) mRNA expression level of *MLL-AF4* (and *AF4-MLL*) fusion variants identified in presentation ALL and relapse AML samples of patients LS01, LS03, LS04 and LS10, and in lymphoid and myeloid fraction of both MPAL patients. mRNA expression level, presented as junction read counts, has been identified based on RNAseq data, with application of STAR-fusion 1.8.0 and ENSMBL Variant Effect Predictor v92.3 pipelines.

Figure S4

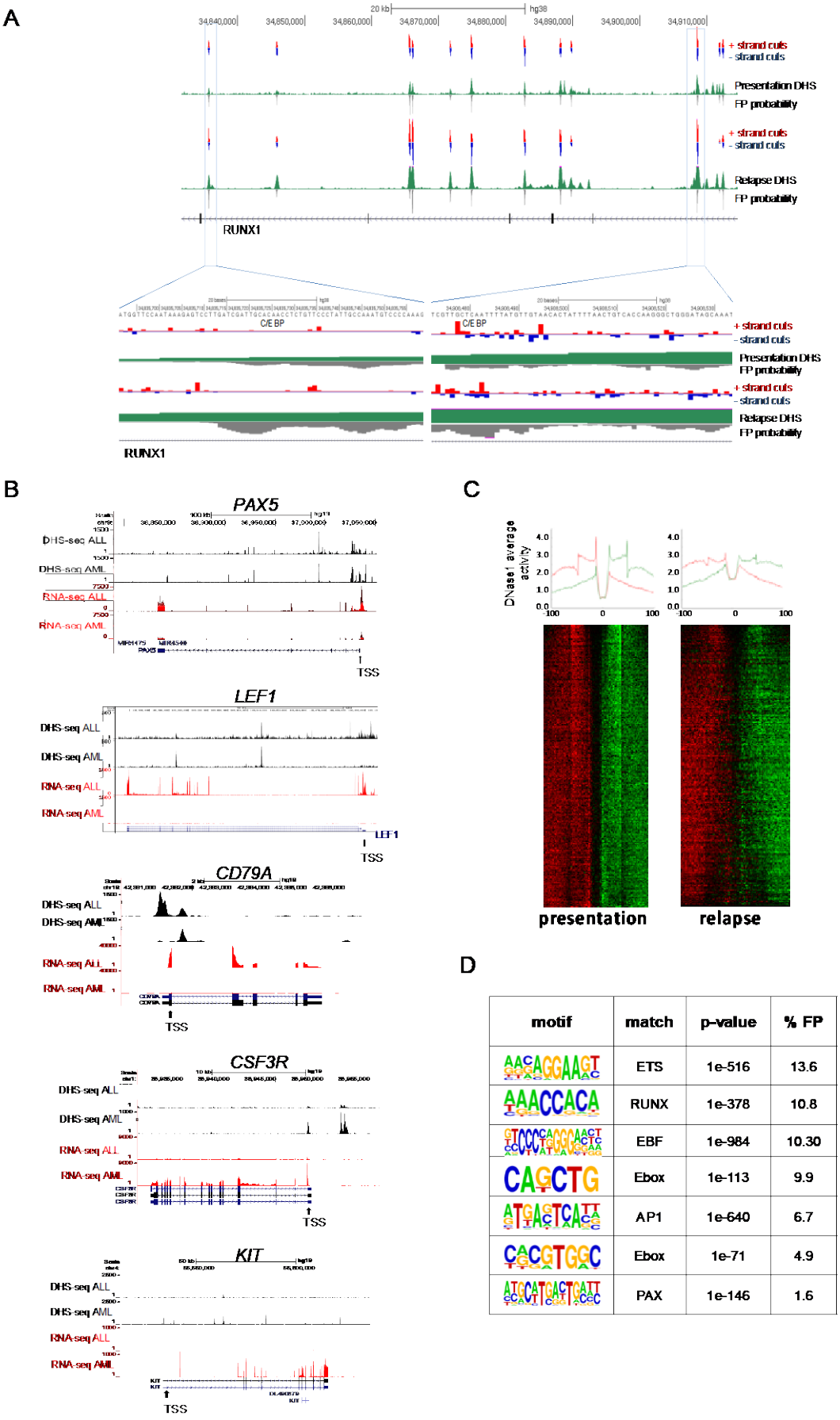
A
+ strand cuts
- strand cuts
Presentation DHS
FP probability
Relapse DHS
RUNX1
C/E BP
B
PAX5
LEF1
CD79A
CSF3R
KIT
DHS-seq ALL
DHS-seq AML
RNA-seq ALL
RNA-seq AML
TSS
C
DNase I average activity
presentation
relapse
D
motif
match
p-value
% FP
ETS
1e-516
13.6
RUNX
1e-378
10.8
EBF
1e-984
10.30
Ebox
1e-113
9.9
AP1
1e-640
6.7
Ebox
1e-71
4.9
PAX
1e-146
1.6

Figure S4. High resolution DNase hypersensitivity sequencing. (A) UCSC genome browser screenshot for *RUNX1* focused on an AML-associated DHS with C/EBP occupation as indicated by high resolution DHS-seq and Wellington analysis. FP - footprint. (B) Differential promoter accessibility is associated with higher expression of *PAX5*, *LEF1*, and *CD79A* in presentation ALL cases, and of *CSF3R* and *KIT* in relapse AML cases. TSS – transcriptional start site. (C) Heat maps showing proximal DHS regions specific for ALL presentation on a genomic scale. Red and green indicate excess of positive and negative strand cuts, respectively, per nucleotide position. Sites are sorted from top to bottom in order of decreasing Footprint Occupancy Score. (D) *De novo* motif discovery in proximal DHSs unique to ALL as compared to AML relapse, as shown on the table, right panel.

Figure S5

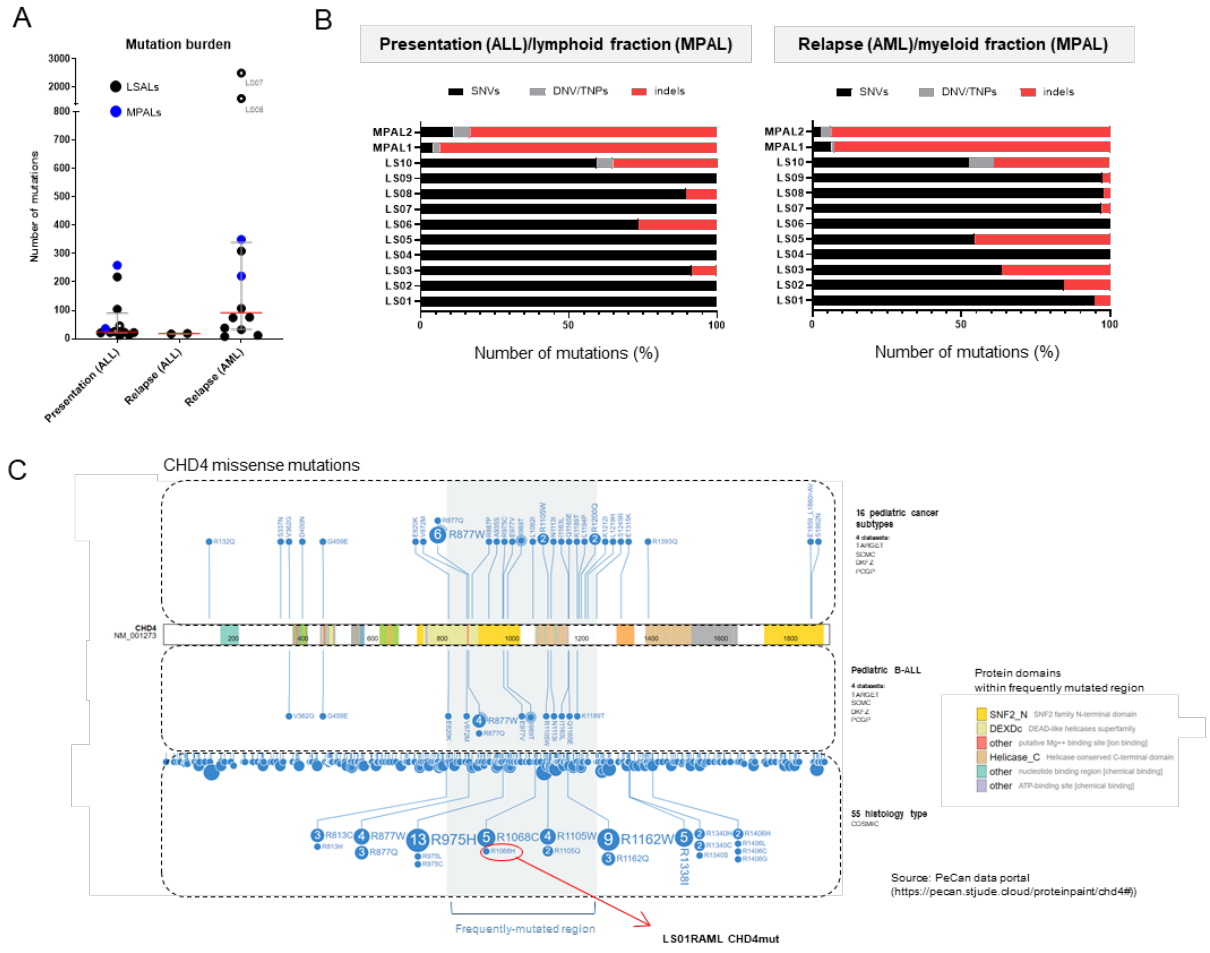

Figure S5. Mutational burden of lineage switched *MLL/AF4* leukaemias. (A) Total number of mutations identified in the analysed cohort. Individual LSAL patients presented are as black dots (black circles for patients LS07 and LS08), MPALs presented are as blue dots. Red lines correspond to the median value of mutations identified at the presentation and relapse stages of all analysed cohort. The initial ALL relapse of patients LS02 and LS03 was followed by secondary relapse with lineage switch; mutation burdens for both initial relapses is also presented. (B) Mutational composition of every patient sample analysed at lymphoid and myeloid leukaemia stage/fraction in LSALs and MPALs, respectively. Presented is percentage of single-, double-and triple-nucleotide variants (SNVs, DNV/TNVs) vs insertion and deletion (indels) variants in every analysed sample. (C) Protein domains of NuRD complex member CHD4 and identified missense mutations. Within published cases of childhood ALL, CHD4 is the only NuRD complex member to show recurrent mutations. These cluster within the

frequently mutated region, containing SNF2_N/DEAD-like helicase and ATPase/Helicase domains. Mutation identified in AML relapse of patient LS01 (LS01RAML) is located within this region and has been previously reported in colon cancer patient (COSMIC database). Figure adapted with permission from St Jude Cloud using the ProteinPaint portal (proteinpaint.stjude.org) [24].

Figure S6

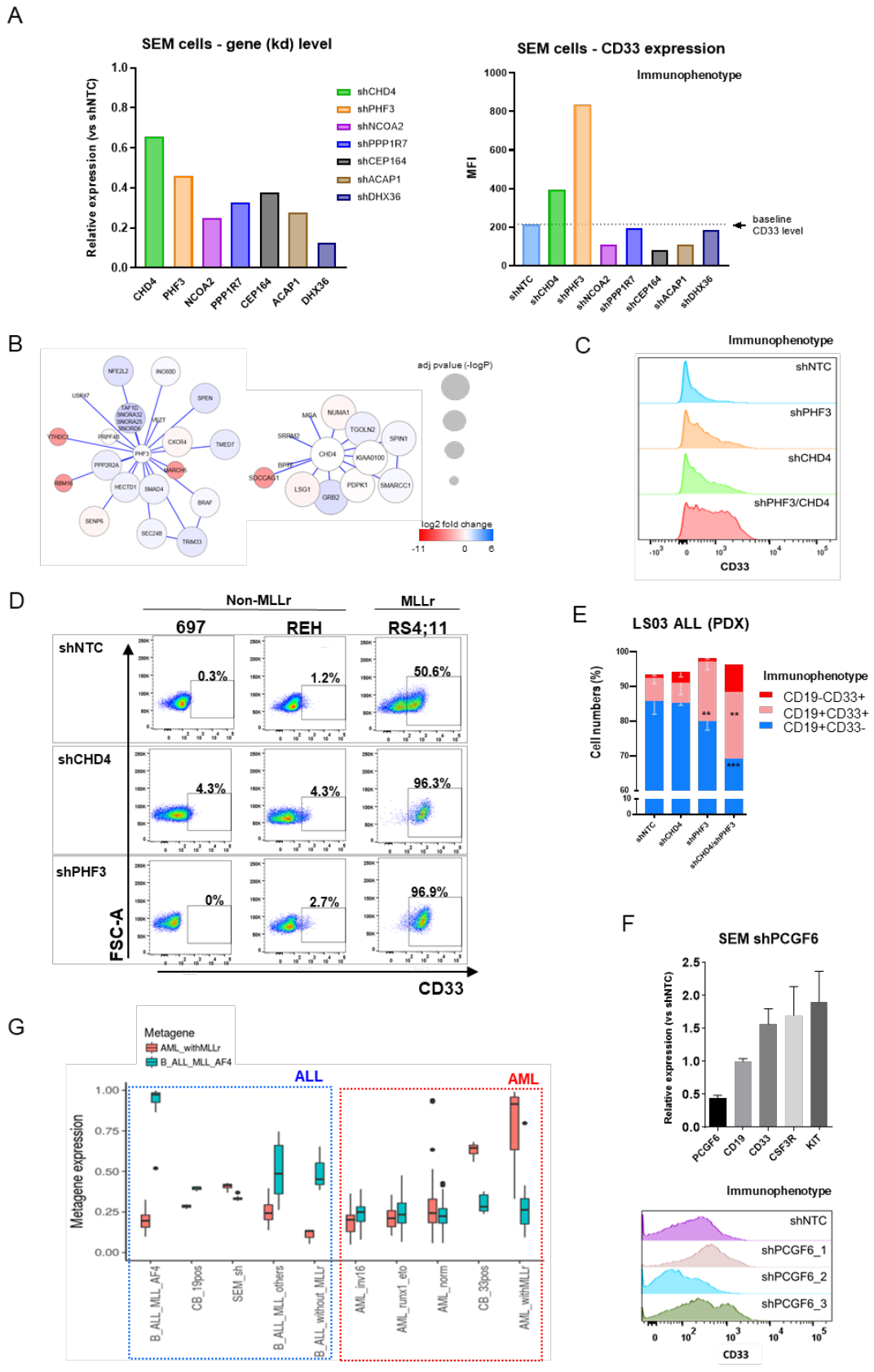
A
SEM cells - gene (kd) level
Relative expression (vs shNTC)
shCHD4
shPHF3
shNCOA2
shPPP1R7
shCEP164
shACAP1
shDHX36
CHD4
PHF3
NCOA2
PPP1R7
CEP164
ACAP1
DHX36
SEM cells - CD33 expression
Immunophenotype
MFI
baseline CD33 level
shNTC
shCHD4
shPHF3
shNCOA2
shPPP1R7
shCEP164
shACAP1
shDHX36
B
adj pvalue (-logP)
log2 fold change
C
Immunophenotype
shNTC
shPHF3
shCHD4
shPHF3/CHD4
CD33
D
Non-MLLr
MLLr
697
REH
RS4;11
shNTC
shCHD4
shPHF3
0.3%
1.2%
50.6%
4.3%
4.3%
96.3%
0%
2.7%
96.9%
FSC-A
CD33
E
LS03 ALL (PDX)
Cell numbers (%)
Immunophenotype
CD19-CD33+
CD19+CD33+
CD19+CD33-
shNTC
shCHD4
shPHF3
shCHD4/shPHF3
F
SEM shPCGF6
Relative expression (vs shNTC)
PCGF6
CD19
CD33
CSF3R
KIT
Immunophenotype
shNTC
shPCGF6_1
shPCGF6_2
shPCGF6_3
CD33
G
Metagene
AML_withMLLr
B_ALL_MLL_AF4
ALL
AML
Metagene expression
B_ALL_MLL_AF4
CB_19pos
SEM_sh
B_ALL_MLL_others
B_ALL_without_MLLr
AML_inv16
AML_runx1_eto
AML_norm
CB_33pos
AML_withMLLr

Figure S6. The functional role of epigenetic modifiers on leukaemia lineage switching. (A) Knockdown (kd) level of genes identified to be potentially driving lineage switch in LS01 patient (left) and corresponding expression level of CD33 marker (right) analysed by flow cytometry in SEM cells. Gene expression levels were normalised by *GAPDH* and presented as fold change vs non-targeting control (shNTC). CD33 expression level is presented as median fluorescence value (MFI) determined by flow cytometry. (B) Mutual information sub-networks illustrate high centrality of mutated genes *PHF3* (left panel) and *CHD4* (right panel) within the AML/ALL transcriptional network. Size and colour of nodes reflect differential expression between AML and ALL. (C) Flow cytometric analysis of the surface CD33 expression following knockdown of *PHF3*, *CHD4* or the combination in *MLL/AF4* positive ALL cell line, SEM. (D) Myeloid marker CD33 expression level upon knockdown of *CHD4*, *PHF3*, and non-targeting control (NTC) in *MLL*r (t(4;11)) cell line RS4;11 and non-*MLL*r cell lines, 697 and REH cells. (E) CD19 and CD33 surface expression level change upon knockdown of *CHD4* and *PHF3* in the PDX sample generated from the first relapse of patient LS03 (ALL relapse). (F) mRNA expression of lineage specific surface markers following *PCGF6* knockdown in SEM cells. Knockdown of *PCGF6* results in upregulation of *CD33*, *CSFR3* and *KIT* mRNA (left panel). Surface CD33 expression increases substantially with shPCGF6_1 and shPCGF6_3, compared with shNTC (right panel). (G) Boxplot of NMF projections of CD19+ cord blood (CB_19pos) and CD33+ cord blood (CB_33pos) populations relative to the derived AML with *MLL*r metagene (red boxes) and B precursor ALL with *MLL*r metagene (turquoise boxes).

Figure S7

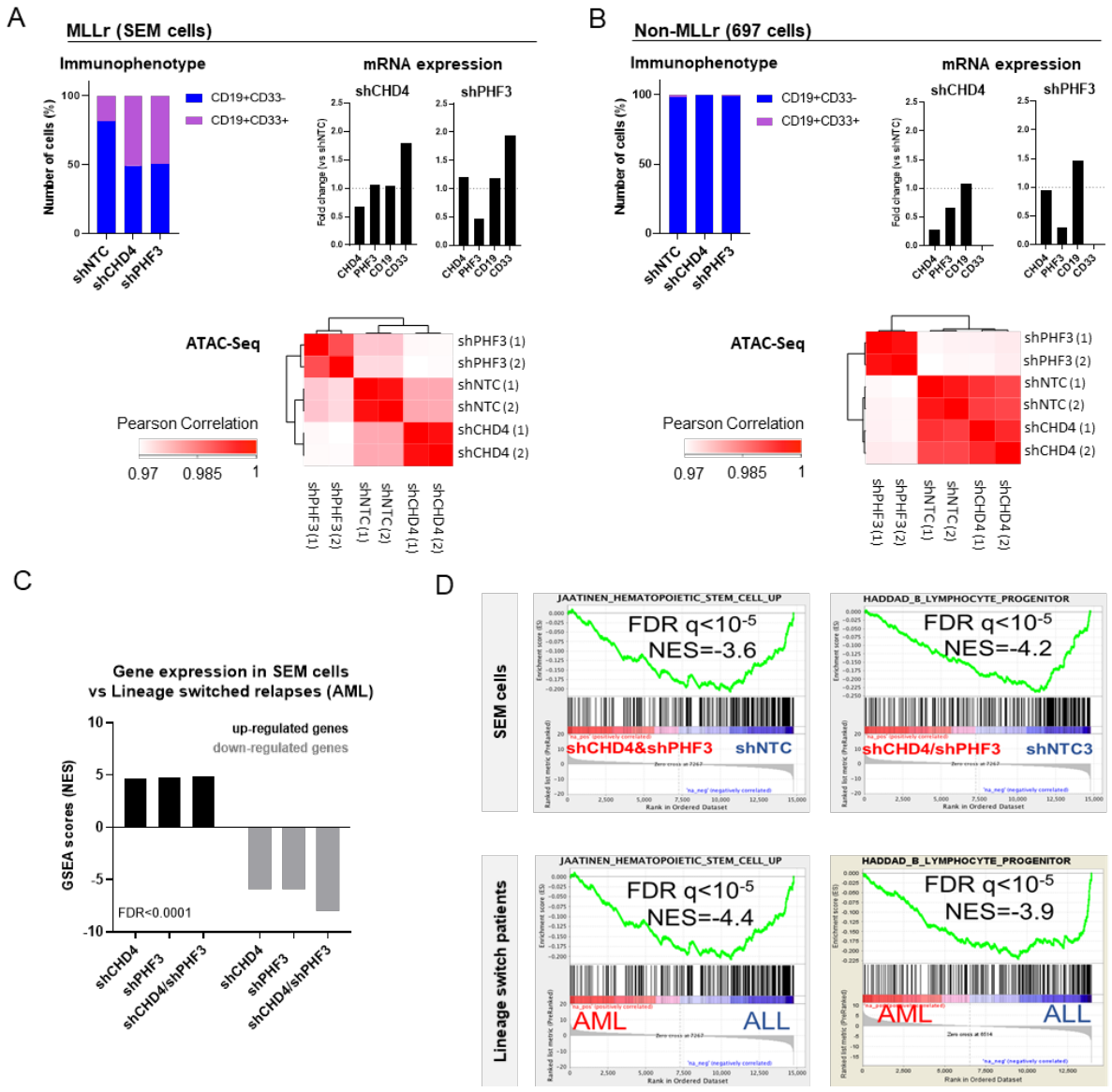

Figure S7. CHD4 and PHF3 disruption affects chromatin accessibility and gene expression in MLLr positive cells. (A) Expression of lineage specific surface markers (left upper panel) and their corresponding mRNA levels (right upper panel) following *CHD4* or *PHF3* knockdown in *MLL*r SEM cells and (B) non-*MLL*r 697 cells. Transposition experiments (ATAC-seq) have been performed with each cellular variant (shNTC, shCHD4 and shPHF3 of SEM and 697 cells) in duplicate and their comparison is presented as hierarchical clustering plots (lower panel). (C) GSEA comparison of transcriptomic changes under specific gene knockdown in SEM cells analysed against the up and downregulated gene sets of lineage

switch patient samples. Shown are NES values (normalised enrichment scores) depicting similarities in compared RNA-seq datasets. (D) Gene set enrichment analysis of RNA sequencing data derived from knockdown of *PHF3* and *CHD4* in the SEM cell line (upper panel), and from differential gene expression upon lineage switch (lower panel). Shown is negative correlation of shCHD4/shPHF3 and relapse samples with Jaatinen haematopoietic stem cell signature[25] (left panel) and Haddad B-lymphocyte progenitor signature [26] (right panel).

Figure S8

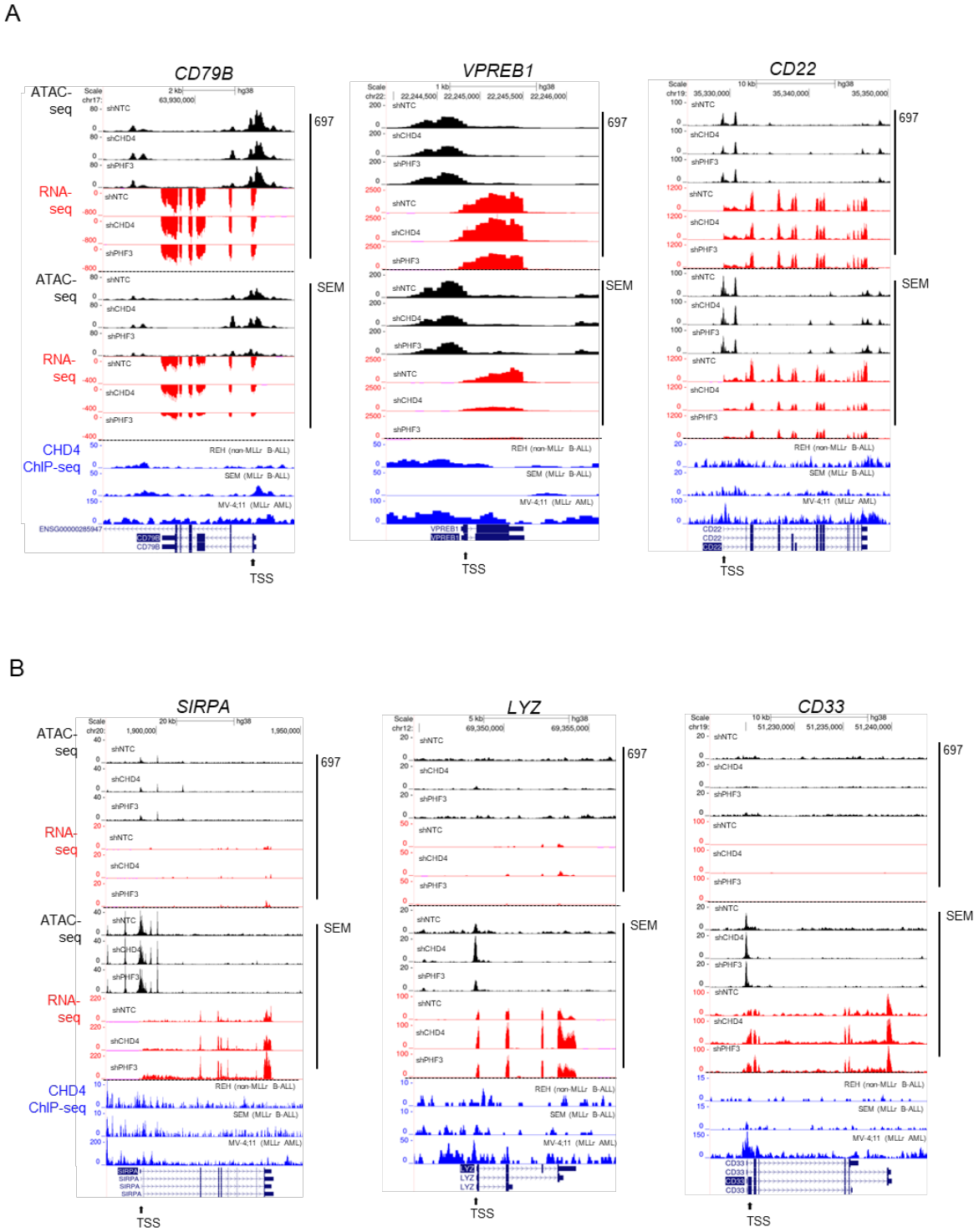

Figure S8. Differential expression of lymphoid and myeloid genes in MLLr cells with disrupted *CHD4* or *PHF3*. UCSC genome browser screenshots representing differential promoter accessibility (ATAC-seq) and gene expression level (RNA-seq), following *CHD4* and *PHF3* knockdown in *MLL*r SEM cells and non-*MLL*r 697 cells. Presented are (A) lymphoid

genes (e.g. *CD79B, VPREB1, CD22*) with reduced chromatin accessibility/gene expression and (B) myeloid genes (e.g. *SIRPA, LYZ, CD33*) with increased chromatin accessibility/gene expression, occurring specifically under gene knockdown of *CHD4* or *PHF3*. ChIP-seq traces representing normal CHD4 occupancy in non-*MLL*r B-ALL (REH cells), *MLL*r B-ALL (SEM cells) and *MLL*r AML cells (MV-4;11) are shown as a reference at the bottom of each screenshot. TSS – transcriptional start site is presented for each gene.

Figure S9

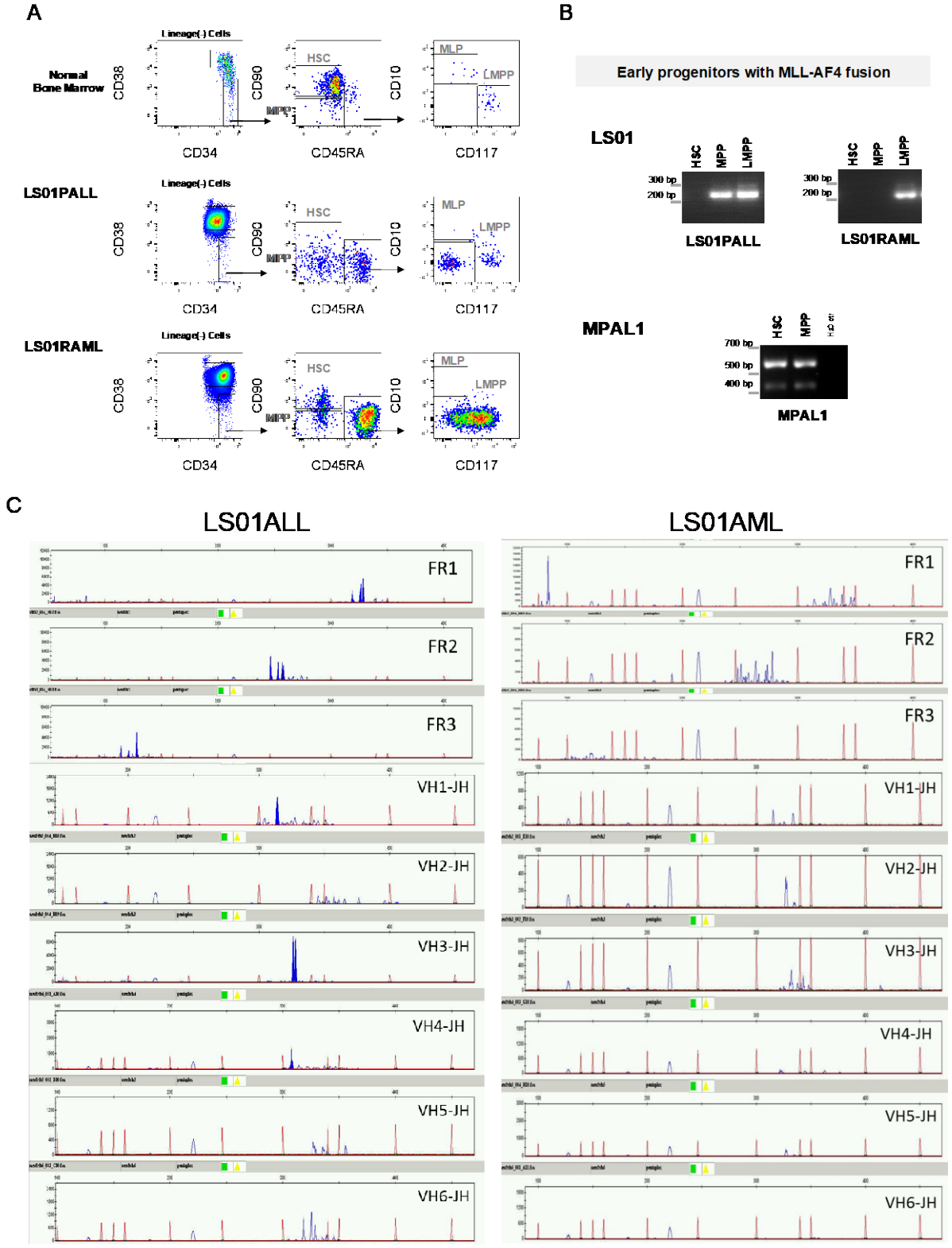
A
Normal Bone Marrow
LS01PALL
LS01RAML
Lineage(-) Cells
CD38
CD34
CD45RA
CD90
CD10
CD117
HSC
MPP
LMPP
MLP
B
Early progenitors with MLL-AF4 fusion
LS01
HSC
MPP
LMPP
300 bp
200 bp
LS01PALL
LS01RAML
MPAL1
700 bp
500 bp
400 bp
C
LS01ALL
LS01AML
FR1
FR2
FR3
VH1-JH
VH2-JH
VH3-JH
VH4-JH
VH5-JH
VH6-JH

Figure S9. Cellular origin of the *MLL/AF4* leukaemic relapse. (A) Flow cytometric sorting strategy for haematopoietic stem/progenitor cell (HSPC) populations. (B) PCR identification of the specific *MLL/AF4* fusion within sorted HSPC populations LS01 and MPAL1 cases. (C) Immunoglobulin rearrangement on ALL and AML LS01 according to the standard protocol BIOMED-2. Clonal peaks are coloured in dark blue, as indicated in the multiplex amplification FR1, FR2, and FR3 of the VH segments LS01PALL (left panel). The multiplex results were confirmed by single amplification VH-JH primer sets, which further showed two clonal VH3, one clonal VH1, and clonal VH4 rearrangement. No clonal peaks are seen in LS01RAML (right panel).

Supplementary References

1. Lin S, Luo RT, Ptasinska A, 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.

2. Cauchy P, James SR, Zacarias-Cabeza J, et al. Chronic FLT3-ITD Signaling in Acute Myeloid Leukemia Is Connected to a Specific Chromatin Signature. *Cell Rep*. 2015;12(5):821-836.

3. Piper J, Elze MC, Cauchy P, Cockerill PN, Bonifer C, Ott S. Wellington: a novel method for the accurate identification of digital genomic footprints from DNase-seq data. *Nucleic Acids Res*. 2013;41(21):e201.

4. Piper J, Assi SA, Cauchy P, et al. Wellington-bootstrap: differential DNase-seq footprinting identifies cell-type determining transcription factors. *BMC Genomics*. 2015;16:1000.

5. Corces MR, Trevino AE, Hamilton EG, et al. An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nat Methods*. 2017;14(10):959-962.

6. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods*. 2013;10(12):1213-1218.

7. Amemiya HM, Kundaje A, Boyle AP. The ENCODE Blacklist: Identification of Problematic Regions of the Genome. *Sci Rep*. 2019;9(1):9354.

8. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009;25(16):2078-2079.

9. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics (Oxford, England)*. 2014;30(7):923-930.

10. Heinz S, Benner C, Spann N, 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.

11. Grinev VV, Barneh F, Ilyushonak IM, et al. RUNX1/RUNX1T1 mediates alternative splicing and reorganises the transcriptional landscape in leukemia. *Nature Communications*. 2021;12(1):520.

12. Hehir-Kwa JY, Koudijs MJ, Verwiel ETP, et al. Improved Gene Fusion Detection in Childhood Cancer Diagnostics Using RNA Sequencing. *JCO Precis Oncol*. 2022;6:e2000504.

13. Haas BJ, Dobin A, Li B, Stransky N, Pochet N, Regev A. Accuracy assessment of fusion transcript detection via read-mapping and de novo fusion transcript assembly-based methods. *Genome Biol*. 2019;20(1):213.

14. Haas BJ, Dobin A, Stransky N, et al. STAR-Fusion: Fast and Accurate Fusion Transcript Detection from RNA-Seq. *BioRxiv*. 2017.

15. Liao Y, Smyth GK, Shi W. The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. *Nucleic Acids Res*. 2019;47(8):e47.

16. Meyer C, Schneider B, Reichel M, et al. Diagnostic tool for the identification of MLL rearrangements including unknown partner genes. *Proc Natl Acad Sci U S A*. 2005;102(2):449-454.

17. Meyer C, Hofmann J, Burmeister T, et al. The MLL recombinome of acute leukemias in 2013. *Leukemia*. 2013;27(11):2165-2176.

18. van Dongen JJ, Langerak AW, Bruggemann M, et al. Design and standardization of PCR primers and protocols for detection of clonal immunoglobulin and T-cell receptor gene recombinations in suspect lymphoproliferations: report of the BIOMED-2 Concerted Action BMH4-CT98-3936. *Leukemia*. 2003;17(12):2257-2317.

19. Margolin AA, Nemenman I, Basso K, et al. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*. 2006;7 Suppl 1:S7.

20. Shannon P, Markiel A, Ozier O, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*. 2003;13(11):2498-2504.

21. Schwarzer A, Emmrich S, Schmidt F, et al. The non-coding RNA landscape of human hematopoiesis and leukemia. *Nat Commun*. 2017;8(1):218.

22. Bomken S, Buechler L, Rehe K, et al. Lentiviral marking of patient-derived acute lymphoblastic leukaemic cells allows in vivo tracking of disease progression. *Leukemia*. 2013;27(3):718-721.

23. Livak KJ, Schmittgen TD. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods*. 2001;25(4):402-408.

24. Zhou X, Edmonson MN, Wilkinson MR, et al. Exploring genomic alteration in pediatric cancer using ProteinPaint. *Nat Genet*. 2016;48(1):4-6.

25. Jaatinen T, Hemmoranta H, Hautaniemi S, et al. Global gene expression profile of human cord blood-derived CD133+ cells. *Stem Cells*. 2006;24(3):631-641.

26. Haddad R, Guardiola P, Izac B, et al. Molecular characterization of early human T/NK and B-lymphoid progenitor cells in umbilical cord blood. *Blood*. 2004;104(13):3918-3926.