Supplemental Data file

Supplemental Excel files
The first set of 6 files (MACE-data) contain the gene entries with gene identifier, gene symbol, gene name, mean exp, mean mock, fold change log2, pvalue, -log10(pvalue), var-log2, chromosome number (with beginning and end), classification (into pseudogene (PG), non-annotated gene (NA), LINC RNA gene (LINC), MIR gene (MIR), SNO gene (SNO), MT gene (MT) and protein coding gene (PCG)). All data are given as in the Bioconductor output file with the exception of var-log2 which was simply calculated by ln(mean exp/mean mock)/ln(2). The following files are provided:
1. MLL-AF6 gene signature.xlsx
2. AF6-MLL gene signature.xlsx
3. CO1 gene signature.xlsx
4. exMLL-AF6 gene signature.xlsx
5. AF6-shMLL gene signature.xlsx
6. CO2 gene signature.xlsx

The second set of 6 files (ATAQ-data) contain the gene entries with gene identifier, gene symbol, gene name, mean exp, mean mock, fold change log2, pvalue, var-log2, chromosome number (with beginning and end), classification (into pseudogene (PG), non-annotated gene (NA), LINC RNA gene (LINC), MIR gene (MIR), SNO gene (SNO), MT gene (MT) and protein coding gene (PCG)). All data are given as in the Bioconductor output file with the exception of var-log2 which was simply calculated by ln(mean exp/mean mock)/ln(2). The following files are provided:
7. ATAC MLL-AF6.xlsx
8. ATAC AF6-MLL.xlsx
9. ATAC CO1.xlsx
10. ATAC exMLL-AF6.xlsx
11. ATAC AF6-shMLL.xlsx
12. ATAC CO2.xlsx

Additional Excel files provided are the gene sets used for the heatmaps:
13. Heatmap Set 1
14. Heatmap Set 2

Methods

QPCR experiments
Total RNA was extracted using RNeasy® Mini Kit (Qiagen). Complementary DNA was synthesized using SuperScript II reverse transcriptase (Invitrogen). An aliquot of the produced cDNA was used for quantitative PCR. Quantitative PCR were performed using ORA™ qPCR Probe ROX H Mix (highQu). Fluorescence signals above the threshold (CT) were determined with the aid of the StepOnePlus™ Real-Time PCR System (Life Technologies, Darmstadt, Germany). The fluorescence intensity of the samples was normalized to the amplification value of the reference gene glyceraldehyde-3-phosphate dehydrogenase (GAPDH). $2^{-\Delta\Delta CT}$ method was used to calculate the expression of the target mRNAs. For qPCR, the following primers were used: CCND1_For (5’-GCAGGCTGAACCTGAGG-3’) in combination with CCND1_Rev (5’-CAGGCTGAACCTGAGG-3’), KLF8_For (5’-CTGGATGGTGCTCC-TGG-3’) with KLF8_Rev (5’-AGGGACAGGTGCAGAACC-3’), ZNF544_For (5’-GAAGCAGGTCACGCACAG-3’) with ZNF544_Rev (5’-CAGGCTGAACCTGAGG-3’), NOV_For
(5'-ACCGTCAATGTGAGATGCTG-3') with NOV_Rev (5'-CAGTTCTTGACACTGAGG-3'), PEG10_For (5'-CCCACCTCTTGATCTTG-3') with PEG10_Rev (5'-GACACCTCACCAGGTTG-3'), CD44_F (5'-ATCTTTGACATCCCTTTG-3') with CD44_R (5'-GAGACCTCTGAGG-3'), and GAPDH.3 (5'-CTTCACCACATGGGAAGG-3') with GAPDH.5 (5'-CCTGCTTTGACACTGAGG-3')

**Bioinformatic analyses and MACE-Seq experiments**

All data received from the Bioconductor software from our RNA-Seq or ATAC-Seq experiments were incorporated into a FILEMAKER database program which allowed filtering the data, and cross-comparison analyses between different data sets. Moreover, gene ontology and other databases were implemented as well to retrieve the necessary information that has been used throughout this manuscript. Resulting data from 3 biological replicates of all cell lines were compared with 3 biological replicates of mock-transfected cells. The MACE-libraries were prepared at GenXPro GmbH using the Massive Analysis of cDNA Ends (MACE-Seq) Library Preparation Kit (v2.0) from GenXPro GmbH. Briefly, RNA was fragmented and and cDNA was generated using Oligo(dT) primers with distinct Oligo IDs per sample for subsequent pooling of up to 24 samples. After pooling, cDNA was fragmented to an average size of 200 bp using the sonicator Biorupter Plus (Diagenode, Belgium). The distribution of cDNA fragment sizes was monitored using the automated microfluidic electrophoresis station LabChip GXII Touch HT platform (PerkinElmer, USA). The poly(A) containing cDNA fragments were purified using solid phase reversible immobilization (SPRI) beads (Agencourt AMPure XP, USA), end repaired and ligated to distinct 8-base pair UMI Adapters (also called TrueQuant adapters). Then, the library containing labelled and fragmentated cDNA was amplified by PCR, purified by SPRI beads (Agencourt AMPure XP, USA) and strand-specific sequenced using the HiSeq2500 (Illumina, USA). Bioinformatic analysis was performed according to the analysis pipeline for MACE libraries by GenXPro GmbH. Unique Oligo IDs and UMIs on each transcript allowed initial demultiplexing and subsequent removal of PCR-duplicates. The remaining reads were trimmed for high-quality as well as adapter-free sequences and aligned to the human reference genome (Genome Reference Consortium Human Build 38 patch release 13 (GRCh38.p13) using Bowtie 2. Resulting output data were implemented in the database program FileMaker for further analysis. All final data sets were exported from the FILEMAKER Database program as individual Excel documents for publication.

**ATAC sequencing experiments**

Cells were grown in 6-well plate and treated with 1 µg/ml Doxycyclin for 48 hours. After that cells were harvested, and viability was checked in every sample. In all samples viability was more than 90%. Cells were then resuspended in cold PBS and cell numbers in each sample were counted. 50,000 cells were centrifuged at 500 RCF for 5 minutes at 4°C in a fixed angle centrifuge. After centrifugation, 900 µl of supernatant was aspirated with P1000 pipette and the remaining 100 µl of supernatant was carefully aspirated by pipetting with a P200 pipette tip to carefully avoid the cell pellet. Cells were resuspended in 50 µl cold ATAC-Seq Resuspension Buffer (RSB; 10 mM Tris-HCl pH 7.4, 10 mM NaCl, and 3 mM MgCl2 in water) containing 0.1% NP40, 0.1% Tween-20, and 0.01% Digitonin, pipetting up and down 3 times. Cells were incubated on ice for 3 minutes and then 1 ml of cold ATAC-Seq Resuspension Buffer (RSB) containing 0.1% Tween-20 but no NP-40 and Digitonin was added. Tubes were inverted 3 times to mix and then centrifuged at 500 RCF for 10 minutes at 4°C in a fixed angle centrifuge to pellet down the nuclei. Supernatant was removed with two
pipetting steps, as described before, and nuclei were resuspended in 50 µl of transposition mix (25 µl 2× TD buffer (20 mM Tris-HCl pH 7.4, 10 mM MgCl₂, 20 % Dimethyl Formamide in water), 2.5 µl transposase 26 (100 nM final), 16.5 µl PBS, 0.5 µl 1% digitonin, 0.5 µl 10% Tween-20, and 5 µl water) by pipetting up and down six times. Transposition reactions were incubated at 37°C for 30 min in a thermomixer with shaking at 1000 rpm. The final ATAC-Seq experiment was performed at GenXPro (https://genxpro.net/) that used an Illumina HiSeq for the analysis, and the resulting data were analysed by bioinformatics tools.

Figure legends

Figure S1: Workflow for the bioinformatic analyses
Upper panel: construction of the 6 stable cell lines together with the mock cell line. Transfection with Sleeping Beauty vectors usually took 7-12 days in order to obtain the selected and stable cell lines. Nucleic acids (RNA or DNA) were harvested 48h after the transgene expression was induced by 1 µg/ml Doxycyclin.Middle panel: MACE or ATAC-Seq experiments were performed and resulting data were analyzed by Bioconductor software to create output Excel files. Various other bioinformatic tools were used to analyze these data (volcano plots with huygens.science.uva.nl; heatmaps with biit.cs.ut.ee/clustvis; pathway analysis with bioinformatics.sdstate.edu). Circos plots have been used to display the genome-wide changes identified by RNA-Seq or ATAC Seq experiments. Finally, all Bioconductor output files have been imported to the Filemaker database program, which allowed to create novel modules for refined analyses not provide by the Bioconductor package (GUDC, DAGT, DAGE & ST). These modules allow to analyze the data beyond conventional data handling.

Figure S2: QPCR experiments with identified target genes
Target genes identified in CO1 cells by MACE were validated by QPCR experiments. Here, an example of 6 different analyses is shown. The QPCR data is shown for mock, MLL-AF6, AF6-MLL and CO1 cells, while the MACE reads are listed below. These analyses were made only for selected genes in order to demonstrate the high concordance between MACE-data and Q-RT-PCR data.

Figure S3: Detailed analyses of protein coding genes in the gene signatures of CO1 and CO2 cells
The identified gene signature in CO1 and CO2 cells were investigated for commonly and idiosyncratic protein coding genes. All these signatures (green colors: upregulated genes; red colors: downregulated genes). These subsets were then investigated by pathway analyses. CO1 (980 up- and 480 downregulated genes) and CO2 cells (655 up- and 74 downreguklated genes) share 256 protein coding genes, while displaying 243 and 135 idiosyncratic protein coding genes. The same data sets display 10 protein coding genes that were significantly downregulated, while having 341 and 8 protein coding genes downregulated as idiosyncratic signatures. Subsequently performed pathway analyses revealed interesting pathways which are shown by the different colors and identical numbers. Of interest, the up-regulated idiosyncratic signature of CO2 cells revealed a lymphoid-specific, and more specifically, a T-cell specific gene signature (see e.g. CD4, CD74, LAT2, IKZF1 and LMO2).

Figure S4: Signature analysis using to the GUDC module
Investigation of the distribution of activated and repressed genes on different chromosomes to generate patterns that demonstrate the genome-wide activity of the individual fusion proteins. The diagrams show two things: (1) mean transcription in percentage of genes located on all chromosomes (e.g. MLL-AF6 with 0,26% of all genes, while CO1 cells expressed 3,18% of all encoded genes); (2) the deviation from the mean transcription in the graph on the right. These
deviations display whether more or fewer genes became activated or repressed when compared to the mean of genes activated or downregulated on whole chromosomes. Thus, this module displays a fingerprint pattern of gene usage per chromosome by the individual fusion proteins. The analysis demonstrates again that effective up- and downregulation that occur only when MLL-AF6 and AF6-MLL are co-expressed in cells (C01: synergistic effect), while exMLL-AF6 alone was capable to perform the upregulation of genes, while a down-regulation is nearly missing. The pattern only slightly changed in CO2 cells, arguing for an additive effect of both fusion proteins exMLL-AF6 and AF6-shMLL, respectively.

**Figure S5: Dissection of the ATAC-Seq data**
The obtained ATAC data from our experiments were summarized. Upper panel: left column displays names of all fusion genes or their combination (CO1 or CO2). The following columns contain information about number of gene entries and the number of totals up- and down-regulated genes. The last 4 columns represent the total reads of the up- and down-regulated gene signature. These were calculated for up- and downregulated chromatin regions by a minimum of 2 reads, a p-value < 0.05 combined with log2 changes of > 1 or 2 in case of up-regulated genes, while downregulated genes were identified by a minimum of 2 reads in the mock sample, a p-value < 0.05 combined with a log2 value of < -1 or -2. Lower panel: all data were dissected for the number of pseudogenes, non-annotated genes, LincRNA genes, microRNA genes, SNO genes, mitochondrial genes and protein coding genes, respectively.

**Figure S6: Signature analysis using to the DAGT module**
Cross-comparison between MACE and ATAC-data. Left table summarizes the gene signatures of all 6 cell lines with precise numbers for pseudogene/non-annotated genes (PG/NA) and protein coding genes (PCG) in the up- (green) or down-regulated gene signatures (red). Right graphic: the signatures of highly deregulated genes (log2 = ±1) found in the MACE experiments was compared to more or less accessible chromatin found in the ATAC-Seq experiment. Numbers in the green and red rectangles show e.g. that 198 out of 203 identified up-regulated genes in MLL-AF6 cells could be attributed to 151 accessible and 47 less accessible chromatin fragments. *Vice versa*, 31 out of 31 MACE downregulated genes could be associated with 9 accessible and 22 less accessible chromatin fragments. In addition, we analyzed the gene type distribution for all 4 possible scenarios. Below the Table, the circle plot explained. All signatures were subdivided into protein coding genes (PCG), SNO RNA genes (SNO); MIR genes (MIR), LINC genes (LINC), non-annotated genes (NA) and pseudogenes (PG). Since the most abundant gene types are protein coding genes, as well as the class of pseudogene/non-annotated genes, we summed up the latter two (pink numbers) and compared it to PCG’s (blue numbers) by indicating their percentages in the center of each circle plot. As an example, the 151 genes identified for the upregulated gene signature in MLL-AF6 cells derived and deriving from the accessible chromatin fraction (green rectangle) contained 72.2% PCG’s but only 19.8% pseudogenes/non-annotated genes. This type of analysis was performed for all 24 subsections.
Figure S3

**signature protein coding gene analysis**

| Enrichment FDR | # genes | functional category |
|----------------|---------|---------------------|
| 0.000126450630 | 77      | Regulation of nucleosome remodeling |
| 0.000126450630 | 71      | Animal organ development |
| 2.19837531058   | 66      | Regulation of developmental process |
| 3.03315683058   | 51      | Regulation of cell differentiation |
| 7.027321058     | 37      | Negative regulation of transcription |

**Figure S4**

**Enrichment FDR**

| Gene       | FDR  |
|------------|------|
| MA6        | 0.26 |
| A6M        | 0.93 |
| CO1        | 2.18 |

**Enrichment FDR**

| Gene       | FDR  |
|------------|------|
| exMA6      | 1.89 |
| AtshM      | 0.15 |
| CO2        | 2.06 |

**Enrichment FDR**

| Gene       | FDR  |
|------------|------|
| MA6        | 0.71 |
| A6M        | 0.94 |
| CO1        | 1.57 |

**Enrichment FDR**

| Gene       | FDR  |
|------------|------|
| exMA6      | 0.25 |
| AtshM      | 0.04 |
| CO2        | 0.23 |
Figure S5

| gene entries | total up | total down | log2 > 1 | log2 > 2 | log2 < -2 | log2 < -1 |
|--------------|----------|------------|----------|----------|----------|----------|
| MLL-AF6      | 51,348   | 27,875     | 23,472   | 258      | 124      | 44       | 137      |
| AF6-MLL      | 51,625   | 29,188     | 22,437   | 318      | 16       | 72       |          |
| CO1          | 51,785   | 28,906     | 22,879   | 97       | 35       | 113      |          |
| exMLL-AF6    | 51,933   | 25,675     | 25,258   | 302      | 74       | 78       | 187      |
| AF6-shMLL    | 51,113   | 26,55      | 24,563   | 272      | 132      | 62       | 151      |

Figure S6

| gene entries | PG | NA genes | LINC | MIR | SNO | MT | PrCodG |
|--------------|----|----------|------|-----|-----|----|--------|
| MLL-AF6      | 52,236 | 9,78 | 17,211 | 1,839 | 1,192 | 241 | 37 | 21,955 |
| AF6-MLL      | 51,348 | 9,463 | 16,796 | 1,82 | 1,122 | 219 | 37 | 21,904 |
| CO1          | 51,625 | 9,519 | 16,968 | 1,819 | 1,131 | 238 | 37 | 21,93 |
| exMLL-AF6    | 51,785 | 9,607 | 17,028 | 1,817 | 1,152 | 233 | 37 | 21,937 |
| AF6-shMLL    | 50,933 | 9,312 | 16,623 | 1,818 | 1,077 | 215 | 37 | 21,867 |
| CO2          | 51,113 | 9,357 | 16,729 | 1,815 | 1,066 | 222 | 37 | 21,901 |