Supplementary Materials for

Identification of Functional Cooperative Mutations of *SETD2* in Human Acute Leukemia

Xiaofan Zhu[1,10], Fuhong He[2,10], Huimin Zeng[1,10], Shaoping Ling[2,10], Aili Chen[2-5,10], Yaqin Wang[1], Xiaomei Yan[3,4], Wei Wei[1], Yakun Pang[1], Hui Cheng[1], Chunlan Hua[1], Yue Zhang[1,3,4], Xuejing Yang[2,5], Xin Lu[2,5], Lihua Cao[2], Lingtong Hao[2], Lili Dong[2], Wei Zou[2], Jun Wu[2], Xia Li[2,5], Si Zheng[2,5], Jin Yan[2], Jing Zhou[2], Lixia Zhang[2,5], Shuangli Mi[2], Xiaojuan Wang[1], Li Zhang[1], Yao Zou[1], Yumei Chen[1], Zhe Geng[6], Jianmin Wang[7], Jianfeng Zhou[6], Xin Liu[8,9], Jianxiang Wang[1], Weiping Yuan[1], Gang Huang[3,4], Tao Cheng[1] & Qian-fei Wang[2]

[10]These authors contributed equally to this work.

Correspondence should be addressed to T.C. (chengt@pumc.edu.cn), Q.-f.W. (wangqf@big.ac.cn) or G.H. (gang.huang@cchmc.org).

This file includes:

Supplementary Note (pp 2-4)

Supplementary Figures 1 to 17 (pp 5-26)

Supplementary Tables 1 to 13 (pp 27-34)

Supplementary References (pp 35-36)

Nature Genetics: doi:10.1038/ng.2894

Supplementary Note

Sample description and preparation for the monozygotic twin pair

A 3-year-old female suddenly experienced pain in four limbs of her body and a fever temperature, and had WBC 10.7×10^9/L, HGB 93g/L, RBC 3.28×10^{12}/L, and PLT 21×10^9/L. Bone marrow morphology (Supplementary Fig. 1a) and immunophenotyping (Supplementary Fig. 1c,d) examinations revealed that her bone marrow had leukemia cell populations of immature myeloblasts with extremely active proliferation, different cell sizes, folded nuclei and less cytoplasm, expressing both CD56 and CD64, and other cell surface markers of acute myeloid leukemia (AML) phenotype, occupying 44% of total nucleated cells, and was diagnosed as AML FAB-M5. Fluorescence in situ hybridization (FISH, Supplementary Fig. 1b) showed *MLL* translocation in 35.5% of the cell population. Based on comprehensive clinical manifestations with bone marrow morphology, immunophenotyping and FISH, she was diagnosed as 11q23/*MLL* AML. She was dead with no treatment within approximately 3 months after her diagnosis as 11q23/*MLL* AML. We had confirmed that six common partner genes, *AFF1*, *MLLT4*, *MLLT3*, *MLLT1*, *ELL*, and *MLLT10,* were not *MLL* fusion partner in this patient using regular RT-PCR and nested RT-PCR experiments (data not shown). Chromosomal rearrangements involving *AML1-ETO*, *PML-RARA*, *BCR-ABL*, or *MLL-AF4* were undetectable by FISH (data not shown).

All of her family members, including her parents, her only elder brother, and her only twin sister, were healthy and did not display any features of acute leukemia. The leukemia patient and her healthy twin sister were confirmed to be monozygotic twins by a birth record from the hospital, genotyping of STR loci and Amelogenin loci (Supplementary Fig. 2a), and STR-PCR analysis (Supplementary Fig. 2b). FISH analyses showed that the healthy monozygotic twin sister did not have *MLL* translocation. No chromosomal rearrangements involving *AML1-ETO*, *PML-RARA*, *BCR-ABL*, or *MLL-AF4* were detected. According to the regulations of the institutional ethics review boards, informed consent was signed by their parents.

Peripheral blood mononuclear cells (PBMCs) from the leukemia patient twin (PT), cryopreserved in 10% DMSO, were rapidly thawed at 38°C, washed and stained with CD45+PerCP-Cy5 (Becton, Dickinson and Company), CD56PE (Becton, Dickinson and Company), and CD64FITC (Becton, Dickinson and Company). The blast population highly expressing CD56 and CD64 (CD56+CD64+ cells), and normal cell population negatively expressing CD56 and CD64 (CD56-CD64- cells) were sorted using a BD Aria II flow sorter (Becton, Dickinson and Company) (Supplementary Fig. 1d). In addition, the PBMCs originated from the healthy twin (HT) was used as an ideal matched normal control due to fewer contaminations of leukemia cells. Genomic DNA and total RNA, prepared from CD56+CD64+ leukemia cells (PT_Leukemia), CD56-CD64- normal cells (PT_Normal), and normal PBMCs (HT), were used for whole genome sequencing (WGS), mRNA-seq and mutation validation, respectively.

Generating WGS and mRNA-seq data for the pair of twins

Mate-paired libraries with a 1.5 Kb insert size were constructed using the reagents and protocol provided by Applied Biosystems (SOLiD 4 System Library Preparation Guide). Briefly, genomic DNA (PT_Leukemia: 25 ug, HT: 17 ug) was fragmented by HydroShear to ~1.5 Kb. The fragmented DNA was end-repaired, and adapter-ligated. DNA fragments between 1.0-2.0 Kb were selected, circularized with a biotinylated internal adaptor, and then purified. Fragments containing the target genomic DNA (about 100 bp) and adaptors were cleaved from the purified circularized DNA by single-strand specific S1 nuclease, ligated by P1 and P2 adaptors, and then PCR-amplified for ~12 cycles. Finally, 250-300 bp adapter-ligated fragments were selected to generate mate-pair (MP) sequencing libraries with average target genomic DNA on each end around 90 bp by excision from PAGE gel using an emulsion PCR template. Emulsion PCR and mate-pair sequencing were performed exactly according to the pipeline of Applied Biosystems SOLiD 4.

Total RNA was DNAse I treated, isolated for polyadenylated (poly-A) RNA, and followed by ds-cDNA synthesis. The resulting cDNA was fragmented to approximately ~400 bp, size-selected, end-repaired, and adapter-ligated. After ~15 rounds of PCR amplification, the final product was size-selected (~400 bp), and quantified with qPCR. Bridge PCR clusters were generated and followed by paired-end (PE) sequencing for 2x81 cycles according to the Illumina GA IIx.

All WGS and mRNA-seq data for this pair of twins were summarized in Supplementary Table 1.

Confirming *MLL* translocation via PCR

(1) Validation on DNA

Genomic DNA was extracted from PBMCs from the *MLL* leukemia patient (PT) and her healthy monozygotic twin sister (HT), and from 1 non-*MLL*-rearranged leukemia cell line (U937), respectively. H_2O was used to be a negative control. PCR amplification followed by Sanger sequencing was performed using primers spanning the fusion junctions. Sequences were generated and aligned to the reference genome to confirm the breakpoints and fusion junction sequences detected by CASfus and CASbreak. The primers for PCR validation were designed by Primer 5.0 using the default parameters and are provided in Supplementary Table 11.

(2) RT-PCR validation on RNA

Total RNA was harvested using Trizol in the PBMCs from PT and HT, respectively. Then, total RNA was reversely transcribed, purified, and amplified with a forward primer in *MLL* exon 8 and a reverse primer in *NRIP3* exon 6. Sanger sequencing was performed using this pair of primers. Sequences were generated and aligned to the reference genome to confirm alternative spliced transcripts predicted by mRNA-seq. The primers for RT-PCR validation were designed by Primer 5.0 with the default parameters and are provided in Supplementary Table 11.

(3) TaqMan PCR examination on RNA

Total RNA was harvested using Trizol in the PBMCs from PT_Leukemia, PT_Normal, and HT, respectively. The primers for TaqMan detecting *MLL-NRIP3* fusion transcripts were designed by primer express® software v3.0, and are provided in Supplementary Table 11. Expression level of *MLL-NRIP3* in each sample was examined by TaqMan PCR assay with high specificity and sensitivity. The expression of *MLL-NRIP3* was normalized against the expression of the control gene *PGK1* to adjust for variations in RNA quality and efficiencies of cDNA synthesis. The expression ratios are given as $\log_{10}$(*MLL-NRIP3*×10^9/*PGK1*).

Collecting acute leukemia patients for screening *SETD2* mutations

229 acute leukemia patients, consisting of 145 pediatric with age < 18 years old and 84 adults with age ≥18 years old, were collected from Institute of Hematology & Blood Diseases Hospital, Center for Stem Cell Medicine, Chinese Academy of Medical Sciences & Peking Union Medical College. Ten acute leukemia patients, consisting of 1 pediatric and 9 adults, were collected from TongJi Hospital, TongJi Medical College, HuaZhong University of Science & Technology. Two acute adult leukemia patients were collected from Changhai Hospital, Second Military Medical University. Among all the 241 patient samples collected, AMLs mainly consists of 42 pediatric and 90 adult leukemia patients, and 104 out of 107 ALLs are pediatric patients. Bone marrow and/or PBMCs were obtained for detecting *SETD2* mutation at diagnosis and the absence of antecedent chemotherapy and radiation therapy. For examining somatic status of candidate variants, we collected the genomic DNA from the matched normal tissue (skin, saliva, mesenchymal stem cell, or remission PBC) whose leukemia DNA harbored the variant, unless the matched normal control tissue was not available. DNA was prepared from all tissues using standard protocols. The institutional review boards of all participating institutions approved this study. Double-blinded to group allocation was performed for prognosis comparison between different patient groups. The detailed and summarized information of these patients is provided in Supplementary Tables 3-6.

Cloning sequencing for identifying allelic status of *SETD2* mutation

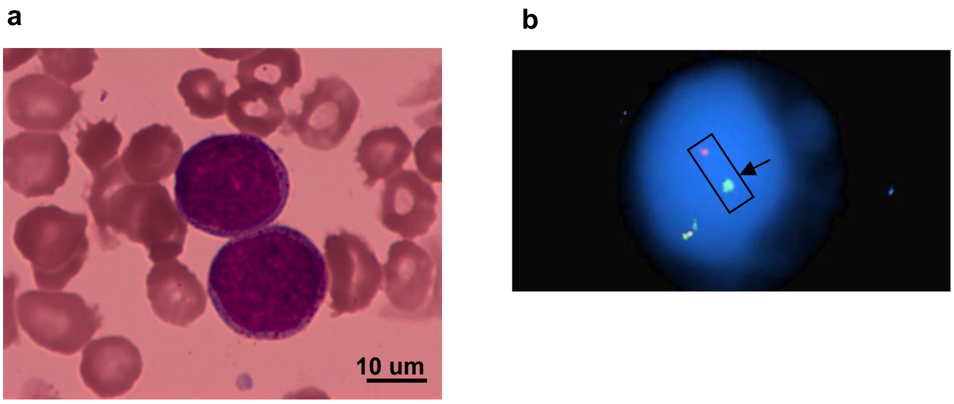
For patients with two *SETD2* mutations, cDNA or DNA was PCR amplified to encompass both identified mutations. PCR products were ligated into a T-vector (pGEM-T easy; Promega) according to the protocol, transferred to competent cells, heat-shocked, cultured and plated onto LB plates with X-gal and IPTG. Finally, positive colonies were selected and individually sequenced with universal primers. The primers were designed by Primer 5.0 with the Dimer and Cross Dimer parameters, and are provided in Supplementary Table 12.

Biological replicates for *Setd2* knockdown mouse assays

Multiple different shRNAs were tested, of which two constructs resulted in a strong reduction of *Setd2* expression and displayed similar effects on mRNA expression profile, H3K36me3 level, and leukemia development (Supplementary Fig. 13). We thus used a representative shRNA for further functional studies and data presentation.

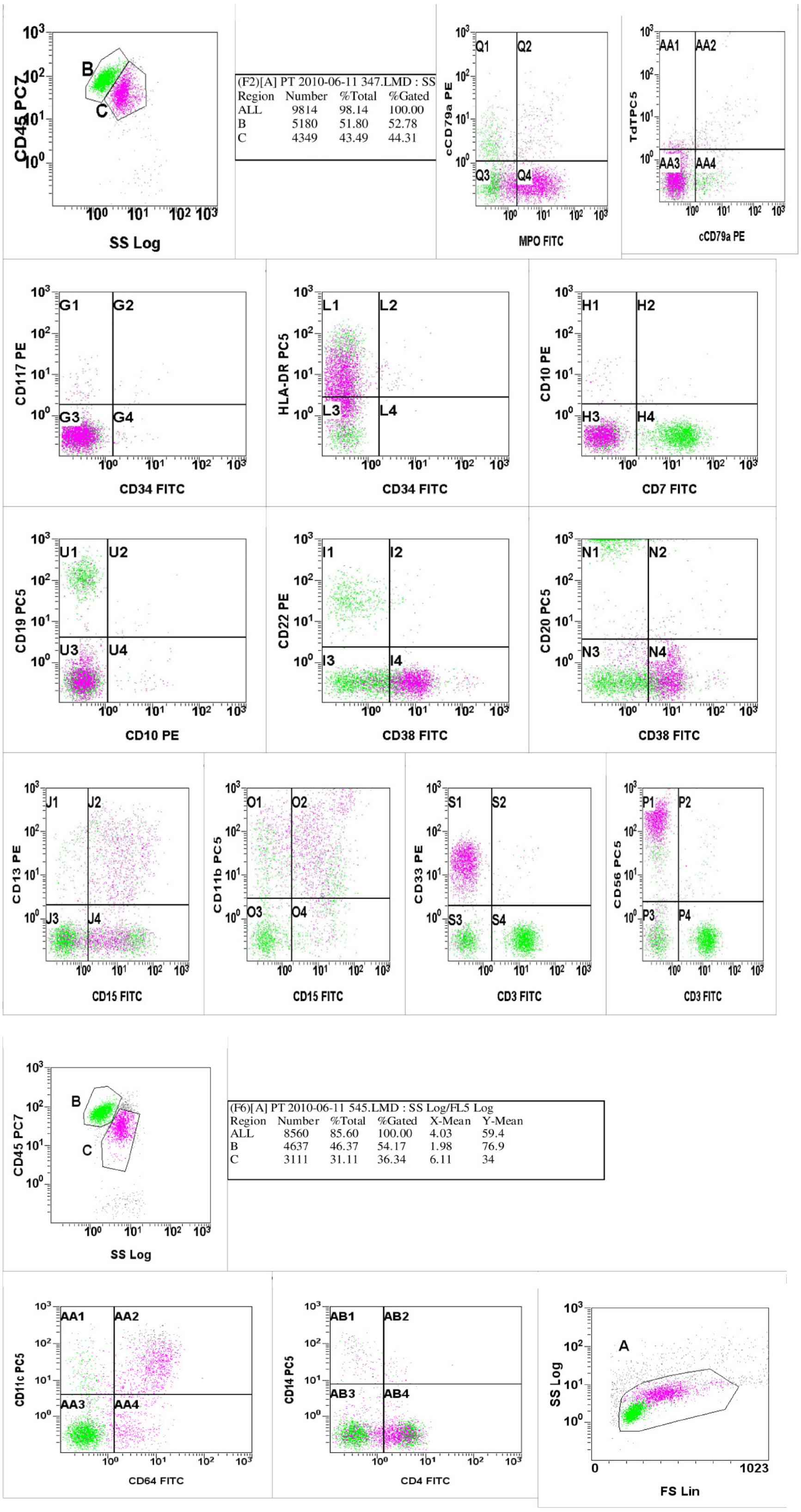
Supplementary Figures

Supplementary Figure 1. Morphology, immunophenotype, and cytogenetics analyses of the leukemia patient twin



a. Morphology of bone marrow biopsy specimen. *MLL* leukemia patient twin had a blast population of immature myeloblasts with extremely active proliferation, different cell sizes, folded nuclei and less cytoplasm.

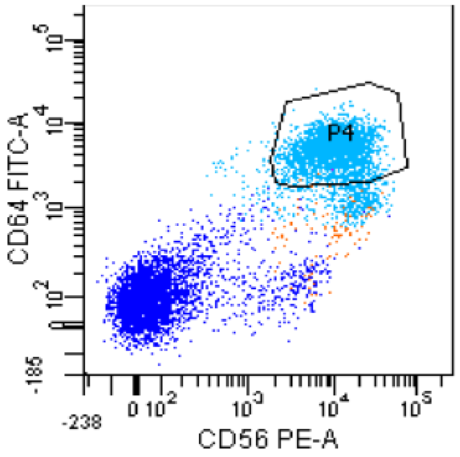
b. FISH analysis of peripheral blood cells. Interphase FISH of leukemic blasts was performed with dual-color break apart probes designed for the *MLL* gene. Green and red signals are not co-localized, which indicates *MLL* rearrangement (boxed and pointed to by an arrow); the single yellow signal represents wild-type *MLL*.



```
(F2)[A] PT 2010-06-11 347.LMD : SS
Region   Number   %Total   %Gated
ALL       9814     98.14    100.00
B         5180     51.80    52.78
C         4349     43.49    44.31
```

```
(F6)[A] PT 2010-06-11 545.LMD : SS Log/FL5 Log
Region   Number   %Total   %Gated   X-Mean   Y-Mean
ALL       8560     85.60    100.00   4.03     59.4
B         4637     46.37    54.17    1.98     76.9
C         3111     31.11    36.34    6.11     34
```

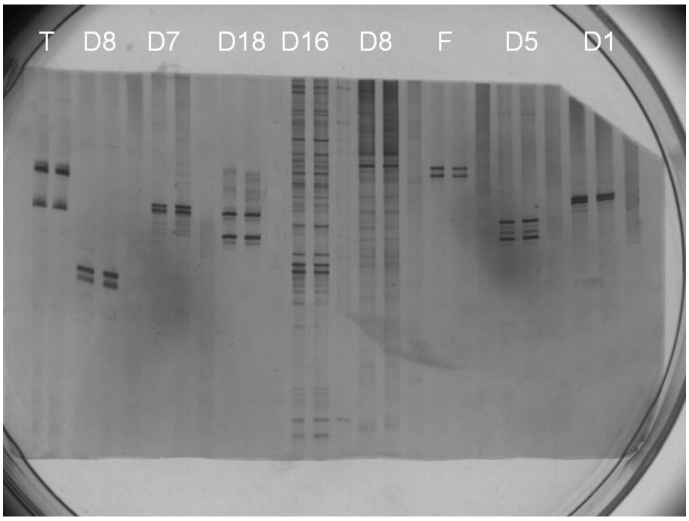
c. Immunophenotyping of peripheral blood cells using flow cytometry. The blast population from the *MLL* leukemia patient twin expressed cell surface markers which represents AML FAB-M5, and occupied 44% of total nucleated cells.



d. Isolation of leukemia cell population using flow cytometry. CD56+CD64+ leukemia cells and CD56-CD64- normal cells from the leukemia twin patient.

Supplementary Figure 2. Monozygosity of the twin sisters confirmed by STR-PCR and Genotyping

a

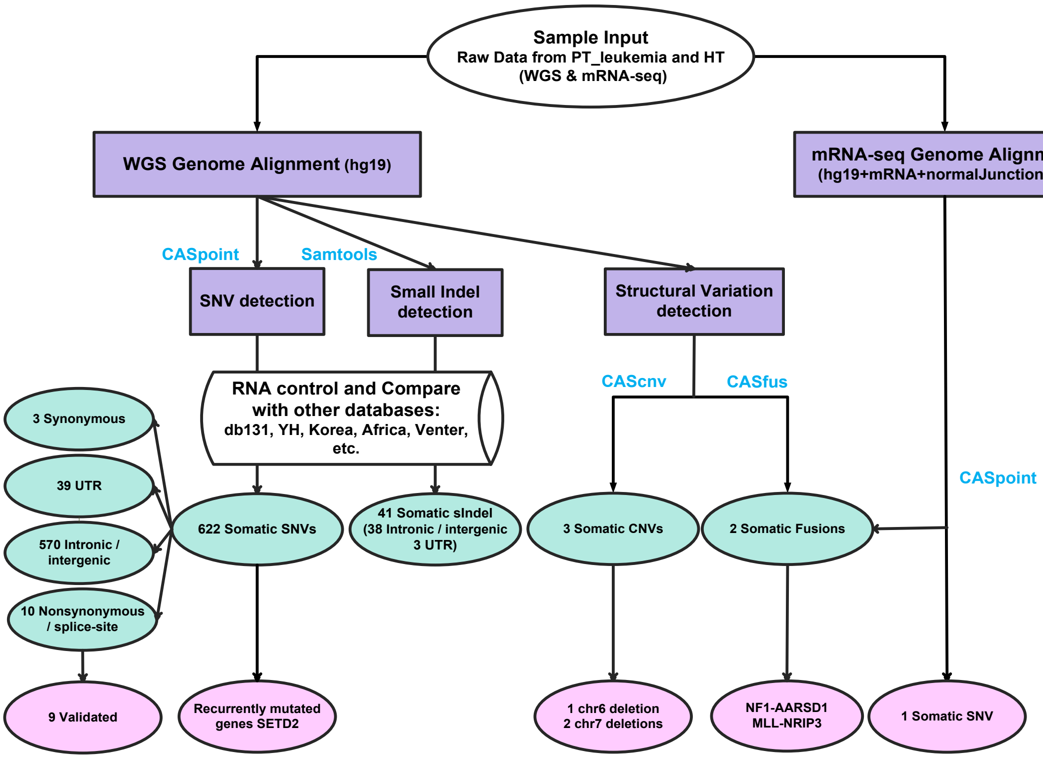


b

sample1			sample2		
Marker	Allele1	Allele2	Marker	Allele1	Allele2
AMEL	X	X	AMEL	X	X
D3S1358	15	16	D3S1358	15	16
D13S317	8	8	D13S317	8	8
D7S820	9	10	D7S820	9	10
D16S539	12	13	D16S539	12	13
Penta E	16	19	Penta E	16	19
TPOX	8	11	TPOX	8	11
TH01	7	9	TH01	7	9
D2S1338	23	23	D2S1338	23	23
CSF1PO	11	12	CSF1PO	11	12
D19S433	16	16	D19S433	16	16
vWA	14	18	vWA	14	18
D5S818	9	10	D5S818	9	10
FGA	19	19	FGA	19	19
D6S1043	14	18	D6S1043	14	18
D8S1179	15	15	D8S1179	15	15
D21S11	31	31	D21S11	31	31
D18S51	14	18	D18S51	14	18

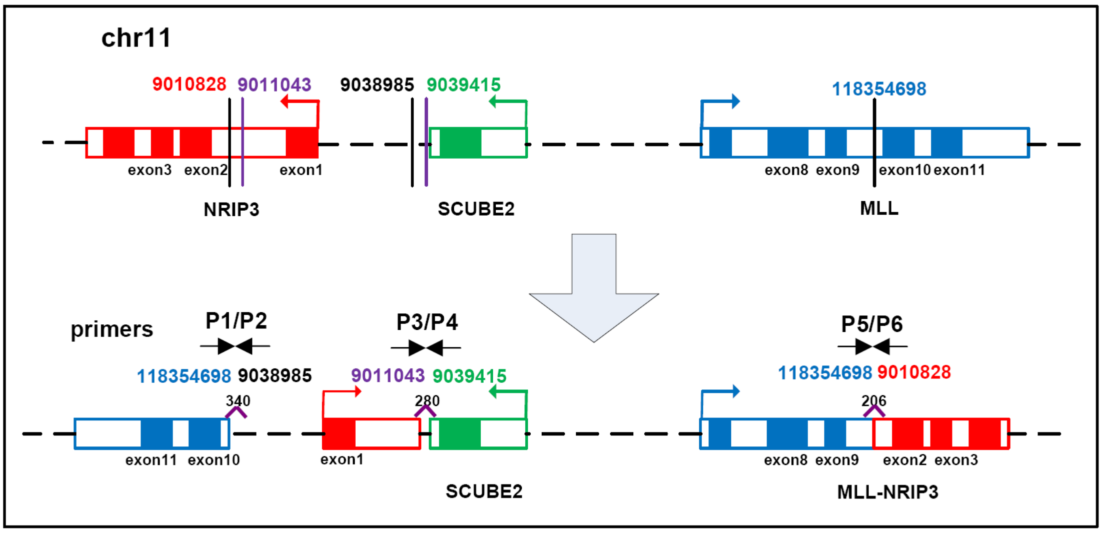
The twin sisters have the same sequence length in each of the 9 short tandem repeat (STR) loci (T, D8, D7, D18, D16, D8 F, D5, D1), as shown by STR-PCR identification (a), and have an identical genotype confirmed by genotyping of 18 STR loci and Amelogenin loci (b). These analyses indicate that the twin sisters are monozygotic with an identical genotype.

Supplementary Figure 3. Bioinformatics pipeline for identification of somatic mutations

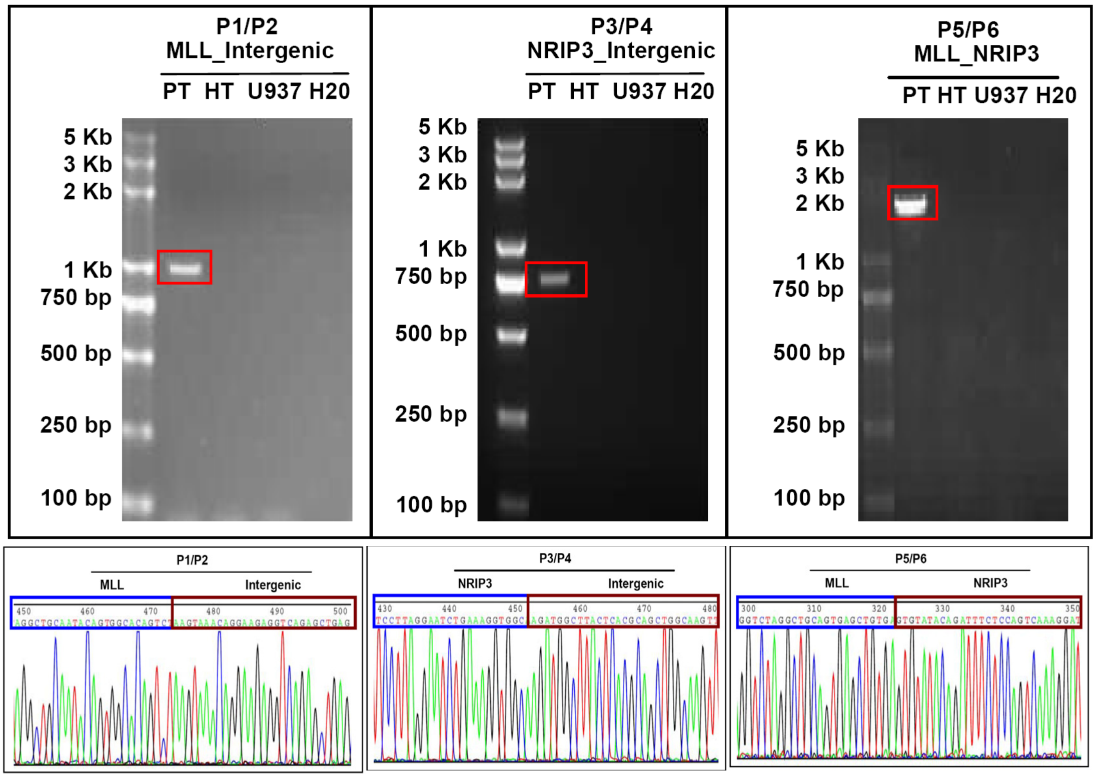

Sample Input
Raw Data from PT_leukemia and HT
(WGS & mRNA-seq)
WGS Genome Alignment (hg19)
mRNA-seq Genome Alignment
(hg19+mRNA+normalJunctionDB)
CASpoint
Samtools
SNV detection
Small Indel detection
Structural Variation detection
RNA control and Compare with other databases:
db131, YH, Korea, Africa, Venter, etc.
CAScnv
CASfus
CASpoint
3 Synonymous
39 UTR
570 Intronic / intergenic
10 Nonsynonymous / splice-site
622 Somatic SNVs
41 Somatic sIndel
(38 Intronic / intergenic
3 UTR)
3 Somatic CNVs
2 Somatic Fusions
9 Validated
Recurrently mutated genes SETD2
1 chr6 deletion
2 chr7 deletions
NF1-AARSD1
MLL-NRIP3
1 Somatic SNV

WGS and mRNA-seq data generated from the twin pair were aligned to the human reference genome sequence (GRCh37/hg19) using Bioscope 1.3 and BWA[1], respectively. CASpoint (an in-house software) and Samtools[2] were used to find single-nucleotide variants (SNVs) and small insertions and deletions (indel), respectively. Candidate SNVs were screened against dbSNP[3] (v131), YanHuang Project[4], Korean human polymorphism database[5], and 1000 human genome SNP database[6]. 622 potential somatic SNVs from WGS data included 9 nonsynonymous and 1 splice-site SNV. Moreover, 1 additional somatic SNV was rescued by analyzing mRNA-seq data. 10 of the 11 SNVs were validated using Sequenom or PCR (Supplementary Table 2 and Supplementary Fig. 8). All 623 somatic SNVs and 41 small indels detected by CASpoint and Samtools respectively (Supplementary Table 10). For structural variations, CAScnv (an in-house software) was used to find 3 somatic deletions, and combined CASfus and CASbreak (both in-house softwares) detected 2 somatic gene fusion events that were verified by mRNA-seq data (Supplementary Fig. 4,6,7).

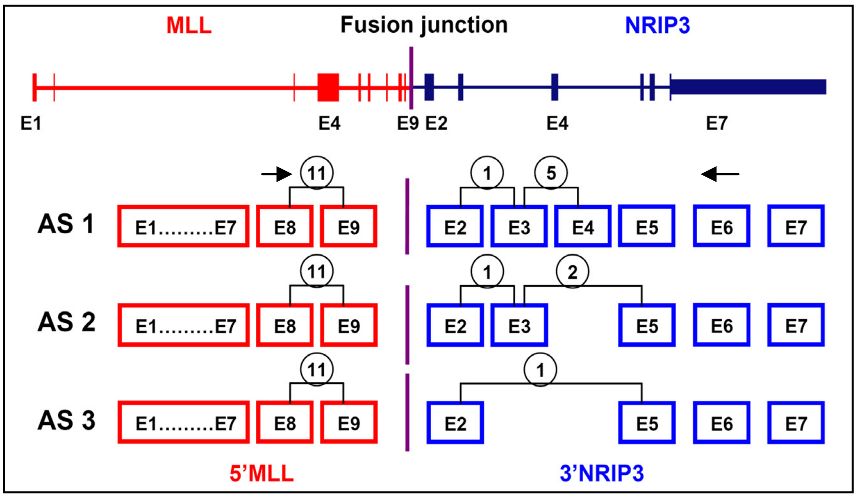
Supplementary Figure 4. Detection and validation of *MLL-NRIP3* fusion gene



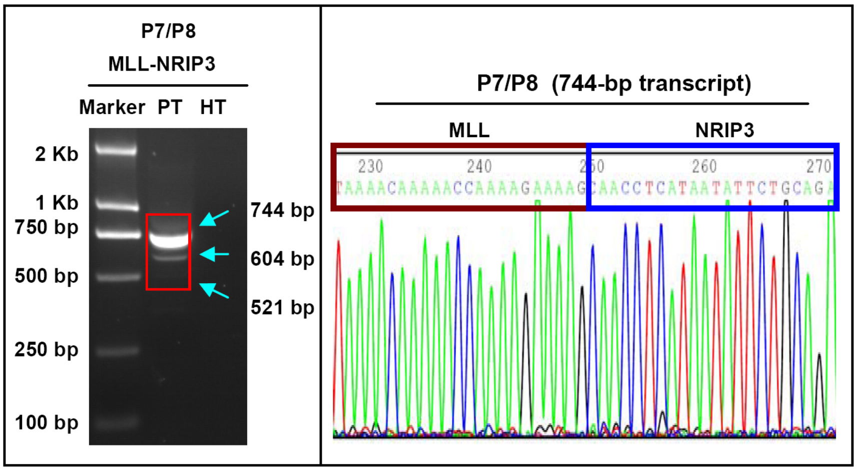
a. Complex chromosomal translocation on chr11 resulting in *MLL-NRIP3* fusion gene. 5 black vertical lines represent 5 chromosomal breakpoints (9010828, 9011043, 9038985, 9039415, 118354698; based on GRCh37/hg19) predicted by CASfus and CASbreak from WGS data. Red, green, and blue arrows represent the transcription orientation of *NRIP3*, *SCUBE2*, and *MLL*, respectively. Three pairs of black head-to-head arrows (P1/P2, P3/P4, P5/P6) represent primers designed for PCR validating chromosomal rearrangements. Each number above the purple cross-bar shows the number of mate-pair reads supporting the fusion junction.



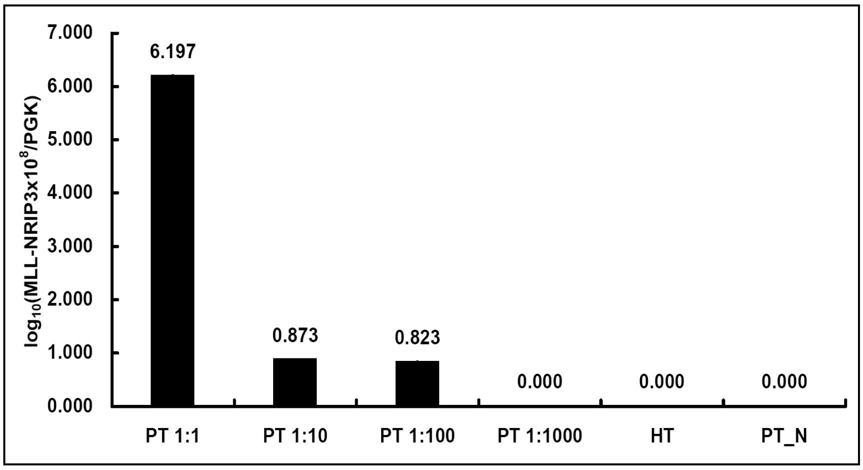
b. PCR-sequencing validation of 3 fusion junctions at the DNA level. By using primers spanning the junction of *MLL*_Intergenic (P1/P2), *NRIP3*_Intergenic (P3/P4), and *MLL_NRIP3* (P5/P6), PCR-sequencing was performed for PT, HT, U937 (non-*MLL*-rearranged leukemia cell line), and H_2O (negative control), respectively. The agarose gel electrophoresis assays (top panel) and their corresponding sequencing traces (lower panel) are shown, respectively.



c. Expression and alternative splicing of *MLL-NRIP3* fusion transcripts predicted by mRNA-seq. Gene structure of *MLL-NRIP3* fusion comprises 5'MLL (exon 1-9 of *MLL*, red), and 3'NRIP3 (exon 2-7 of *NRIP3*, blue), in which the purple vertical line represents the fusion junction of *MLL-NRIP3*. 1 pair of black arrows (P7/P8) represents the primer designed for PCR validation of *MLL-NRIP3* fusion and 3 alternatively spliced fusion transcripts (AS1, AS2, AS3) predicted by mRNA-seq. The numbers of read pairs supporting junction-joining are shown in circles.

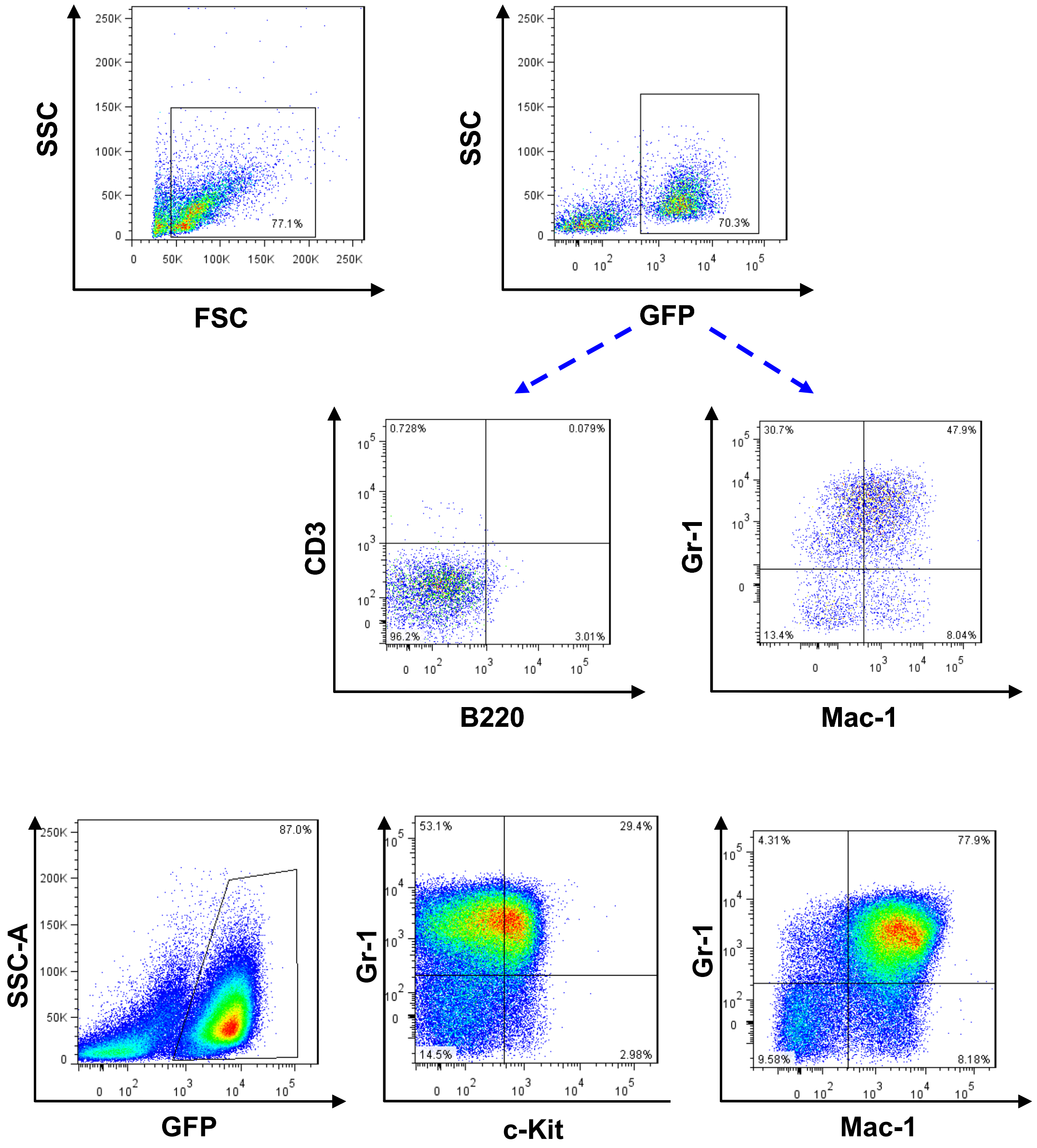


d. RT-PCR sequencing validating expression and alternative splicing of *MLL-NRIP3*. RT-PCR sequencing was performed for PT and HT with a forward primer (P7) in *MLL* exon 8 and a reverse primer (P8) in *NRIP3* exon 6. The agarose gel electrophoresis assay and its corresponding sequencing trace are shown, respectively. Three bands pointed to by 744, 604, and 521 bp represent AS1, AS2, and AS3, respectively.



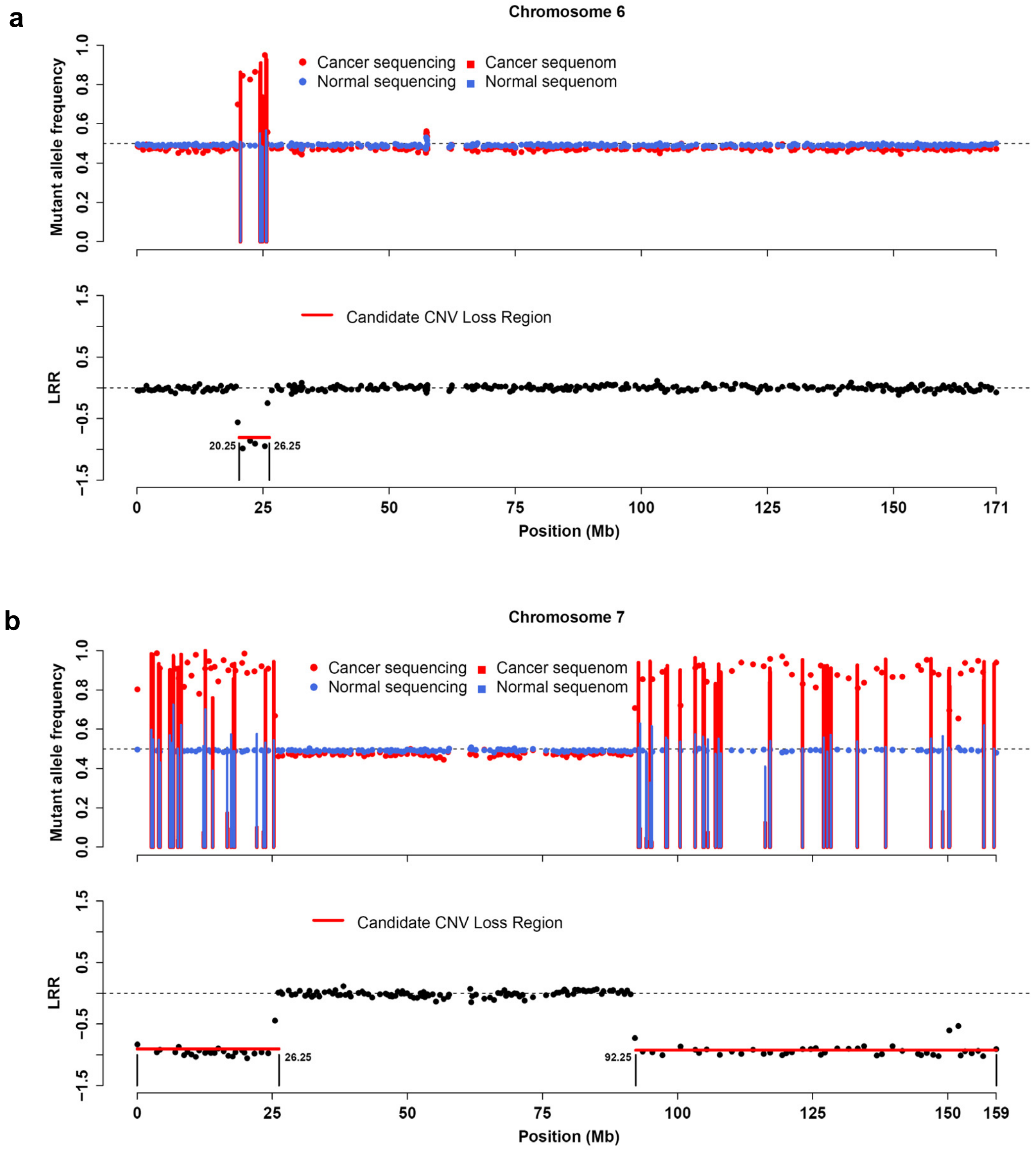
e. TaqMan detection for *MLL-NRIP3* expression. Expression level of *MLL-NRIP3* was detected by TaqMan in serial diluted RNA samples from CD56+CD64+ leukemia cells of PT, RNA samples from CD56-CD64- normal cells of PT (PT_N) and normal PBMCs of HT. Expression level of *MLL-NRIP3* was normalized against *PGK1* as $\log_{10}$(*MLL-NRIP3*×10^9/*PGK1*).

Supplementary Figure 5. Immunophenotyping of spleen cells from *MLL-NRIP3* induced mouse AML



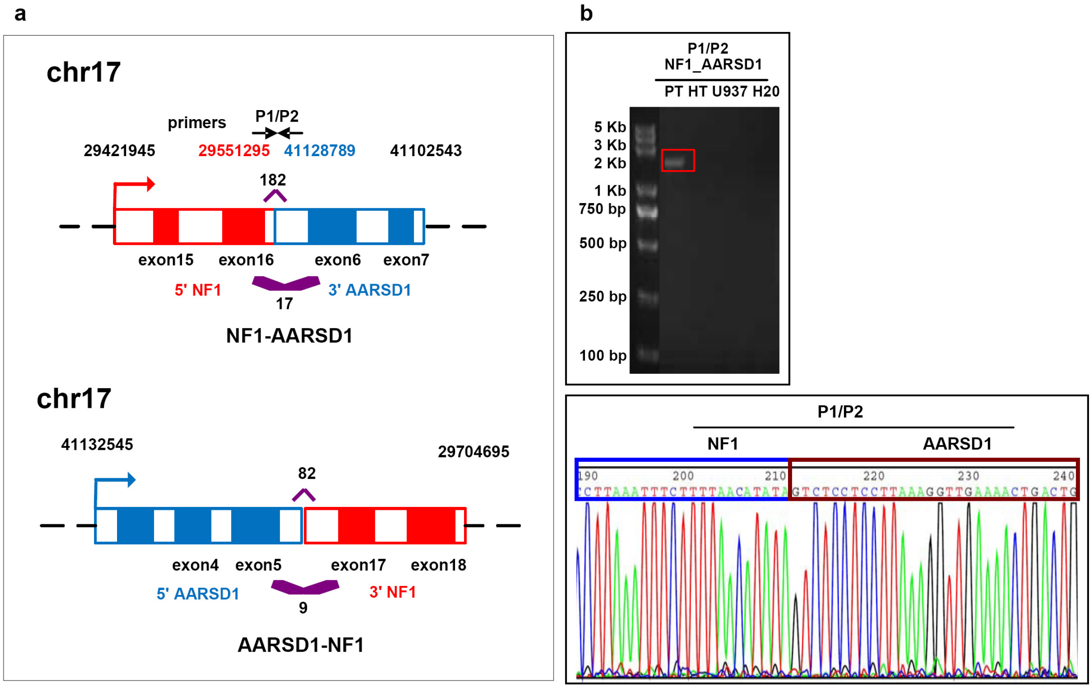
Immunophenotyping of spleen cells using flow cytometry. The spleen cells from the *MLL-NRIP3* leukemia mouse expressed cell surface markers representing monocytic leukemia.

Supplementary Figure 6. Detection and validation of copy number variations (CNVs)



Discovery and validation of CNVs in chr6 (a) and chr7 (b) are shown. Each sub-figure includes MAF (mutant allele frequency, top panel) and LRR (log2 ratio of read depth, lower panel). In each MAF graph, circles and squares represent MAF obtained from sequencing data and Sequenom data, respectively; leukemia and normal samples are colored in red and blue, respectively. In each LRR graph, the red lines represent candidate CNV loss regions predicted by the WGS data; regions with identical sequencing coverage between PT_Leukemia (Cancer) and HT (Normal) are shown as a horizontal dashed line for which LRR equals 0.

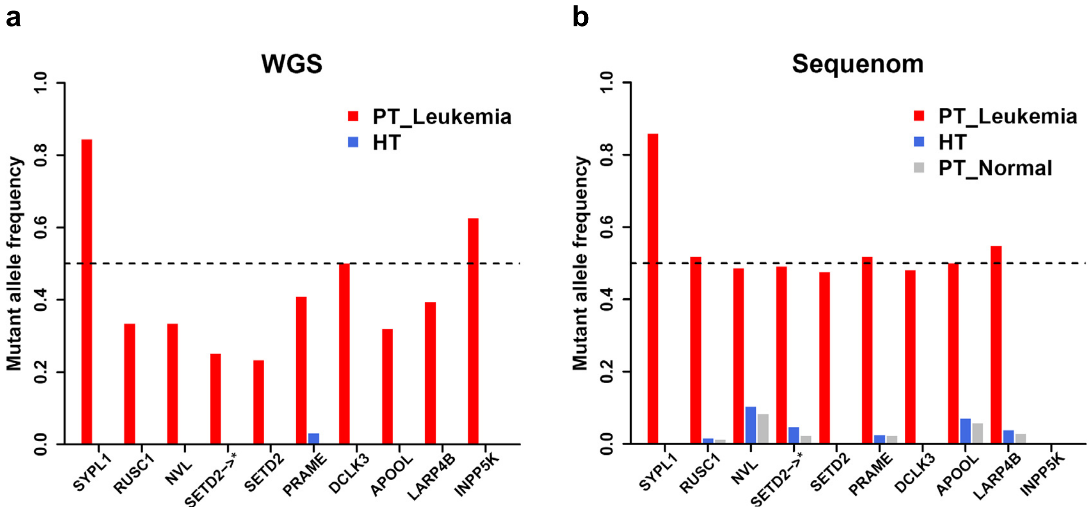
Supplementary Figure 7. Detection and validation of *NF1-AARSD1* fusion gene



a. Complex chromosomal translocation on chr17 resulting in *NF1-AARSD1* and *AARSD1-NF1* fusion genes. Red and blue arrows represent transcription orientation of *NF1-AARSD1* and *AARSD1-NF1* fusion genes, respectively. The black head-to-head arrows (P1/P2) represent primers designed for PCR validation of the *NF1-AARSD1* fusion junction. Each number above and below the purple cross-bar shows the number of mate-pair WGS reads and paired-end mRNA-seq reads supporting the fusion junction, respectively.

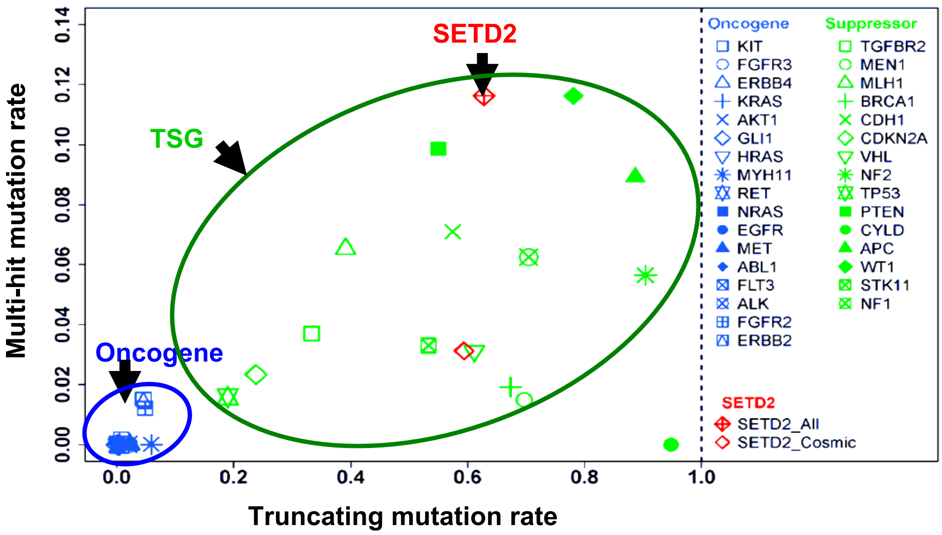
b. PCR-sequencing validation of the *NF1-AARSD1* fusion junction at the DNA level. By using primers spanning the junction of *NF1-AARSD1* (P1/P2), PCR-sequencing was performed for PT, HT, U937, and H_2O, respectively. The agarose gel electrophoresis assay (top panel) and its corresponding sequencing trace (lower panel) are shown, respectively.

Supplementary Figure 8. Mutation frequency of somatic SNVs identified by WGS and Sequenom



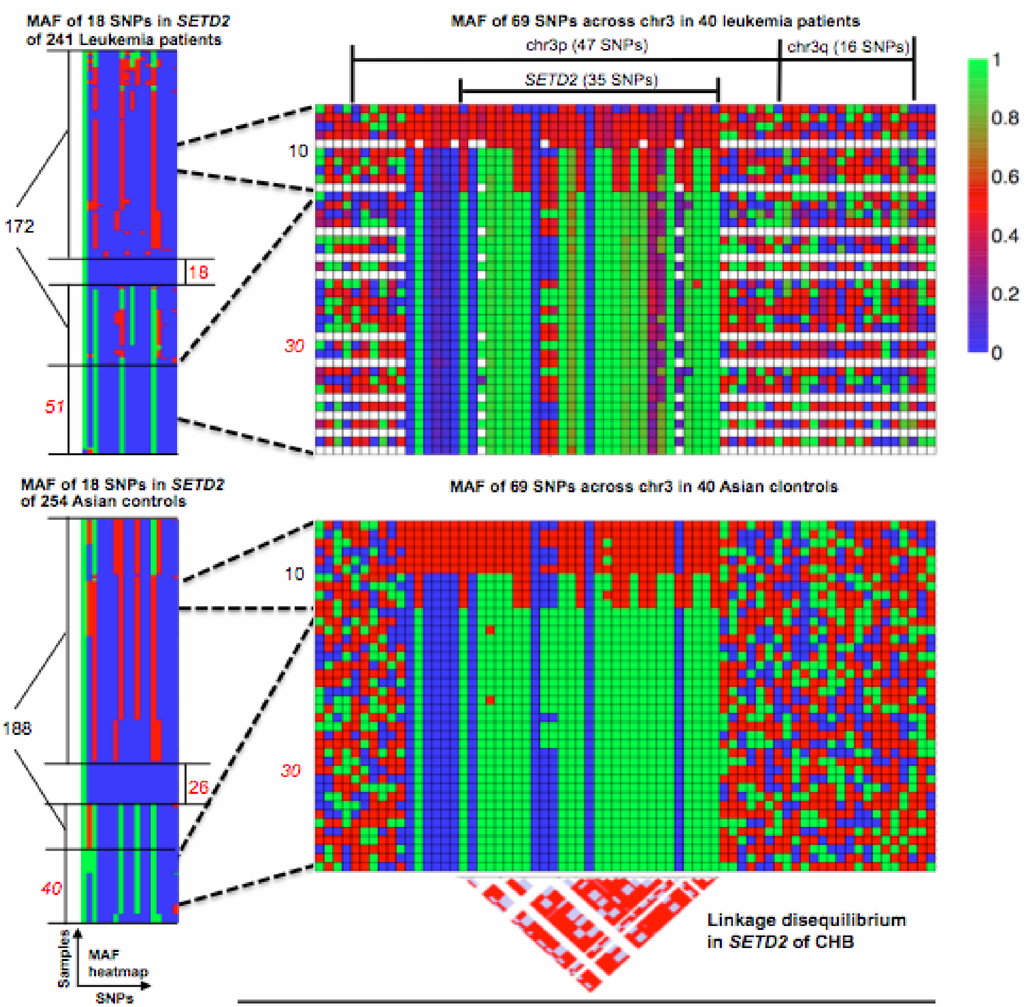
Bar-plots show the mutation frequencies of somatic SNVs as determined by WGS (a), and Sequenom (b), respectively. The horizontal dashed lines represent a mutation frequency of 50%. PT_Leukemia: CD56+CD64+ leukemic cells from PT. PT_Normal: CD56-CD64- normal cells from PT. HT: PBMCs from HT.

Supplementary Figure 9. Mutation spectrum of *SETD2* and cancer causative genes in cancer patients



Gene truncating mutation rate (F_T) is plotted against gene multi-hit mutation rate (F_{MH}) (see Online Methods). Mutational information for each gene was retrieved from the COSMIC[7] database (Release 65). SETD2_Cosmic: *SETD2* mutations reported in the COSMIC database. SETD2_All: *SETD2* mutations reported in COSMIC database, the leukemia study by Zhang J et al[8] and this study. Tumor suppressor genes (TSGs) and oncogenes are circled by green and blue lines, respectively. *SETD2* is pointed by an arrow.

Supplementary Figure 10. ***SETD2*** **inactivation in acute leukemia patients**



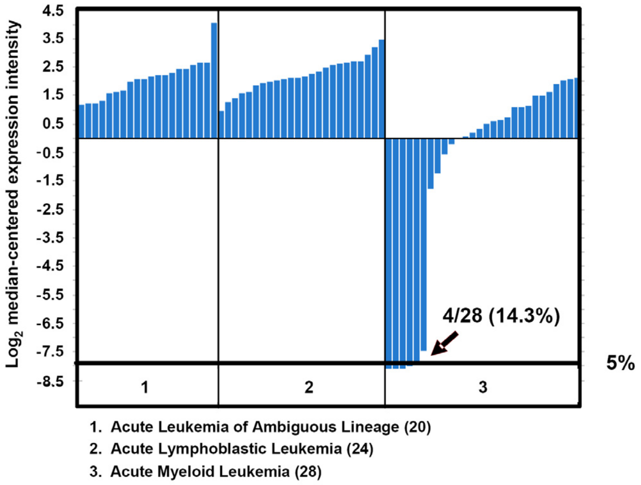
Neutral LOH testing	Leukemia patients	Asian controls	P-value
Candidates with LOH	69 (18+*51*)	66 (26+*40*)	1
Candidates without LOH	172	188	

a. No *SETD2* deletion detected in the leukemia cohort of current study. Top and lower left panels show heatmaps of mutant allele frequencies (MAF) of 18 SNPs obtained from either 241 leukemia patients (top left) or 254 Asian controls in 1000 genomes (lower left). The numbers of candidates with *SETD2* LOH are shown in red: regular (non-italic) numbers indicate candidates with MAF of about 0 or 1; italic numbers indicate candidates with MAF of about 0 in all 18 SNP sites. Numbers shown in black indicate candidates without *SETD2* LOH. Top and lower right panels show heatmaps of MAF of 69 SNP across the chr3p in 40 leukemia patients (top right) and 40 Asian controls (lower right). For both 241 leukemia patients and 254 Asian controls, 10 and 30 samples were selected from candidates without LOH and that with LOHs, respectively. MAF of 69 SNPs in 40 leukemia patients and 40 Asian patients were obtained from Sequenom genotyping and 1000 human genomes polymorphism dataset[6], respectively. The gradient change of green, red, and blue in all four heatmaps represents MAF ranging from 1 to 0. Two linkage disequilibrium blocks across *SETD2* locus among Chinese population (http://hapmap.ncbi.nlm.nih.gov/cgi-perl/gbrowse/hapmap24_B36/#search) are shown under the lower right MAF heatmap. The bottom table shows the neutral LOH testing which compares the number of candidate with LOH observed in 241 leukemia patients (69=51+18) with that (66=26+40) in 254 Asian controls (Fisher testing: *p* value=1).

b

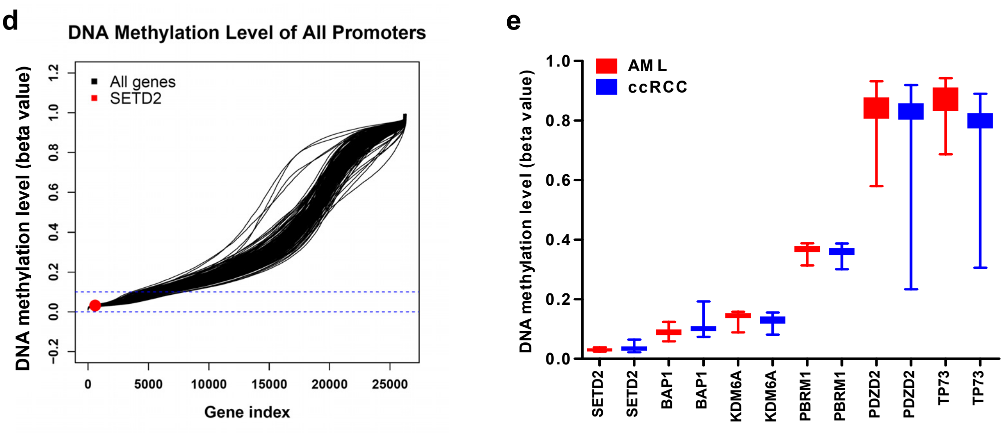
dataset	t-test (p<0.05, FC>2) (cancer vs normal)		COPA (top 25%) (outlier)	
	up	down	up	down (ratio)
Leukemia (35)	0	0	7	14 (0.0-20.0%, 9.0%)
Lymphoma (14)	0	6	3	8 (5.5-30.2%, 17.0%)
Myeloma (9)	0	0	1	1 (26.9%)
Significant Unique Analysis	0	6	12	24

c



b. *SETD2* downexpression in a subset of patients with leukemia and other hematopoietic malignancies. Analytical tools incorporated in Oncomine[9] were used to generate a summary of *SETD2* expression profile in leukemia, lymphoma and myeloma. Cancer Outlier Profile Analysis (COPA)[10] showed that *SETD2* has significant downregulated outlier profile in 14 leukemia, 8 lymphoma, and 1 myeloma datasets. The ranged and median fractions of patients with *SETD2* outlier profile in all datasets are shown.

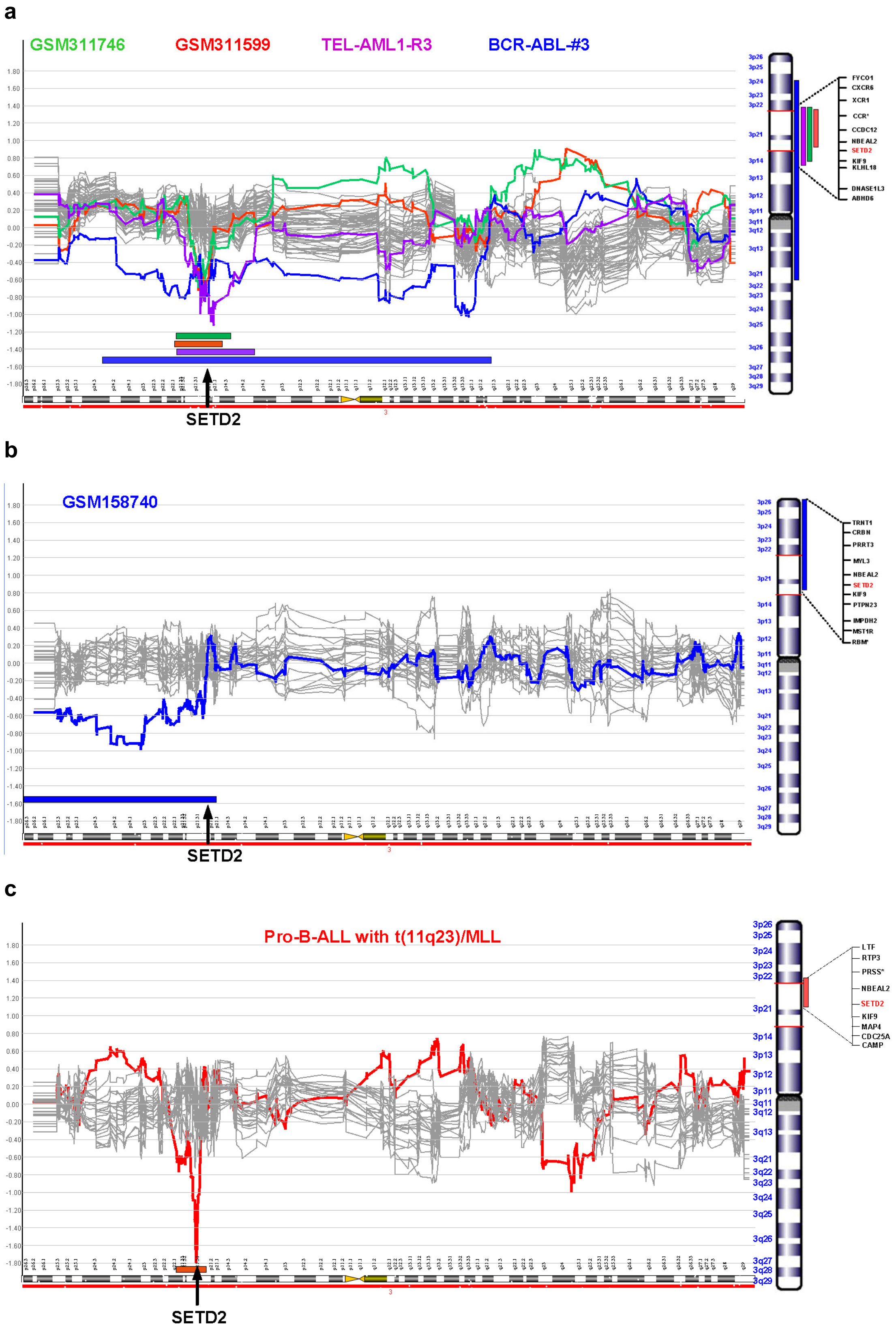
c. Expression suppression identified in a subset of leukemia patients based on Oncomine database. COPA revealed *SETD2* as a gene with outlier expression profile at the 95*th* percentile in Armstrong et al.'s acute leukemia dataset (n=72)[11]. *SETD2* expression is shown from all profiled samples in this dataset. Arrows point to 4 AML samples with under expression of *SETD2* among 72 patients examined. Visualization tools incorporated in Oncomine were used to generate graphical displays. The microarray data indicate that *SETD2* is underexpressed in a subset of AML samples (4/28).

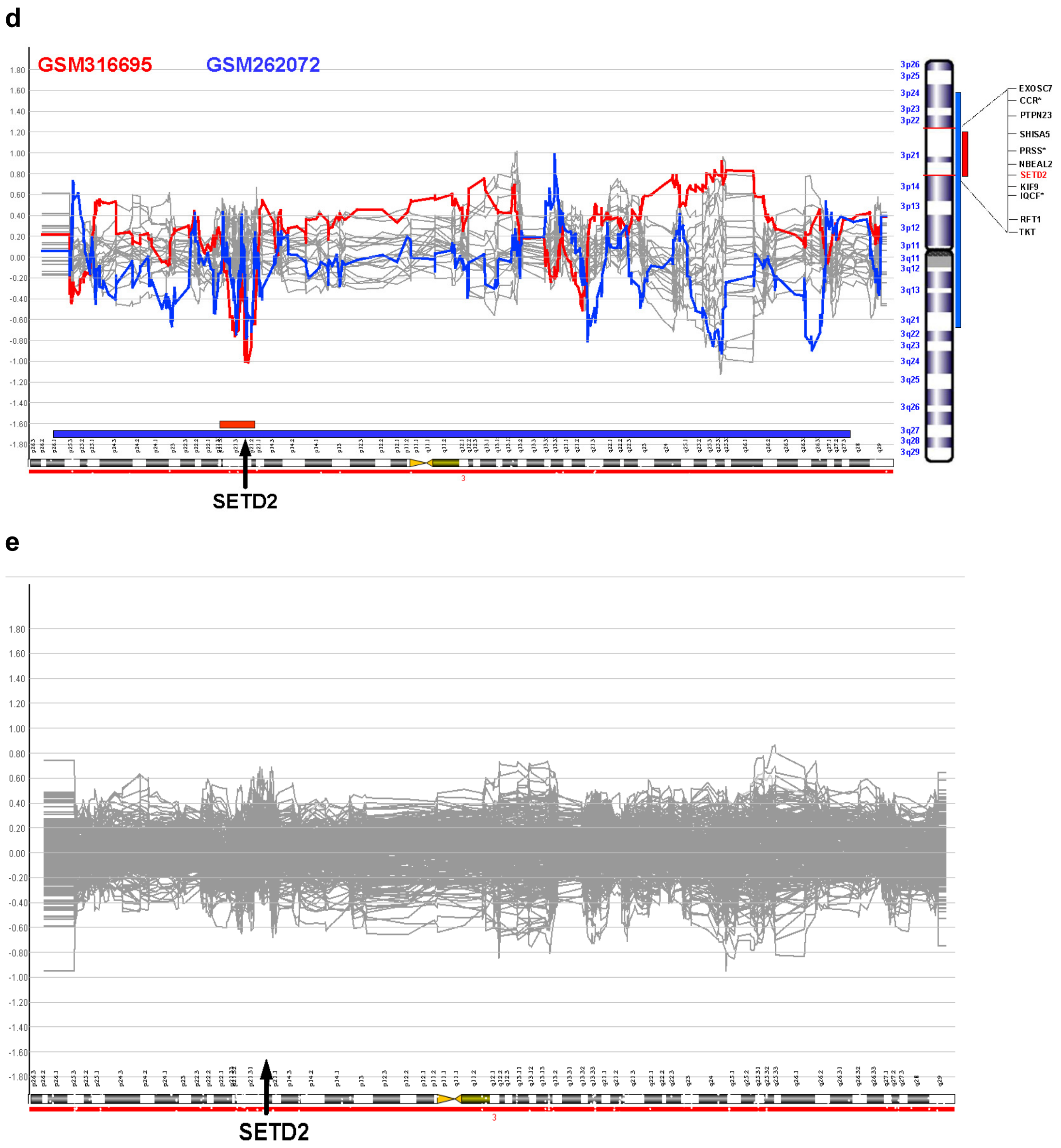


d. Promoter of DNA methylation levels of all genes in 194 AML patients are shown as lines in increasing order, with each line representing one patient and x axis representing gene index. Horizontal blue dashed lines represent DNA methylation level with beta value of 0 and 0.1. The levels of the *SETD2* gene are indicated.

e. Box-plots showing distributions of the DNA methylation level of *SETD2* promoter in AML and ccRCC patients. Medians are denoted by solid lines while the top and bottom box edges denote the first and third quartile. Whiskers denote the largest and smallest data. The level of promoter DNA methylation is shown as beta value, which is defined as (intensity of methylated probe)/(total intensity of methylated and unmethylated probes).

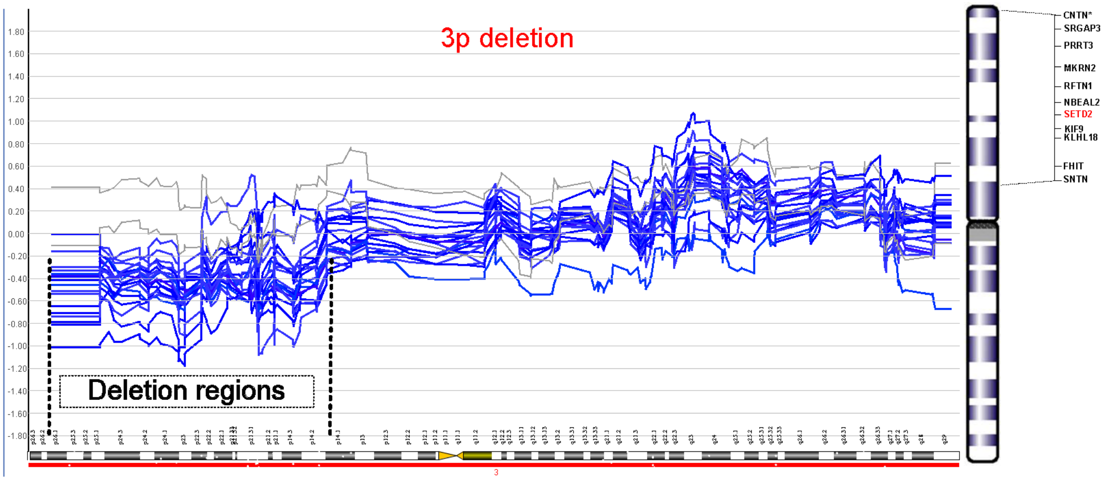
Supplementary Figure 11. Large deletion involving *SETD2* in acute leukemia patients





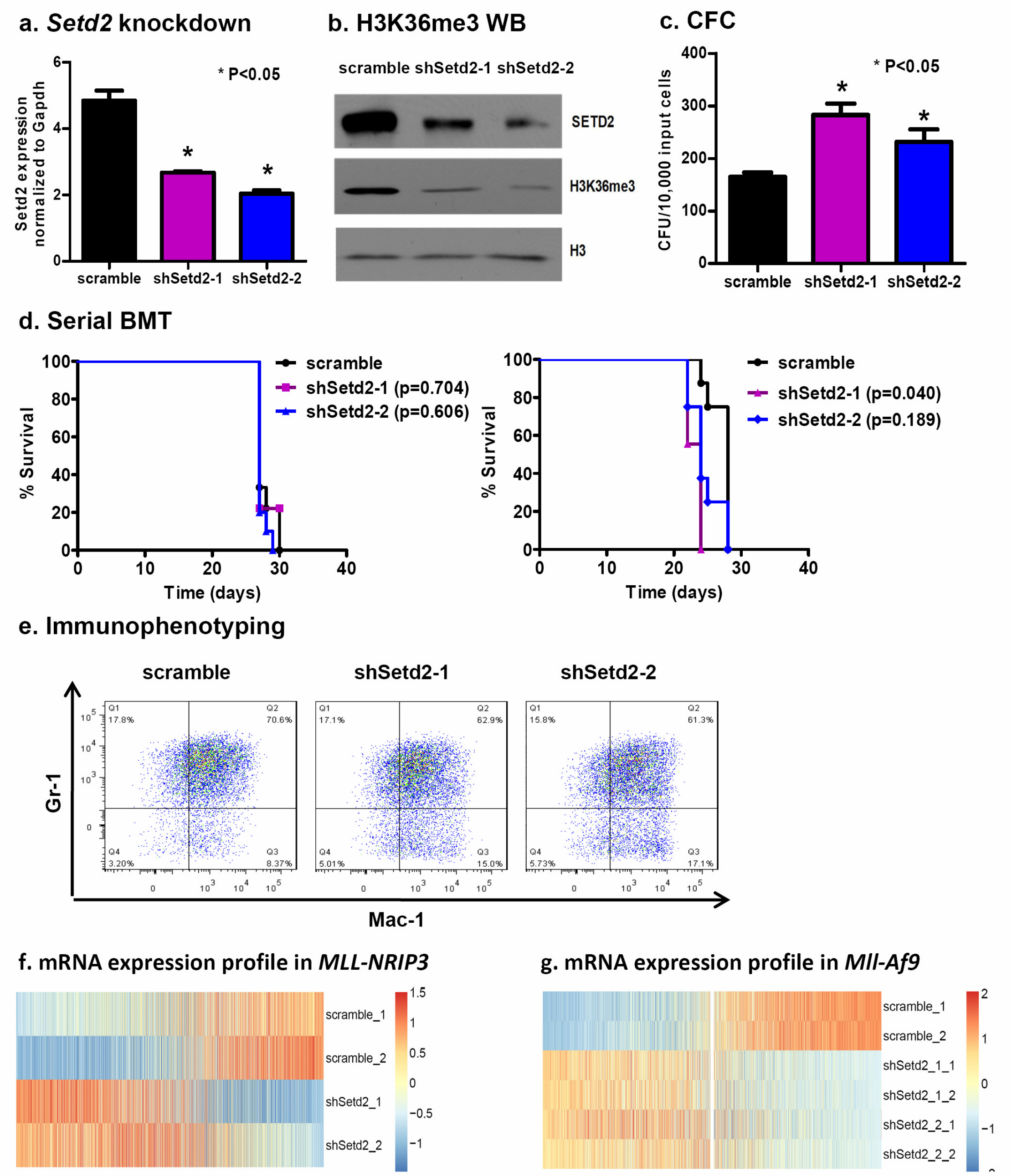
Large deletion involving the *SETD2* gene was detected in acute leukemia patients based on mRNA expression array derived from multiple datasets (in seven cohorts). The gene expression profile moving average plots demonstrate large deletions detected along chromosome 3: a.ROSS2003BLOOD, ROSS2004BLOOD, and GSE12417; b.GSE6891; c.GSE13159; d.GSE10358 and GSE12417; e.GSE1159 respectively. Horizontal colored bars indicate the detected significant chromosomal deletions. Vertical colored bars beside chromosomal ideogram indicate regions of detected deletions. Colored lines show patients have detected *SETD2* large deletion and grey lines show patients don't have detected *SETD2* large deletion. *SETD2* gene locus is pointed by an arrow. The method for detecting large deletion from expression array data referred the pipeline in Mayshar et al's study[12,13].

Supplementary Figure 12. Large deletion involving *SETD2* in ccRCC patients



Gene expression profile moving average plot of mRNA expression array data derived from a Clear Cell Renal Cell Carcinoma patient cohort[14]. Detected large deletions in chromosome 3 are shown and labeled. Vertical black dashed lines represent general boundaries of chromosomal deletions. Blue lines represent patients with detected large deletions and grey lines show patients without detected large deletions. The method for detecting large deletion from expression array data referred the pipeline in Mayshar et al's studies[12,13].

Supplementary Figure 13. Two different shRNAs have a similar effect on *Setd2* expression and leukemia transformation



a. Two different shRNAs efficiently knock *Setd2* down (mean±s.d., n=2).

b. *Setd2* knockdown by two shRNAs displays similar decrease of H3K36me3 level.

c. *Setd2* knockdown in *Mll-Af9* knock-in HSPCs yields a significantly higher number of total colonies in CFC assay (mean±s.d., n=3).

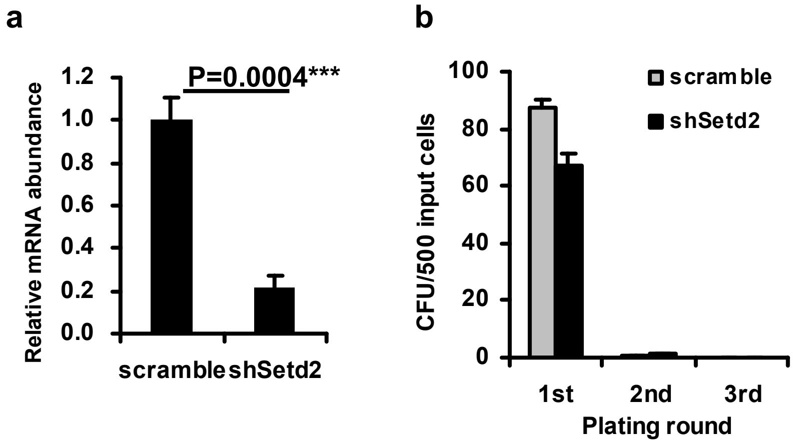
d. *Setd2* knockdown accelerates *MLL-NRIP3* leukemia in serial BMT assays (n≥8 for each group).

e. *Setd2* knockdown has no effect on immunophenotyping of bone marrow cells from *MLL-NRIP3*-induced leukemia mouse as detected by flow cytometry.

f. Similar mRNA expression profiles in *Setd2* knockdown *MLL-NRIP3* leukemia by two different shRNAs.

g. Similar mRNA expression profiles in *Setd2* knockdown *Mll-Af9* leukemia by two different shRNAs.

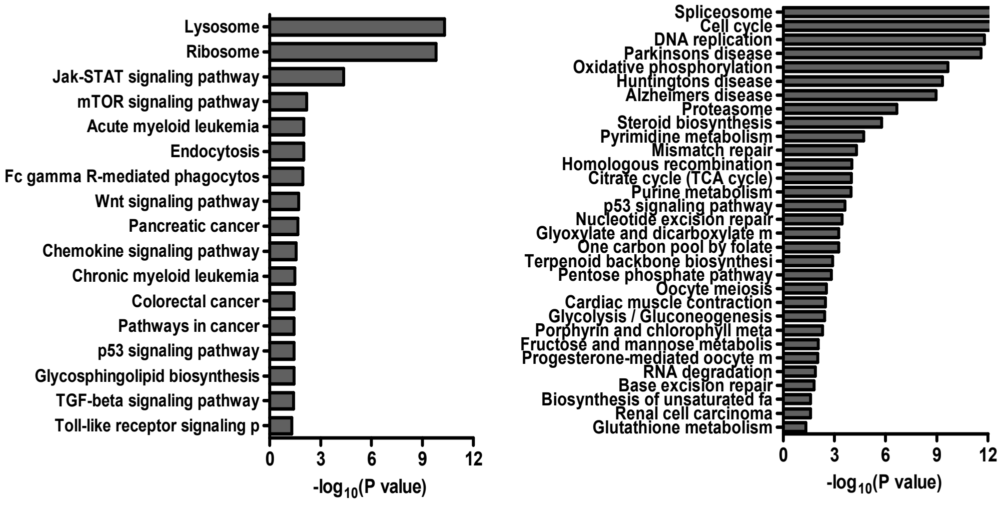
Supplementary Figure 14. *Setd2* knockdown in normal hematopoietic stem and progenitor cells (HSPCs)



a. Knockdown of *Setd2* in normal HSPCs derived from normal mice. Bar-plot represents *Setd2* expression (mean±s.d., n=3) in normal HSPCs transduced with *Setd2* shRNA (shSetd2) or the scrambled control shRNA (scramble).

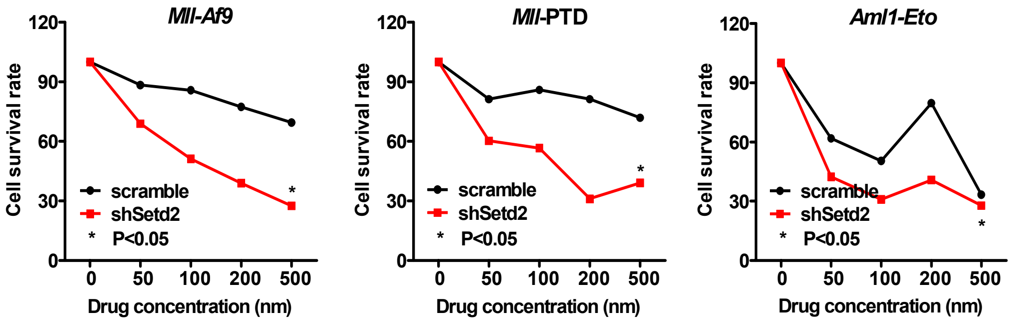
b. Serial CFU replating assays for normal HSPCs derived from normal mice. Cells were transduced with either *Setd2* shRNA (shSetd2) or the control scrambled shRNA (scramble). The plating round and the number of CFU (mean±s.d., n=3) per 10,000 input cells are shown.

Supplementary Figure 15. Functional categories enriched among differentially expressed genes upon *Setd2* knockdown in *Mll-Af9* cells



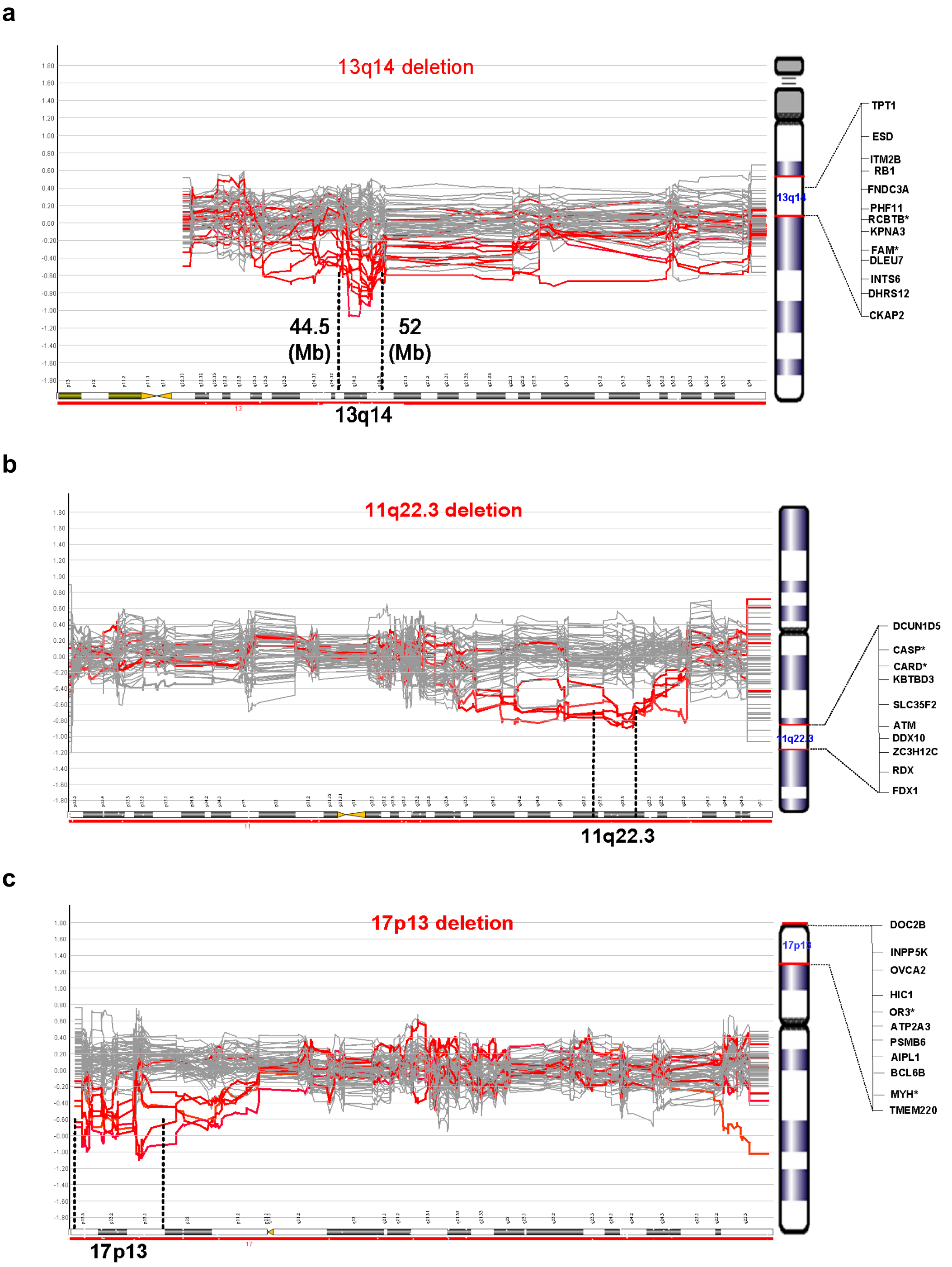
KEGG pathway enrichment analysis was performed on differentially upregulated or downregulated genes using DAVID Bioinformatics resources[15]. Each bar represents a significantly enriched pathway as determined using the *p* value < 0.05. The x axis shows -log$_{10}$-transformed expected *p* values.

Supplementary Figure 16. Rapamycin inhibits cell growth in *Setd2* knockdown pre-leukemic cells



Increased sensitivity to mTOR inhibition in *Setd2* knockdown HSPCs isolated from *Mll-Af9*, *Mll*-PTD, and *Aml1-Eto* knock-in mice. *Setd2* knockdown or scrambled shRNA treated cells were plated. mTOR inhibitor Rapamycin was added at indicated concentrations.

Supplementary Figure 17. Validation of the genomic analysis method for detecting large DNA deletions



Large deletions were detected using the method of genomic analysis for the mRNA expression array data from a chronic lymphocytic leukemia patient cohort[16].The genomic analysis of deletion detection with mRNA expression array was established by Mayshar et al[12,13]. Detected large deletions are shown and labeled as: a.13q14 deletion; b.11q22.3 deletion; c.17p13 deletion. Vertical black dashed lines represent general boundaries of chromosomal deletions. Red lines represent patients with detected large deletions and grey lines show patients without detected large deletions.

Supplementary Tables

Supplementary Table 1. Data summary of WGS and mRNA-seq

Sample	Sample Type	Library	Library size	Sequencer	Reads Length (bp)	Total bases (Gb)	Aligned reads (M)	Aligned Rate (%)	Depth	Q20 coverage (%)
PT_Leukemia	DNA	MP[¶]	1.5 kb	SOLID 4.0	2×50	164	2723	83.07	54.6×	98.48
HT	DNA	MP	1.5 kb	SOLID 4.0	2×50	187	3192	85.29	62.4×	98.46
PT_Leukemia	mRNA	PE[§]	400 bp	GA IIx	2×81	25.35	270	86.18	-	-
PT_Normal	mRNA	PE	400 bp	GA IIx	2×81	18.06	200	89.60	-	-
HT	mRNA	PE	400 bp	GA IIx	2×81	17.82	202	91.75	-	-
MA9[£]_scramble_rep1	mRNA	SE[£]	325 bp	HiSeq-2000	50	0.62	9.45	76.16	-	-
MA9_scramble_rep2	mRNA	SE	325 bp	HiSeq-2000	50	0.94	14.13	74.79	-	-
MA9_shSetd2_sh1_rep1	mRNA	SE	325 bp	HiSeq-2000	50	1.04	15.38	74.18	-	-
MA9_shSetd2_sh1_rep2	mRNA	SE	325 bp	HiSeq-2000	50	1.08	15.38	71.11	-	-
MA9_shSetd2_sh2_rep1	mRNA	SE	325 bp	HiSeq-2000	50	0.41	5.79	69.93	-	-
MA9_shSetd2_sh2_rep2	mRNA	SE	325 bp	HiSeq-2000	50	1.00	14.41	72.19	-	-
MN3[#]_scramble_rep1	mRNA	PE	250 bp	HiSeq-2000	2×101	11.40	92.82	82.25	-	-
MN3_scramble_rep2	mRNA	PE	250 bp	HiSeq-2000	2×101	9.97	83.50	84.58	-	-
MN3_shSetd2_sh1	mRNA	PE	250 bp	HiSeq-2000	2×101	7.85	64.77	83.34	-	-
MN3_shSetd2_sh2	mRNA	PE	250 bp	HiSeq-2000	2×101	7.01	58.04	83.62	-	-

[¶]MP: Mate-pair; [§]PE: Pair-end; [£]SE: Single-end; [£]MA9: *Mll-Af9*; [#]MN3: *MLL-NRIP3*

Supplementary Table 2. SNVs identified and validated in the leukemia twin patient

Gene	DNA nucleotide change	mRNA nucleotide change	Variant type	Amino acid change	Allele expression (reference#:variant#)			SIFT[¶]	CHASM[§]
					PT_Leukemia	HT	PT_Normal		
NVL	chr1:g.224477365T>G	NM_001243146.1:c.829A>C	missense	p.(Thr277Pro)	21:31	8:0	24:0	DAMAGING	0.438
DCLK3	chr3:g.36779306G>A	NM_033403.1:c.845C>T	missense	p.(Ala282Val)	0:0	0:0	0:0	DAMAGING	NA
SETD2	chr3:g.47059146A>T	NM_014159.6:c.7515T>A	missense	p.(Phe2505Leu)	56:39	44:0	16:0	DAMAGING	NA
SETD2	chr3:g.47088102G>A	NM_014159.6:c.6973C>T	nonsense	p.(Gln2325*)	26:21	45:0	125:0	NA	NA
SYPL1	chr7:g.105738276C>T	NM_006754.2:c.316G>A	missense	p.(Val106Ile)	8:126	88:0	209:0	TOLERANT	0.738
LARP4B	chr10:g.882360C>A	NM_015155.1:c.733G>T	nonsense	p.(Glu245*)	33:12	63:0	75:0	NA	NA
PRAME	chr22:g.22892500G>A	NM_206953.1:c.601C>T	nonsense	p.(Arg201*)	0:0	0:0	0:0	NA	NA
APOOL	chrX:g.84342619C>T	NM_198450.3:c.742C>T	missense	p.(Pro248Ser)	32:0	0:0	0:0	DAMAGING	0.76
INPP5K	chr17:g.1399611A>G	NM_001135642.1:c.957+2T>C	splice-site	NA	4:66	11:59	0:80	NA	NA
RUSC1	chr1:g.155296631G>A	NM_001105203.1:c.2122G>A	missense	p.(Gly708Arg)	32:40	18:0	25:0	TOLERANT	NA

[¶]SIFT[17] is a software which predicts whether an amino acid substitution affects protein function.

[§]CHASM[18] is a method that predicts the functional significance of somatic missense mutations observed in the genomes of cancer cells.

Supplementary Table 3. Clinical and Mutation information of all acute leukemia patients

See a separate supplementary file: TableS3.xlsx

Supplementary Table 4. Comparison of clinical and molecular characteristics by *SETD2* mutation status in AML patients

Variable	*Total*	*SETD2-mutated*	*SETD2-wild-type*	*P-value*¶
No. of patients	134	7	127	
Age (years)	28.5 (27.5)	34.8 (41.0)	28.2 (27.0)	0.459
Female	42	5 (11.9%)	37 (88.1%)	0.032 *
Male	91	2 (2.2%)	89 (97.8%)	
PB blast%	55.0 (61.0)	60.3 (56.0)	54.7 (61.0)	0.647
BM blast%	69.1 (73.2)	67.4 (60.0)	69.2 (73.5)	0.779
IM blast%	71.2 (78.8)	78.2 (90.9)	70.8 (77.5)	0.394
Hemoglobin (g/dL)	84.4 (84.0)	80.1 (89.0)	84.6 (83.5)	0.607
WBC (×10^9/L)	55.3 (22.9)	36.0 (36.1)	56.4 (22.6)	0.246
Platelet count (×10^9/L)	44.2 (31.0)	51.6 (21.0)	43.8 (31.5)	0.714
Chemotherapy effect (BM)	101	7	94	-
Complete remission (CR)	80 (79.2%)	4 (5.0%)	76 (95.0%)	0.155
Partial/Non-remission (PR+NR)	21 (20.8%)	3 (14.3%)	18 (85.7%)	
FAB category (n/%)	133	7	126	-
M0	1 (0.8%)	0	1	-
M1	3 (2.3%)	0	3	-
M2	38 (28.6%)	2 (5.3%)	36 (94.7%)	1
M3	16 (12.0%)	0	16	-
M4	14 (10.5%)	1 (7.1%)	13 (92.9%)	0.550
M5	55 (41.4%)	4 (7.3%)	51 (92.7%)	0.447
M6	2 (1.5%)	0	2	-
M7	2 (1.5%)	0	2	-
Missing data	2 (1.5%)	0	3	-
Cytogenetic characteristics				
t(11q23)/*MLL*				
MLL-rearranged	18 (13.4%)	4 (22.2%)	14 (77.8%)	0.010 *
MLL wild type	79 (59.0%)	2 (2.5%)	77 (97.5%)	
t(15;17)/*PML-RARA*				
Positive	16 (11.9%)	0 (0.0%)	16 (100.0%)	0.579
Negative	70 (52.2%)	5 (7.1%)	65 (92.9%)	
t(8;21)/*AML1-ETO*				
Positive	21 (15.7%)	1 (4.8%)	20 (95.2%)	1
Negative	92 (68.7%)	5 (5.4%)	87 (94.6%)	
inv(16)/t(16;16)/*CBFβ-MYH11*				
Positive	8 (6.0%)	0 (0.0%)	8 (100.0%)	1
Negative	88 (65.7%)	5 (5.7%)	83 (94.3%)	
NPM1				
NPM1-mutated	3 (2.2%)	1 (33.3%)	2 (66.7%)	0.125
NPM1-wild type	21 (15.7%)	0 (0.0%)	21 (100.0%)	
***FLT3*-ITD**				
Positive	18 (13.4%)	1 (5.6%)	17 (94.4%)	0.492
Negative	99 (73.9%)	3 (3.0%)	96 (97.0%)	
***FLT3*-TKD**				
Positive	5 (3.7%)	1 (20.0%)	4 (80.0%)	0.204
Negative	109 (81.3%)	4 (3.7%)	105 (96.3%)	
***MDR1* (positive)**	30 (22.4%)	2 (6.7%)	28 (93.3%)	-

¶P-value was calculated using the different test methods, such as Fisher test and t-test (two-sided)

* represents statistics significance of differences between *SETD2-mutated* and *SETD2-wild* type patients (*: P < 0.05).

Supplementary Table 5. Comparison of clinical and molecular characteristics by *SETD2* mutation status in ALL patients

Variable	*Total*	*SETD2-mutated*	*SETD2-wild type*	*P-value*[¶]
No. of patients	107	8	99	
Age (years)	5.7 (4.0)	12.6 (9.0)	5.2 (4.0)	0.255
Female	45	3 (6.7%)	42 (93.3%)	1
Male	62	5 (8.1%)	57 (91.9%)	
PB blast%	46.6 (44.0)	67.1 (61.5)	44.5 (42)	0.004 **
BM blast%	86.4 (90.5)	87.1 (89.0)	86.3 (90.5)	0.814
IM blast%	80.5 (84.5)	80.6 (81.3)	80.5 (84.6)	0.983
Hemoglobin (g/dL)	78.4 (78.5)	74.9 (75.5)	78.6 (78.5)	0.750
WBC ($\times 10^9$/L)	56.4 (12.5)	93.0 (48.6)	53.4 (12.1)	0.306
Platelet count ($\times 10^9$/L)	66.3 (47.0)	55.8 (46.5)	67.2 (47.0)	0.418
Chemotherapy effect (BM)	95	8	87	
Complete remission (CR)	93 (97.9%)	8 (8.6%)	85 (91.4%)	1
Partial/Non-remission (PR+NR)	2 (2.1%)	0 (0.0%)	2 (100.0%)	
FAB category (n/%)	107	8	99	
B-lineage	93 (86.9%)	5	88	0.068
T-lineage	12 (11.2%)	2	10	0.220
Missing data	2 (1.9%))	1	1	0.145
Diagnosis	99	6	93	
Low-risk	36 (36.4%)	3	33	0.665
Moderate-risk	32 (32.3%)	1	31	0.661
High-risk	31 (31.3%)	2	29	1
Cytogenetic characteristics				
t(11q23)/*MLL*				
MLL-rearranged	9 (8.4%)	2 (22.2%)	7 (77.8%)	0.144
MLL wild type	94 (87.9%)	6 (6.4%)	88 (93.6%)	
t(9;22)/*BCR-ABL*				
Positive	10 (9.3%)	1 (10.0%)	9 (90.0%)	0.557
Negative	97 (90.7%)	7 (7.2%)	90 (92.8%)	
t(12;21)/*TEL-AML1*				
Positive	26 (24.3%)	1 (3.8%)	25 (96.2%)	0.676
Negative	81 (75.7%)	7 (8.6%)	74 (91.4%)	
TCR				
Positive	30 (28.0%)	3 (10.0%)	27 (90.0%)	0.680
Negative	61 (57.0%)	4 (6.6%)	57 (93.4%)	
IGH/IGK				
Positive	50 (46.7%)	2 (4.0%)	48 (96.0%)	0.124
Negative	36 (33.6%)	5 (13.9%)	31 (86.1%)	
***MDR1* (positive)**	19 (17.8%)	2 (10.5%)	17 (89.5%)	

[¶] P-value was calculated using the different test methods, such as Fisher test and t-test (two-sided).

* represents statistics significance of differences between *SETD2-mutated* and *SETD2-wild* type patients (**: *P* < 0.01).

Supplementary Table 6. Comparison of clinical and molecular characteristics by *SETD2* mutation status in all acute leukemia patients

Variable	*Total*	*SETD2-mutated*	*SETD2-wild type*	*P-value*¶
No. of patients	241	15	226	
Age (years)	18.4 (10.0)	23.7 (10.5)	18.1 (10.0)	0.355
Female	87	8 (9.2%)	79 (90.8%)	0.173
Male	153	7 (4.6%)	146 (95.4%)	
PB blast%	51.5 (55.0)	64.0 (60.0)	50.6 (54.5)	0.050
BM blast%	76.6 (82.0)	77.9 (82.0)	76.5 (82.0)	0.739
IM blast%	75.3 (82.7)	79.4 (83.1)	75.1 (82.7)	0.365
Hemoglobin (g/dL)	81.7 (82.0)	77.3 (82.0)	82.0 (82.0)	0.514
WBC ($\times 10^9$/L)	55.8 (16.6)	66.4 (39.9)	55.1 (16.20)	0.605
Platelet count ($\times 10^9$/L)	54.0 (37.0)	57.2 (52.0)	53.8 (36.5)	0.324
Chemotherapy effect (BM)	196	15	181	
Complete remission (CR)	173 (88.3%)	12 (6.9%)	161 (93.1%)	0.393
Partial/Non-remission (PR+NR)	23 (11.7%)	3 (13.0%)	20 (87.0%)	
FAB category (n/%)	241	15	226	
AML	134	7	127	0.594
ALL	107	8	99	
Cytogenetic characteristics				
t(11q23)/*MLL*				
MLL-rearranged	**27 (11.2%)**	**6 (22.2%)**	**21 (77.8%)**	**0.005 ****
MLL wild type	**173 (71.8%)**	**8 (4.6%)**	**165 (95.4%)**	
***MDR1* (positive)**	49 (20.3%)	4 (8.2%)	45 (91.8%)	

¶P-value was calculated using the different test methods, such as Fisher test and t-test (two-sided).

* represents statistics significance of differences between *SETD2-mutated* and *SETD2-wild* type patients (**: $P < 0.01$).

Supplementary Table 7. Molecular characteristics of the *SETD2* deletion pattern

Dataset (excluding complex karyotype)	**# total cases**	**# cases with *SETD2* deletion (genetic abnormalities in *SETD2*-deleted cases)**		**Percentage**
Ross2003BLOOD+Ross2004BLOOD +GSE12417 (ALL/AML)	450	4	Cytogeneticly normal: 2 (289) *BCR-ABL*: 1 (15) *TEL-AML1*: 1 (20)	0.0054
GSE6891 (AML)	523	1	-5/7(q): 1 (28)	0.0019
GSE13159 (AML/ALL/Normal)	568	1	Pro-B-ALL with t(11q23)/*MLL*: 1 (70)	0.0018
GSE10358_GSE12417 (AML)	340	2	Cytogeneticly normal: 2 (196)	0.0059
GSE1159 (AML)	293	0		0

Supplementary Table 8. GSEA for expression of stem cell signatures in *Setd2* knockdown *MLL-NRIP3* leukemia cells

Gene Set Name	Original Size	# hits to Dataset	NES	Nominal p-value	FDR q-value	FWER p-Value	Upregulated in class	Gene set Description	References
HUMAN_LEUKEMIC_STEM_CELL	133	87	1.445	0.031	0.034	0.029	KD	Genes up-regulated in LSC¶, defined as CD34+CD38-cells from AML compared to the CD34+CD38+ cells	**Gal et al 2006**[19]
HUMAN_CANCER_RELATED_ESC	335	260	1.429	0.005	0.005	0.005	KD	The 'core ESC-like gene module': genes coordinately up-regulated in a compendium of mouse ESC§which are shared with the human ESC-like module.	**Wong et al 2008**[20]
MYELOID_CELL_PROLIFERATION_AND_SELF_RENEWAL	129	97	1.501	0.008	0.008	0.007	KD	Genes defining proliferation and self renewal potential of the bipotential myeloid cell line FDB	**Brown et al 2006**[21]
MUELLER_PLURINET	229	224	1.39	0.011	0.039	0.194	KD	Genes constituting the PluriNet protein-protein network shared by the pluripotent cells (ESC, embryonical carcinomas and induced pluripotent cells).	**Muller et al 2008**[22]
KOHOUTEK_CCNT2_TARGETS	58	30	1.401	0.067	0.044	0.184	KD	Genes down-regulated in E14 ESC upon knockdown of CYCT2 by RNAi.	**Kohoutek et al 2009**[23]
GUO_HEX_TARGETS_DN	65	49	1.382	0.075	0.036	0.203	KD	Genes down-regulated in day 6 embryoid bodies derived from ESC with HEX knockout	**Guo et al 2003**[24]
DURAND_STROMA_MAX_UP	296	181	1.45	0.005	0.067	0.116	KD	Up-regulated genes discriminating stromal cells that can support HSC£ from those that cannot.	**Link**#

¶LSC: leukemic stem cells, §ESC: embryonic stem cells, £HSC: hematopoietic stem cells,

#Link: http://www.broadinstitute.org/gsea/msigdb/geneset_page.jsp?geneSetName=DURAND_STROMA_MAX_UP&keywords=DURAND

Supplementary Table 9. Limiting dilution assay

Cell dose (cells)	shSetd2 (n=4)	scramble (n=4)
100,000	4¶	4
10,000	4	4
1,000	4	1
100	2	0

¶The numbers of dead mice in total BMT mice are shown.

Supplementary Table 10. All point mutations detected using WGS data

See a separate supplementary file: TableS10.xlsx

Supplementary Table 11. Primers for confirming chromosomal translocation events on DNA or RNA

MLL/NRIP3		**Forward primer**		**Reverse primer**	
PCR	**Intergenic_MLL**	P1	5’-CGCTAAACCCACCTGCTA-3’	P2	5’-CTTCAGGGCTAAGAAGATTAC-3’
	Intergenic_NRIP3	P3	5'-TATCTGAGTTGGTCCTGGTAA-3'	P4	5'-AGGAGCAGGTTCGGGATAT-3'
	MLL_NRIP3	P5	5’-TATTCTAAAGGCCATTTGGC-3’	P6	5’-GCTGATTGAGTTTGTTCGTCTT-3’
RT-PCR	***MLL-NRIP3***	P7	5’-AGTGAAGAAGGGAATGTCTCG-3’	P8	5’-CCGGTGCTTATCCAAGTTTA-3’
TaqMan	***MLL-NRIP3***	P9	5’-GCCTCCACCACCAGAATCAG-3’	P10	5’-GCGCCTCTGCAGAATATTATGA-3’
NF1/AARSD1					
PCR	***NF1-AARSD1***	P1	5'-AACTTCAACTCTAGGCATCTG-3'	P2	5'-TTTCGATTCCTCCAATAAGG-3'

Supplementary Table 12. Primer sequences used for screening *SETD2* mutations in acute leukemia patients

Amplification Primers		Sequencing Primers	
Primer ID	Sequence (5'->3')	Primer ID	Sequence (5'->3')
5'UTR+exon1/R3	CCGCCCTCGGCTGGGGATAAGGC	5'UTR+exon1/R3	CCGCCCTCGGCTGGGGATAAGGC
5'UTR+exon1/F1	CGACGAGCGAGGTAGCGACG	5'UTR+exon1/F2	TAGCGACGCGGGCCGCCCCTG
exon2/F2	CCTGTAGGTAGTAAGTATCCCAA	exon2/F2	CCTGTAGGTAGTAAGTATCCCAA
exon2/R2	TTCAGCTTTTACCCAATTTC	exon3/F1-1	GGCTTTTCATTTCTCCAGTAAAC
exon3/F1-1	GGCTTTTCATTTCTCCAGTAAAC	exon3/F2-2	CATCTGCTTCATGTAACATCCAG
exon3/F2-2	CATCTGCTTCATGTAACATCCAG	exon3/F3-1	GAATGAGCAAGCAGATATTTCC
exon3/F3-1	GAATGAGCAAGCAGATATTTCC	exon3/F4-1	CTCACTCTAGGTCTGAGAGAGGC
exon3/F4-1	CTCACTCTAGGTCTGAGAGAGGC	exon3/F8-1	GGCTCTTCTGAAAGTTCAAATG
exon3/F5-1	CCCGTTATAAATCTACCCTTTCA	exon3/F9-1	AACCGTGAAAGCCAAAATAC
exon3/F6-1	GATTTCAGAATATTAGTAGGTGCA	exon3/R5-4	TTTTGCAGCAAGAAACCCTCGTA
exon3/F7-1	ATCTCTTCAGAGTCTTCCACCAG	exon3/R6-4	CATTATTACGCCTGTTCTCCCTGG
exon3/F8-1	GGCTCTTCTGAAAGTTCAAATG	exon3/R7-4	TCTCTCATCTTCCCAATGGTC
exon3/F9-1	AACCGTGAAAGCCAAAATAC	exon3-F10-4	ACTGGCAAGGCAATGGTTACTGG
exon3/R10-1	TAAAGGCTTTTTCTAACAACAA	exon4/F1	AAAACCCAAAAGAATCTAATGAG
exon3/R1-1	GAGGGCGGTGAGTCTACAG	exon5/F3	CAGTTCTAAGGAATCCCTTTGTG
exon3/R2-2	ACTTGAAGAAGTCCGTACAGAATC	exon6/F1	TCCACATTGCAGTATTTATTTA
exon3/R3-1	GCGTCCTCTCTCGATAAGG	exon7/R1	TTCATGAGTACCTTAGATATGGG
exon3/R4-1	AATTCACTACCTTTTGAACAAGG	exon8/R1	AAGGGTCAGAAGTGTCATACAGTAG
exon3/R5-4	TTTTGCAGCAAGAAACCCTCGTA	exon9/R2	GGTAACAACTCATCTGATCTTGG
exon3/R6-4	CATTATTACGCCTGTTCTCCCTGG	exon10-F3	GGTATTTTTATTTGCTGTGGTAG
exon3/R7-4	TCTCTCATCTTCCCAATGGTC	exon10-R3	AAAACAAACATTAAAGTGTTCAC
exon3/R8-3	GAGAGAAGTCCCAACCTAAGTTTC	exon11/F1	CTCCTCGGGCTCTTTGTAATA
exon4/F1	AAAACCCAAAAGAATCTAATGAG	exon12/F2-1	GGATGGCAAAGAGGATCTTGA
exon4/R1	GTCATCCATAGGTAGGAGAAAGG	exon12-1/R1-4	TCACTATCAACTTTGCATTCAG
exon5/F3	CAGTTCTAAGGAATCCCTTTGTG	exon13-F3	TTTTAAGGGGCCAGGATATATTC
exon5/R3	TAGAGGTTCCAGTGAGCCAAGAT	exon13-R3	AAAGCGACAACAAAACAGTGTAAG
exon6/F1	TCCACATTGCAGTATTTATTTA	exon14-F3	ACAAACCACCCTTTTCCCTTAGC
exon6/R2	ATCAAATCAGTATCAATGGCTCC	exon14-R3	ACTCACACAGGCCACTTACCTG
exon7/F1	AGATGTATGTAGTTTTGCAGGTAA	exon15/F2-2	CTGCCCCTCCACCAGTACCAGTG
exon7/R1	TTCATGAGTACCTTAGATATGGG	exon15/R1-3	CAGGAGAGTACTGCTGCTGTACAC
exon8/F1	ACCTTCATTGCCTAAAGACCCA	exon15/R1-4	TATGGTTGCTTGAGACTGTGCAGG
exon8/R1	AAGGGTCAGAAGTGTCATACAGTAG	exon16-F3	AAAAAGAAACTTTGGGGTTTGAA
exon9/F2	AGTAATAGCAGACTTGTTGGAATC	exon17/R2	AGTGAAAAGCTAATGACAAGAAG
exon9/R2	GGTAACAACTCATCTGATCTTGG	exon18/R1	AAAGAATCCCAAGACCATGCG
exon10/F1	TTCCCAGTTTGGTCTTCATTTGTG	exon18/R2	TCATCTCTACACCTTGATACATGC
exon10/R1	ACAGAACAAGACCCCATCTAAAC	exon19/R2	ACCTTGTGTTCCTGCCTATCTTGG
exon11/F1	CTCCTCGGGCTCTTTGTAATA	exon20-R3	AAAATACTTTCTATGATGAAAAGG
exon11/R1	AACCGAGGCAATCAATATAAC	exon21+3'UTR/F1-1	ACCCCCTTATTAACTTCCTG
exon12/F1-1	GAGCCAAGATCATGCCATTGC	exon21+3'UTR/F2-3	TGGGAGGATGGGTGGTCAGGTAAG
exon12/F2-1	GGATGGCAAAGAGGATCTTGA		
exon12/R2-1	AATTTTGCTTAGGTTTGTAGAAC		
exon12-1/R1-5	CCGTTGCTCTCTTTGGGCTCTAT		
exon13-F3	TTTTAAGGGGCCAGGATATATTC		
exon13-R3	AAAGCGACAACAAAACAGTGTAAG		
exon14/F1	GGAGAAAAGTCCACCAAATAAAC		
exon14/R1	CTTGGGTACTTTTTAACTTGCTC		
exon15/F1-2	CTTATTAGTGCTTTGATTGTGTC		
exon15/F2-2	CTGCCCCTCCACCAGTACCAGTG		
exon15/R1-4	TATGGTTGCTTGAGACTGTGCAGG		
exon15/R2-1	TCCTCCCCTATACCAGATTTAAC		
exon16/R1	AACCCAAAACAAATCCAAACCA		
exon16-F3	AAAAAGAAACTTTGGGGTTTGAA		
exon17/F1	GTGGGCTATTTTGTCCTATTCAG		
exon17/R2	AGTGAAAAGCTAATGACAAGAAG		
exon18/F1	ATTCTCTGTTTCCTAGCCCTGAC		
exon18/R2	TCATCTCTACACCTTGATACATGC		
exon19/F1	AGGGCATCACTAAACTGACTTCTC		
exon19/R2	ACCTTGTGTTCCTGCCTATCTTGG		
exon20/F2	GATTGGCTGGCTTTACTCACTACC		
exon20-R4	ACGCATCCCTCCCCAAACCTTC		
exon21+3'UTR/F1-1	ACCCCCTTATTAACTTCCTG		
exon21+3'UTR/F2-3	TGGGAGGATGGGTGGTCAGGTAAG		
exon21+3'UTR/R1-1	CTAGGTAATCACTTGTAGATGG		

Supplementary Table 13. Mutation information about known oncogenes and tumor suppressor genes

All cancers	Gene	Mutation Rate (%)	Mutated Patients#	Truncated Patients#	F_T¶ (%)	Multi-hit Patients#	F_{MH}§ (%)
Oncogene	*FGFR3*	27	2707	7	0.3	1	0
	KRAS	23	22101	3	0	0	0
	KIT	21	2608	19	0.7	5	0.2
	ABL1	19	779	2	0.3	1	0.1
	EGFR	16	5220	11	0.2	1	0
	RET	10	432	2	0.5	0	0
	NRAS	7	2447	3	0.1	0	0
	ERBB4	3	45	1	2.2	0	0
	HRAS	3	743	2	0.3	0	0
	FLT3	3	1012	2	0.2	0	0
	FGFR2	3	83	4	4.8	1	1.2
	GLI1	2	18	0	0	0	0
	MYH11	2	17	1	5.9	0	0
	AKT1	1	136	0	0	0	0
	MET	1	134	3	2.2	0	0
	ALK	1	156	3	1.9	0	0
	ERBB2	1	66	3	4.5	1	1.5
Suppressor	*TP53*	31	13617	2583	19	216	1.6
	VHL	18	1060	648	61.1	33	3.1
	APC	18	2196	1947	88.7	196	8.9
	CDKN2A	15	1799	428	23.8	42	2.3
	NF2	14	531	480	90.4	30	5.6
	PTEN	11	1632	897	55	161	9.9
	MEN1	8	201	140	69.7	3	1.5
	CDH1	7	169	97	57.4	12	7.1
	WT1	5	370	289	78.1	43	11.6
	STK11	4	212	113	53.3	7	3.3
	MLH1	3	46	18	39.1	3	6.5
	SETD2_COS	3	32	19	59.4	1	3.1
	SETD2_All	3	43	27	62.8	5	11.6
	TGFBR2	2	27	9	33.3	1	3.7
	BRCA1	1	52	35	67.3	1	1.9
	CYLD	1	19	18	94.7	0	0
	NF1	1	240	169	70.4	15	6.3

¶F_T: Gene truncating mutation rate is the ratio of the number of patients with truncating mutation in a certain gene to the number of patients with nonsynonymous or frameshift small indel mutations in that gene in the COSMIC database[7].

§F_{MH}: Gene multi-hit rate is the ratio of the number of patients with at least one truncating mutation and one other coding mutation including missense, nonsense or frameshift small indel mutations of a certain gene to the number of patients with nonsynonymous and frameshift small indels in that gene in the COSMIC database.

Supplementary References

1. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-60 (2009).
2. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-9 (2009).
3. Sherry, S.T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* **29**, 308-11 (2001).
4. Li, G. *et al.* The YH database: the first Asian diploid genome database. *Nucleic Acids Res* **37**, D1025-8 (2009).
5. Ahn, S.M. *et al.* The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group. *Genome Res* **19**, 1622-9 (2009).
6. Abecasis, G.R. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-73 (2010).
7. Bamford, S. *et al.* The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. *Br J Cancer* **91**, 355-8 (2004).
8. Zhang, J. *et al.* The genetic basis of early T-cell precursor acute lymphoblastic leukaemia. *Nature* **481**, 157-63 (2012).
9. Rhodes, D.R. *et al.* Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles. *Neoplasia* **9**, 166-80 (2007).
10. MacDonald, J.W. & Ghosh, D. COPA--cancer outlier profile analysis. *Bioinformatics* **22**, 2950-1 (2006).
11. Armstrong, S.A. *et al.* MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nat Genet* **30**, 41-7 (2002).
12. Mayshar, Y. *et al.* Identification and classification of chromosomal aberrations in human induced pluripotent stem cells. *Cell Stem Cell* **7**, 521-31 (2010).
13. Ben-David, U., Mayshar, Y. & Benvenisty, N. Large-scale analysis reveals acquisition of lineage-specific chromosomal aberrations in human adult stem cells. *Cell Stem Cell* **9**, 97-102 (2011).
14. Dondeti, V.R. *et al.* Integrative genomic analyses of sporadic clear cell renal cell carcinoma define disease subtypes and potential new therapeutic targets. *Cancer Res* **72**, 112-21 (2012).
15. Huang da, W., Sherman, B.T. & Lempicki, R.A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4**, 44-57 (2009).
16. Mosca, L. *et al.* Integrative genomics analyses reveal molecularly distinct subgroups of B-cell chronic lymphocytic leukemia patients with 13q14 deletion. *Clin Cancer Res* **16**, 5641-53 (2010).
17. Ng, P.C. & Henikoff, S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res* **31**, 3812-4 (2003).
18. Wong, W.C. *et al.* CHASM and SNVBox: toolkit for detecting biologically important single nucleotide mutations in cancer. *Bioinformatics* **27**, 2147-8 (2011).
19. Gal, H. *et al.* Gene expression profiles of AML derived stem cells; similarity to hematopoietic stem cells. *Leukemia* **20**, 2147-54 (2006).
20. Wong, D.J. *et al.* Module map of stem cell genes guides creation of epithelial cancer stem cells. *Cell Stem Cell* **2**, 333-44 (2008).

21. Brown, K.A. *et al.* Neutrophils in development of multiple organ failure in sepsis. *Lancet* **368**, 157-69 (2006).

22. Muller, F.J. *et al.* Regulatory networks define phenotypic classes of human stem cell lines. *Nature* **455**, 401-5 (2008).

23. Kohoutek, J. *et al.* Cyclin T2 is essential for mouse embryogenesis. *Mol Cell Biol* **29**, 3280-5 (2009).

24. Guo, Y. *et al.* The homeoprotein Hex is required for hemangioblast differentiation. *Blood* **102**, 2428-35 (2003).