Molecules and Cells

Identification of Novel Functional Variants of SIN3A and SRSF1 among Somatic Variants in Acute Myeloid Leukemia Patients

Jae-Woong Min[1,7], Youngil Koh[2,3,7], Dae-Yoon Kim[3], Hyung-Lae Kim[4], Jeong A Han[5], Yu-Jin Jung[6], Sung-Soo Yoon[2,3,*], and Sun Shim Choi[1,*]

[1]Division of Biomedical Convergence, College of Biomedical Science, Institute of Bioscience & Biotechnology, Kangwon National University, Chuncheon 24341, Korea, [2]Department of Internal Medicine, Seoul National University Hospital, Seoul 03080, Korea, [3]Cancer Research Institute, Seoul National University College of Medicine, Seoul, 03080, Korea, [4]Department of Biochemistry, School of Medicine, Ewha Woman's University, Seoul 03760, Korea, [5]Department of Biochemistry and Molecular Biology, School of Medicine, Kangwon National University, Chuncheon 24341, Korea, [6]Department of Biological Sciences, Kangwon National University, Chuncheon 24341, Korea, [7]These authors contributed equally to this work.
*Correspondence: ssysmc@snu.ac.kr (SSY); schoi@kangwon.ac.kr (SSC)
http://dx.doi.org/10.14348/molcells.2018.0051
www.molcells.org

The advent of massively parallel sequencing, also called next-generation sequencing (NGS), has dramatically influenced cancer genomics by accelerating the identification of novel molecular alterations. Using a whole genome sequencing (WGS) approach, we identified somatic coding and noncoding variants that may contribute to leukemogenesis in 11 adult Korean acute myeloid leukemia (AML) patients, with serial tumor samples (primary and relapse) available for 5 of them; somatic variants were identified in 187 AML-related genes, including both novel (SIN3A, C10orf53, PTPRR, and RERGL) and well-known (NPM1, RUNX1, and CEPBA) AML-related genes. Notably, SIN3A expression shows prognostic value in AML. A newly designed method, referred to as "hot-zone" analysis, detected two putative functional noncoding variants that can alter transcription factor binding affinity near PPP1R10 and SRSF1. Moreover, the functional importance of the SRSF1 noncoding variant was further investigated by luciferase assays, which showed that the variant is critical for the regulation of gene expression leading to leukemogenesis. We expect that further functional investigation of these coding and noncoding variants will contribute to a more in-depth understanding of the underlying molecular mechanisms of AML and the development of targeted anti-cancer drugs.

Keywords: acute myeloid leukemia, somatic variants, whole genome sequencing

INTRODUCTION

Acute myeloid leukemia (AML) is a representative hematologic malignancy (Greenberg et al., 1997). The incidence of AML is increasing, possibly due to increases in life expectancy (Juliusson et al., 2009). With cytotoxic chemotherapy and/or allogeneic stem cell transplantation (ASCT), a certain proportion of AML patients are cured. However, more than 50% of AML patients eventually die due to this aggressive disease (Hulegårdh et al., 2015), and the five-year survival rate of elderly AML patients is less than 10% (Oran and Weisdorf, 2012), indicating that strategies need to be developed to improve treatment outcomes.

Treatment outcomes might be improved by at least two approaches. The first approach is to increase the precision of risk stratification in AML patients. ASCT, which is a potent

Received 29 January, 2018; revised 25 February, 2018; accepted 8 March, 2018; published online 15 May, 2018

eISSN: 0219-1032

but toxic treatment modality, could be more appropriately applied to the high-risk group with an improved stratification system (Ferrara and Schiffer, 2013). The second approach is to advance the understanding of carcinogenic mechanisms in AML, which could reveal novel chemotherapeutic agents. The development of an FLT3 inhibitor (Lee et al., 2017) and an IDH2 inhibitor (Stein et al., 2017) exemplify the second approach. These two approaches will soon become feasible through extensive genomic studies in conjunction with the evaluation of clinical associations.

From a genetic viewpoint, the number of somatic mutations varies significantly depending on the primary origin of the cancer tissue as well as the carcinogenic processes, including the influences of environmental mutagens (Lawrence et al., 2013). The number of somatic mutations that occur in AML patients is low compared with other types of tumors, with the average number of coding somatic variants being 10-15/sample (Cancer Genome Atlas Research Network, 2013). The well-established coding variants observed in AML include mutations in DNMT3A, NRAS, BCOR, SSRP1, YY2, SRSF2, TET2, IDH1/2, CEBPA, RUNX1, and GATA2 (DiNardo et al., 2015; Patel et al., 2012). Furthermore, the prognostic significance of these variants and their genomic categorization have been well studied (Papaemmanuil et al., 2016).

Additionally, the advent of next-generation sequencing (NGS) technologies has dramatically influenced cancer genomics by accelerating the identification of novel molecular alterations. While cancer research has focused on coding regions based on whole exome sequencing (WES), the recent production of whole genome sequencing (WGS) data has facilitated the detection of noncoding variants. TERT promoter mutations are one of the recent findings of analyzing noncoding variants from WGS data (Cancer Genome Atlas Research Network, 2015). In fact, approximately 90% of mutations associated with phenotypic traits are known to be located outside of coding regions, and to obtain a more in-depth understanding of cancer biology, major discoveries in noncoding regions remain to be made. However, justification of the functional validity of noncoding variants is still a challenging task (Ward and Kellis, 2012). A widely accepted way to test the function of noncoding sequences is to determine whether noncoding variants are located within gene expression control regions, such as promoters, enhancers, and transcription factor (TF) binding sites (Chen et al., 2016; Ong and Corces, 2011). In addition, it is important to consider evolutionary conservation when searching for functional noncoding variants (Kircher et al., 2014). Several bioinformatic databases and tools, such as HaploReg (Ward and Kellis, 2011) and Genomic Region Enrichment of Annotations Tool (GREAT) (McLean et al., 2010), have been developed and applied to estimate putative functional noncoding variants, and further improvements in noncoding variant analysis are ongoing.

In this study, we analyze both coding and noncoding mutations in AML using a WGS approach. The functionality of novel noncoding variants is verified through a novel "hot-zone" analysis as well as luciferase assays. External validation of the novel variants using public databases, such as The Cancer Genome Atlas (TCGA), is also performed.

MATERIALS AND METHODS

WGS and variant calling

DNA was extracted using a QuickGene DNA whole blood kit S (Kurabo Industries, Japan) according to the manufacturer's recommendations. For WGS, we used the Solexa sequencing technology platform (HiSeq X Ten, Illumina, USA) following the manufacturer's instructions. FASTQ files were aligned to the human reference genome (human_g1k_v37.fasta) using the Burrows-Wheeler Aligner (BAM) (Li and Durbin, 2009) to generate a SAM file. "SortSam" in the Picard toolset was employed to convert the file to a BAM file and sort by chromosome, and the data were then subjected to a PCR duplicate marking process, which enabled the Genome Analysis Toolkit (McKenna et al., 2010) to ignore duplicates in subsequent processing.

Somatic variants were called with VARSCAN2 (Koboldt et al., 2012), and germline variants from the saliva sample of each patient were called with genome analysis toolkit (GATK) (McKenna et al., 2010). The variant positions were all mapped to the "hg19 (GRCh37)" reference genome. The characteristics of the called variants were annotated with "ANNOVAR" (Wang et al, 2010). Common polymorphic variants were further excluded using the following filtration criteria: (1) variants matched dbSNP variants (version 138), (2) variants had a minor allele frequency greater than 0.01 in the 1000 Genomes Project, and (3) variants matched paired germline variants.

Sanger sequencing

A total of six single nucleotide variants (SNVs) were verified through conventional Sanger sequencing using dye-terminator chemistry and were analyzed with an ABI 3730 automatic sequencer (Applied Biosystems, USA). Oligo primers were designed using Primer3 software (Rozen and Skaletsky, 2000) to amplify the genome fragments containing the mutations from bone marrow samples. PCR was performed using DNA polymerase (SolGent, Korea) under optimized thermal conditions. The PCR products were evaluated in 2% agarose gels, purified and sequenced in both directions using BigDye® Terminator (Applied Biosystems) reactions and subsequently loaded into a capillary sequencer. Sequence variants were verified with chromatograms using the SeqMan® feature of Lasergene® software (DNASTAR, USA).

Obtaining regulatory marks for defining specific genomic regions referred to as "hot-zones"

ENCODE (ENCODE Project Consortium, 2004) information for four types of regulatory elements [transcription factor binding sites (TFBSs), DNase hypersensitivity sites (DHSs), active promoters (APs), and histone modification marks (HMMs)] was downloaded from the UCSC Table Browser (https://genome.ucsc.edu/cgi-bin/hgTables). These regulatory elements and their genomic locations were acquired based on the "hg19" reference genome. We chose to download the information regarding these regulatory marks, with the exception of TFBSs, in a lymphoblastoid cell line, GM12878. It should be noted that data on APs, DHSs, and

HMMs are cell-type specific information, whereas TFBS data are not. Information on ten types of HMMs, including H3k4me1, H3k4me3, H3k27ac, H3k336me3, H3k4me2, H3k79me2, H3k9ac, H3k27me3, H3k9me3 and H4k20me1, were downloaded. These regulatory positions were overlaid onto the "hg19" reference genome, and we then searched for genomic regions exhibiting all of these regulatory marks in the same area, which we defined as "hot-zone" regions.

Other external sources of variant data

The overall procedure for data preparation is shown in Supplementary Fig. S1. A total of 30 TCGA WGS-derived variants (TWG30) and 193 TCGA WES-derived variants (TES193) were downloaded from the TCGA database (https://gdc-portal.nci.nih.gov/); 134 WES-derived variants (Lawrence 134) were obtained from a study by Lawrence et al. (2013) and 532 cancer-associated variants were downloaded from the *catalogue of somatic mutations in cancer* (COSMIC) database (http://cancer.sanger.ac.uk/census). TWG30 had been called by the TCGA team using CaVEMan (Stephens et al., 2012), and we further filtered out the variants with a somatic mutation probability of less than 0.95 or a germline mutation probability of at least 0.05. Since the 193 TCGA WES-derived variants had been mapped onto the "hg18" reference genome by the TCGA team (Cancer Genome Atlas Research Network, 2013), it was necessary to convert the variant locations to the "hg19" genome using the "liftOver" tool from UCSC (https://genome.ucsc.edu/util.html). Subsequently, filtration procedures were performed as described for the TWG30 variants (https://gdc-portal.nci.nih.gov/). Gene expression data from 24 types of cancer, provided as RNA-Seq results and patient clinical information, were downloaded from TCGA (https://tcga-data.nci.nih.gov/).

Luciferase assay

A luciferase assay was conducted to investigate the functional role of a noncoding variant detected upstream of the SRSF1 gene. For transfection, the pGL4.10[luc2] (Promega, Cat No. E6651) vector and the pRL-TK (Promega, Cat No. E2241) vector, with a 7-base pair (bp) deletion from both ends of the sequence based on the mutation locus, were used. The cloning vector was transfected into the NB4 (Lanotte et al., 1991) cell line using a Gene Pulser 2 Electroporator (Bio-Rad, Cat No. 165-2108). After transfection, NB4 cells were incubated for 40-42 h at 37℃ under 5% CO_2. The luciferase assay was performed using a GloMax® 20/20 Luminometer (Promega) and a Dual-Luciferase assay kit (Promega, E1910) following the manufacturer's protocol. We measured both Renilla and firefly activity to determine the effect of the modification and compared the ratio between the two activities in cell lines subjected to three different transfections using the control vector, the SRSF1 wild-type vector, and the SRSF1 mutation vector. Activity values obtained by performing six repeated experiments were employed in this analysis.

Codes used for statistical tests and batch work

All statistical tests were performed with R studio (https://www.rstudio.com/). The relationships between the expression levels of genes and patient prognosis were investigated using Kaplan-Meier (KM) plots combined with log-rank tests (Efron, 1988) and Cox regression analyses (Royston and Altman, 2013). The transcription factor binding motif score was calculated using bioconductor (https://www.bioconductor.org/) from the R package TFBSTools (Tan and Lenhard, 2016) combined with JASPAR2014 (Mathelier et al., 2014) and JASPAR2016 (Mathelier et al., 2016). The other scripts necessary for several batch jobs were executed using an in-house built *Perl* script.

Other bioinformatics web tools used to determine the functional importance of variants

The categorical functions or gene ontology (GO) classifications of the genes carrying somatic mutations were estimated with a web-based DAVID tool (https://david.ncifcrf.gov/). Changes in the stability of secondary protein structures caused by nonsynonymous mutations were measured using DUET (http://bleoberis.bioc.cam.ac.uk/duet/stability), and the "MutationMapper" tool from cBioPortal (http://www.cbioportal.org/mutation_mapper.jsp) was employed to visualize the location of any variant within each corresponding translated protein structure. IGV (http://software.broadinstitute.org/software/igv/) and UGENE (http://ugene.net/) were used to view the NGS reads and Sanger sequencing results, respectively.

RESULTS

WGS statistics

Genomic DNA was extracted from a total of 11 Korean AML patient samples to conduct WGS. Tumor blood and skin tissues were retrieved from each of the cytogenetically normal AML patients. The patients included three females and eight males with ages ranging from 21 to 74 years (Supplementary Table S1). The achieved sequencing depths were 69.5× (ranging from 61.9 to 76.0×) and 32.3× (ranging from 24.1 to 37.1×) on average for the tumor and matched saliva samples, respectively. Approximately 99.3% of the reference human genome was covered at least once, with approximately 98.6% and 85.5% of the normal and tumor samples combined showing ≥10× coverage and ≥30× coverage, respectively. A detailed summary of the sequencing statistics for all samples is provided in Supplementary Table S2. The overall procedure for variant annotations and filtrations is depicted in Supplementary Fig. S1. A total of 30 somatic mutations in exons were detected in each sample after extensive filtration (Supplementary Table S3).

Identification of somatic mutations in 11 Korean AML patients

Two different datasets were produced from the 11 Korean AML patient samples: (1) the "SNU-p11" dataset, comprising only the 11 primary AML samples, and (2) the "SNU-pr5" dataset, containing five paired primary and relapse samples obtained from each of five individuals (Supplementary Table S1). A WGS-based analysis was conducted, and somatic SNVs for the samples in these two datasets were identified (see "Materials and Methods"). Other SNVs derived from external public databases, including TWG30, TES193, Law-

rence 134, and COSMIC census genes, were also included in the present work (see "Materials and Methods") (Supplementary Fig. S1).

Our basic strategy for analyzing the SNVs detected in the SNU-p11 and SNU-pr5 datasets was as follows. (1) Identify SNVs occurring around genic regions, including exons, introns, 3'- and 5'- untranslated regions (UTRs), and regions containing various regulatory elements within 2 kb up- and down-stream of genes, i.e., well-known regions in which most functional elements are located (Farré et al., 2007; Keightley et al., 2005; Zhang et al., 2017). (2) Compare the SNVs derived from the SNU samples with those from the TWG30, TES193, and Lawrence134 datasets. (3) Identify noncoding variants located in "hot-zone" regions, which were newly defined in the present work. (4) Demonstrate the functional relevance of the validated variants through survival analysis.

The numbers of SNVs detected in different genic regions in each patient sample are summarized in Table S3. Consistent with previous studies, the number of mutations estimated for our AML samples was relatively small (Kandoth et al., 2013); for example, we identified approximately 30 exonic mutations per patient on average (Supplementary Table S3 and Supplementary Fig. S2A). The average number of mutations per patient was less than that estimated from the TWG30 dataset (a WGS-based dataset) but slightly higher than the number estimated from the Lawrence134 dataset by Lawrence et al. (2013), which might be primarily due to the stringency of the threshold for SNV calling (Supplementary Table S3). Despite differences in the number of mutations, the mutation architecture was very similar among these three AML datasets; for example, in all three, C to T (or G to A) transition mutations represented the highest proportion of mutations, whereas A to T (or T to A) transversion mutations represented the lowest proportion (Supplementary Fig. S2B).

Analysis of nonsynonymous SNVs detected in the SNU-p11 samples

A total of 187 genes with nonsynonymous SNVs were detected in the 11 primary AML samples (SNU-p11). Thirty-seven of these genes were also detected in the TWG30 dataset, 34 in the TES193 dataset, eight in the Lawrence134 dataset, and 13 in the list of COSMIC census genes, whereas 17 were detected in at least two other datasets, indicating that mutations in these genes might play critical roles in AML pathogenesis (Supplementary Table S4). Interestingly, several well-established AML-associated genes, such as NPM1, RUNX1, and WT1, were also found (Supplementary Table S4). However, the depiction of the mutations derived from the SNU-p11 and TWG30 samples shown in Fig. 1 demonstrates the well-known concept of tumor heterogeneity, i.e., different AML patients carried different sets of variants. Some genes, such as NPM1, WT1, and PABPC3, were found to be mutated in more than two patients in both the SNU-p11 and TWG30 datasets (Fig. 1).

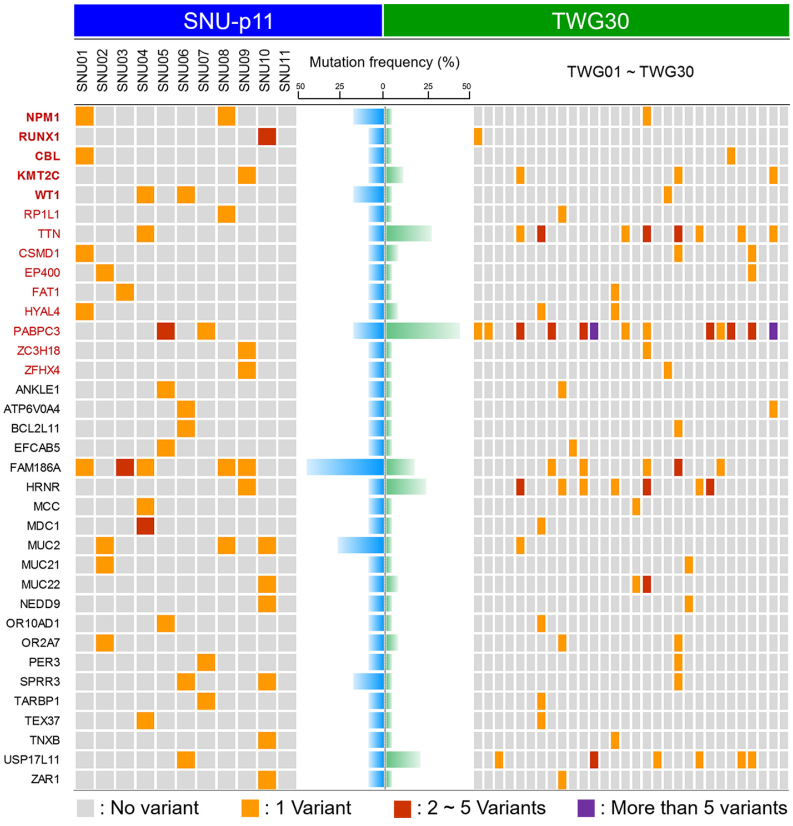

Fig. 1. Comparison of common variants between the SNU-p11 and TWG30 datasets. The patient samples with somatic nonsynonymous variants are represented by colored boxes. The patient samples on the left and the right are from the SNU-p11 and TWG30 datasets, respectively. Gene symbols are provided in the leftmost column only for genes carrying any somatic variant in any patient sample. The gray, red, orange, and purple boxes represent no variant, only one variant, two to five variants, and more than five variants per sample, respectively. The blue and green bars in the center indicate the frequency of mutations estimated from the SNU-p11 and TWG30 samples, respectively. The gene symbols in bold font are genes that overlap with the COSMIC census genes (Methods), whereas genes that overlap with previously identified AML-associated genes from three datasets (TWG30, TES193, and Lawrence134) are shown in red. The genes that overlap both COSMIC census and previously known genes are shown in bold red font.

These genes generally carried only one nonsynonymous SNV; however, a few genes, including RUNX1 and PABPC3, carried two to five (or even more than five) nonsynonymous SNVs per patient (Fig. 1). To investigate the most frequently mutated genes in the SNU-p11 samples, genes exhibiting mutations in more than two patients or carrying more than five mutations in a single patient were selected and are depicted in Supplementary Fig. S3. A total of 25 frequently mutated genes were identified from the SNU-p11 samples, and two of these, MUC4 and CEBPA, were mutated in more than three of the 11 total patient samples.

Analysis of nonsynonymous SNVs detected in the SNU-pr5 samples

As described above, we also generated WGS data for five of the 11 patients with paired primary-relapse samples (SNU-pr5). These data provided an opportunity to investigate mutational changes during relapse. A total of 41 mutated genes were shared between the primary and relapse tumors within a single patient or between the primary and relapse tumor samples derived from different patients (Fig. 2). Seventeen of the 41 genes had previously been identified as AML-associated genes in other studies (Fig. 2). We classified the mutated genes detected in SNU-pr5 into three categories, namely, "primary-specific genes", "commonly mutated genes", and "relapse-specific genes." Figure 3 shows a graphical representation of the mutated genes corresponding to each category.

Subsequently, GO classification terms were assigned to each category. A main function assigned to the "primary-specific genes" was "negative regulation of apoptotic process", as NPM1 was found in this category (top panel of Fig. 3). Several well-known AML-associated genes were found in the "commonly mutated genes" category, and these included CEBPA and GATA2, whose GO functional categories include "embryonic development", "notch signaling pathway", and "homeostatic process" (middle panel of Fig. 3). In contrast, the "relapse-specific genes" assigned to the GO terms "chromatin organization" and "negative regulation of transcription" were found to include several novel genes, such as KDM4B and PTMA, as well as known cancer-associated genes, such as RB1 and ARID1A (bottom panel of Fig. 3). All of the GO terms assigned to the three categories of genes identified in the present work have consistently been reported to be altered during cancer pathogenesis (Lee et al., 2006; Li et al., 2014; Rapin et al., 2014).

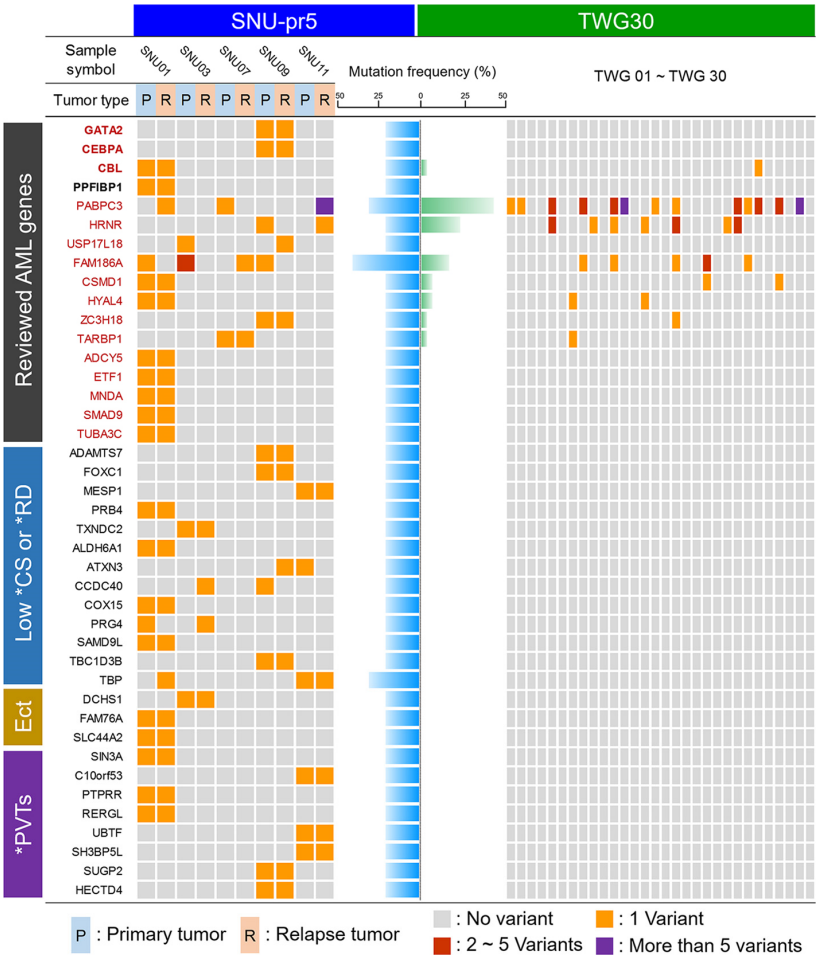

Fig. 2. Profiles of commonly detected coding mutations between primary and relapse samples. A total of 41 mutated genes that were commonly detected between the primary and relapse samples of a single patient or between the primary and relapse tumor samples from different patients are shown here. Please refer to the legend of Fig. 1 for a description of the colored boxes and gene symbol labeling.

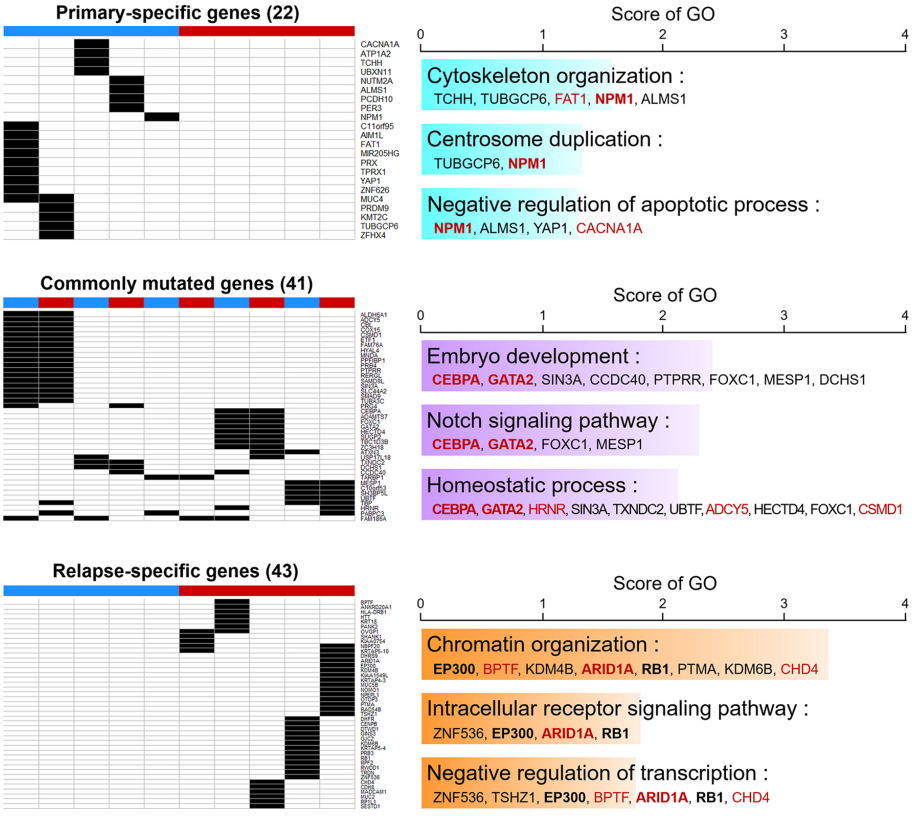

Fig. 3. Profiles of somatic nonsynonymous variants detected in both the paired primary and relapse samples. The 106 nonsynonymous variants obtained from the SNU-pr5 dataset, which was generated from 10 WGS runs for five primary samples and their paired relapse samples, were grouped into three categories: "primary-specific", "commonly mutated", and "relapse-specific". The presence or absence of any variant in any patient sample is represented in a matrix box, and a filled box indicates that the patient sample carries any nonsynonymous variant. The numbers of variant-harboring genes allocated to each group are shown in parentheses for each matrix box category. The bar in the uppermost portion of the matrix boxes represents the samples under each bar, and the boxes under the blue bar and the red bar represent primary and relapse samples, respectively. Gene symbols corresponding to any variant-carrying gene are shown on the right side of the matrix boxes. The bar graphs on the right show the results of a GO analysis using DAVID (Methods). The blue, purple, and orange colors on the bar graphs represent each classified category. The size of the bars indicates the score of each GO category [$-\log_{10}$(P value)]. Please refer to the legend of Figure 1 for the labeling scheme of the gene symbols inside the colored bars.

Investigation of the functional importance of nonsynonymous variants

Using the list of recurrent nonsynonymous variants, we then attempted to identify putative functional genetic changes attributable to leukemogenesis via *in silico* prediction with the following criteria: (1) commonly mutated genes, (2) genes carrying nonsynonymous substitutions with high variant allele frequencies (VAFs), (3) genes carrying novel SNVs, and (4) genes carrying highly conserved SNVs with a PhyloP score > 1.0. Four genes, SIN3A, C10orf53, PTPRR, and RERGL, were chosen based on these criteria (Fig. 3 and Supplementary Table S5), and their variants were validated via Sanger sequencing (Supplementary Fig. S4).

We then investigated whether the genetic changes in these four genes exhibit prognostic value for AML using the variants and clinical information deposited in external databases such as TCGA. Although no TCGA samples were found to harbor the same mutations in these four genes, expression information based on RNA-Seq data was available, which allowed us to investigate the prognostic significance of these four genes. As shown in Fig. 4, the TCGA AML (labeled LAML in TCGA) patients with SIN3A and C10orf53 expression levels in the top 20% had a significantly worse prognosis than the LAML patients with expression levels in the bottom 20%. Interestingly, the same SIN3A variant (Y325C) was detected in a uterine corpus endometrial carcinoma (UCEC) sample deposited in TCGA, suggesting functional importance of these variants in other cancers. Moreover, the mutation was calculated by the DUET web tool (http://bleoberis.bioc.cam.ac.uk/duet/stability) to cause a "Predicted Stability Change (PSC)" of -1.433 Kcal/mol, indicating a significant disruption of the secondary protein structure. The genic locations of the four validated SNVs are schematically represented at their corresponding positions within each protein structure (Supplementary Fig. S5).

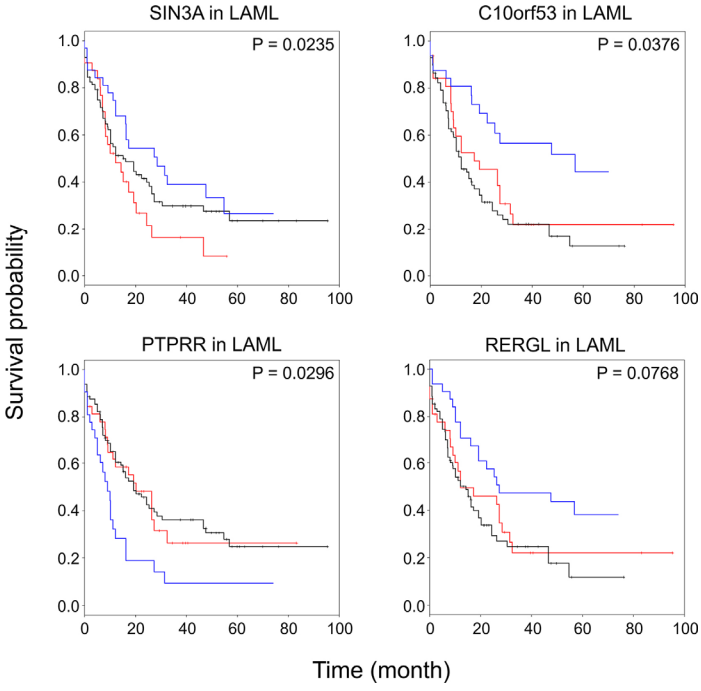

Fig. 4. Survival differentiated by the expression levels of the four mutated genes. RNA-Seq and patient clinical information for each type of cancer was obtained from TCGA (Methods). Patients with each type of cancer were classified based on the expression levels of the four genes carrying somatic coding variants validated by Sanger sequencing, as shown in Supplementary Figs. S4A-S4D. To compare patient survival among each cancer type, KM plots were generated for each group; the red, black, and blue lines represent the "top 20% (high)" group, "intermediate" group, and "bottom 20% (low)" group, respectively. The KM plot in the "SIN3A in LAML" panel shows the patient prognosis depending on the expression levels of SIN3A in acute myeloid leukemia. The numbers of patients in each group of LAML grouped by gene expression were as follows: high (32), low (32), and intermediate (97).

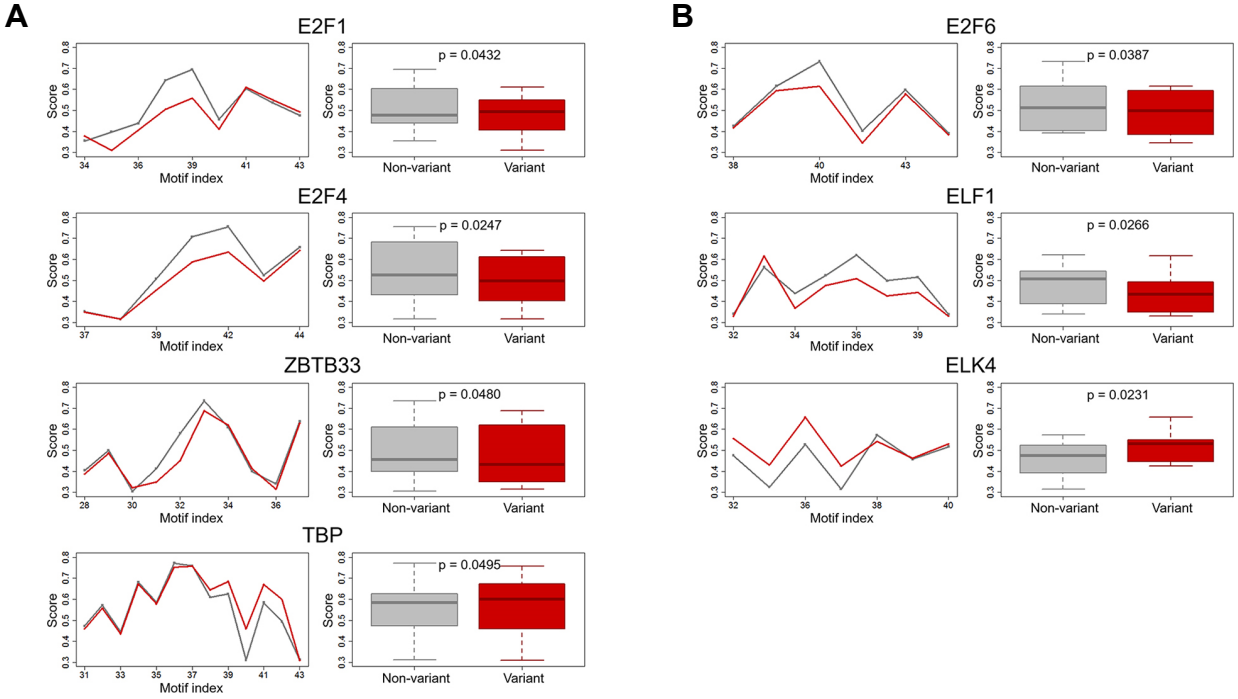

Fig. 5. Alteration of TF motif binding by two "hot-zone" variants. Alterations in TF binding activity caused by variants located in the "hot-zone" regions of two genes, PPP1R10 and SRSF1, were analyzed using TFBSTools (Supplementary Methods). Briefly, to investigate the influence of the variants on TF binding activity, a sequence with 50 bases before and after each variant (i.e., a total of 101 bases) was selected as a window for sliding-window analysis across genomic locations shifted by one base pair. The TF binding motif score was measured for each window based on the motif matrix of each TF. The scores for non-variant reference (i.e., hg19) queries were also measured for each window. The significance of the score differences between the non-variant reference window and the variant window was then tested via a t-test. The scores obtained from the sliding-window analysis are shown as line graphs on the left for each gene; the red and gray lines represent the TF binding activity of the variant and non-variant, respectively. The box plots on the right constitute a conversion of the line graphs. (A) Plots for the PPP1R10 variant. (B) Plots for the SRSF1 variant.

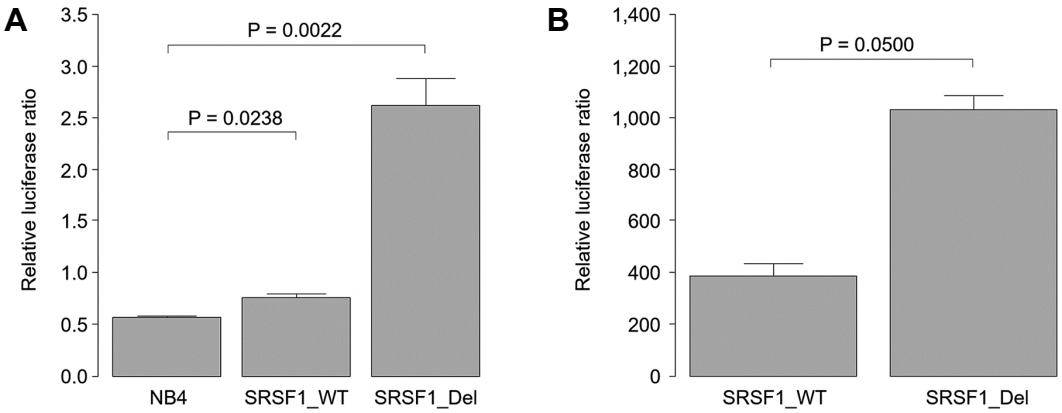

Fig. 6. Functionality of the SRSF1 "hot-zone" mutation verified through a luciferase assay. The activity of the SRSF1 promoter with a deletion mutation spanning the site of the "hot-zone mutation" (-44 G to A) was measured through a luciferase activity. The luciferase activity of the mutant construct was compared with that of the control construct without any mutation. (A) Genomic location of the "hot-zone" mutation of the SRSF1 gene. (B) Schematic representation of the construction of the expression vector for estimating the change in luciferase activity due to the 15-bp deletion that spans the SRSF1 "hot-zone" mutation site (-44 G to A). (C) NB4 cells were transfected with the control vector carrying the normal promoter region (labeled SRSF1_WT) or the deletion vector spanning the "hot-zone mutation" site (-44 G to A) of the SRSF1 gene (labeled SRSF1_del). (D) The same assay was conducted with the CHO-K1 cell line. The same labeling scheme as that in (C) was applied. The error bars represent the means ± standard deviations for six repeats of the experiments. P values were estimated by Mann-Whitney tests.

Definition of "hot-zone" regions for detecting noncoding regulatory mutations

As shown in Supplementary Table S3, many more SNVs were identified outside of coding regions, such as in UTRs and within 2 kb up- and down-stream of genes. Since there is no definitive bioinformatics-based strategy or algorithm for identifying functional noncoding mutations, we designed a new strategy referred to as the "hot-zone" method. The rationale for this method is that variants located in genomic regions harboring multiple regulatory signals, such as HMMs, DHSs, and TFBSs, are likely to be more functionally important than variants located in genomic regions without these signals. Using this rationale, regulatory marks were downloaded from ENCODE (https://genome.ucsc.edu/ENCODE/) and mapped to their corresponding reference sequence positions ("Material and Methods"), and the "hot-zone" regions overlapping multiple regulatory marks were defined. The variants found in the "hot-zone" regions were then selected for further analyses. Two "hot-zone" variants near the PPP1R10 (-31 G to A) and SRSF1 (-44 A to T) genes were selected for verification by Sanger sequencing and were validated as real noncoding somatic mutations that might play functional roles in AML (Supplementary Fig. S4E and S4F). It is highly likely that these two variants alter the promoter function of these genes because the two variants are located near the transcription start sites (TSSs) of the two genes.

Functional implications of the validated mutations

We further investigated whether these two variants (-31 G to A of PPP1R10 and -44 G to A of SRSF1) altered TF binding activity. The "TFBSTools" algorithm (https://www.bioconductor.org/) predicted that the PPP1R10 "hot-zone" mutation perturbs the binding activities of E2F1, E2F4, ZBTB33, and TBP (Fig. 5A), whereas the SRSF1 "hot-zone" mutation alters the binding activities of E2F6, ELF1, and ELK4, ELK4 (Fig. 5B). This result strongly suggests that the "hot-zone" mutations of these two genes are likely to lead to alterations in downstream gene expression. Interestingly, the SRSF1 and PPP1R10 genes are interacting partners in a protein-protein interaction network (Supplementary Fig. S6).

The gene expression regulation functionality of the validated "hot-zone" noncoding mutation of SRSF1 (-44 G to A) was then investigated using a luciferase assay in the human AML cell line NB4. A schematic representation of the design of the expression vectors is shown in Figs. 6A and 6B ("Materials and Methods"). Interestingly, we observed that luciferase activity was significantly increased in NB4 cell lines transfected with deletion mutant constructs spanning the "-44 G to A" site compared with NB4 cell lines transfected with wild-type constructs (P value < 0.05 by Mann-Whitney test) (Supplementary Table S6), strongly suggesting that the noncoding "hot-zone" variant might be involved in gene expression regulation (Fig. 6C). The same luciferase assay performed as a duplicative experiment in Chinese hamster ovary K1 (CHO-K1) cells led to the same observation (Fig. 6D), namely, that the cell line transfected with the mutant construct showed significantly increased luciferase activity (Supplementary Table S7) (P value ≤ 0.05 by Mann-Whitney test). This result not only confirms the functional role of the nucleotide located in the "hot-zone" area but also indicates functional conservation between humans and other species.

DISCUSSION

AML is a prototypical hematologic malignancy that is chemo-sensitive with the potential to be cured but exhibits

high relapse rates, which reduces cure rates (Deschler and Lübbert, 2006). As in other cancers, somatic mutation causes AML, and some cancer predisposition-associated germline alterations also contribute to its pathogenic mechanism (de Voer et al., 2015). Anecdotal genomic studies of AML have been published by TCGA collaborators (Cancer Genome Atlas Research Network, 2013) in which AML-associated mutations were analyzed from a 50-variant WGS dataset and a 150-variant WES dataset. This study revealed several important features of the AML genomic landscape (Cancer Genome Atlas Research Network, 2013), and a total of 23 genes were identified as significantly mutated genes. Additionally, an extensive genomic study by Pappaemmanuil et al. (2016) categorized AML into nine genetic categories, including transcription factor fusion, DNA methylation-related genes, chromatin-modifying genes, and tumor suppressor genes. Although numerous AML genomic studies have contributed to broadened knowledge of AML genomics, we still believe that low-frequency leukemogenic mutations remain to be discovered in AML patients.

One distinguishable feature of our study is that we designed a novel method for narrowing down the number of noncoding variants that should be validated from "immense" to "several". As noted above, we refer to the new method as "hot-zone" analysis. Basically, "hot-zone" analysis works by defining noncoding regions that are highly likely to be functional based on overlaying multiple epigenetic regulatory signals, and any given noncoding variant located in the "hot-zone" is subsequently determined to be sufficiently valuable that it should be validated through further experiments. In fact, we identified a total of 45 noncoding "hot-zone" variants, two of which (from the PPP1R10 and SRSF1 genes) were selected based on certain criteria and further validated as real somatic mutations. It can therefore be stated that "hot-zone" analysis is very effective for detecting putative functional noncoding variants.

Among the two validated "hot-zone" variants, the SRSF1 gene was selected for further functional verification through luciferase assays. SRSF1 plays an important role in splicing and has recently been recognized as one of the key oncodriver genes in small cell lung cancer (Jiang et al., 2016) and breast cancer (Das and Krainer, 2014). As shown in Fig. 6, the "hot-zone" mutation of SRSF1 was involved in enhancing SRSF1 expression. Interestingly, up-regulated expression of SRSF1 was recently suggested to stimulate mTOR, which plays a key role in carcinogenesis (Jiang et al., 2016; Malakar et al., 2017). A luciferase assay for the PPP1R10 variant was not performed because the roles of PPP1R10 in cancer and leukemia have not yet been well elucidated, and biological interpretation of the results would therefore be very difficult. However, it is still plausible that the "hot-zone" variant of PPP1R10 might have leukemogenic potential because these two proteins, SRSF1 and PPP1R10, were found to be interacting partners in a protein-protein interaction network (Supplementary Fig. S6).

In addition, several novel coding variants were also identified. We not only validated these variants through Sanger sequencing but also provided some theoretical evidence of the functional importance of the variants. In particular, SIN3A Y325C, which was detected in both primary and relapse tumors, was located in a region with a high PhyloP conservation score of > 1.0 and was predicted to perturb the stability of the secondary structure of SIN3A. Notably, SIN3A expression is associated with AML patient survival in the TCGA cohort (Fig. 4). SIN3A is a regulator of histone deacetylases and is essential for the regulation of hematopoiesis (Heideman et al., 2014); furthermore, SIN3A is known to suppress the function of tumor suppressor genes, such as CDKN2A (Jiang et al., 2014). Hence, it is possible that a functional change in SIN3A via the Y325C mutation would contribute to leukemia development.

Cancers consist of a founding clone harboring key mutations and sub-clones carrying various mutations that confer a survival advantage to tumor cells in the complex context of the tumor microenvironment. These heterogeneous genetic properties of cancers are considered one of the main mechanisms leading to chemotherapy resistance. In particular, disease relapse after chemotherapy is always related to clonal evolution, and several types of clonal evolution have been suggested (Graham and Sottoriva, 2017). Ley et al. (2008) analyzed clonal evolution history in AML and revealed that three mutations, specifically mutations in FLT3, NPM1, and DNMT3A, tend to co-occur in the founding clones of patients. Because a prerequisite for analyses of clonal evolution is high-depth sequencing, our study design (WGS of 60× depth) was unfortunately not suitable for clonal evolution analysis. However, we revealed and confirmed some genomic aspects of AML using serial samples. We found that NPM1 was mutated in primary AML samples, whereas RB1 and ARID1A were mutated in relapse samples. The findings obtained for NPM1 are in accordance with the results obtained by Ley et al., 2008. Moreover, mutations in tumor suppressor genes, such as RB1 and ARID1A, are commonly found in the late stage of cancer (Lee et al., 2017), and our findings confirmed these biological insights from previous studies.

Unfortunately, RNA sequencing was not possible in the SNU-p11 or SNU-pr5 groups due to the lack of RNA samples from these patients, limiting further functional assays of the mutants found in our study. Although we performed external validation using TCGA data and luciferase assays, it is likely that RNA sequencing of our patient samples could have provided strong evidence for the functionality of the mutations identified in the present study.

In conclusion, we report six novel somatic mutations in coding and noncoding regions that might play critical roles in AML pathogenesis. Further experimental evidence is necessary to reveal their full involvement in cancer and to develop prognostic or diagnostic markers or anti-cancer drugs.

Note: Supplementary information is available on the Molecules and Cells website (www.molcells.org).

ACKNOWLEDGMENTS

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (NRF-2016R1D1A1B03930411) and by a Korea

Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI) funded by the Ministry of Health & Welfare (HI14C0072). The authors would like to thank Donghyun Park for his useful comments.

REFERENCES

Cancer Genome Atlas Research Network. (2013). Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. N. Engl. J. Med. *368*, 2059-2074.

Cancer Genome Atlas Research Network. (2015). Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas. N. Engl. J. Med. *372*, 2481-2498.

Chen, L., Wang, W., Cao, L., Li, Z., and Wang, X. (2016). Long non-coding RNA CCAT1 acts as a competing endogenous RNA to regulate cell growth and differentiation in acute myeloid leukemia. Mol. Cells *39*, 330-336.

Das, S., and Krainer, A.R. (2014). Emerging functions of SRSF1, splicing factor and oncoprotein, in RNA metabolism and cancer. Mol. Cancer Res. *12*, 1195-1204.

de Voer, R.M., Hahn, M.M., Mensenkamp, A.R., Hoischen, A., Gilissen, C., Henkes, A., Spruijt, L., van Zelst-Stams, W.A., Kets, C.M., Verwiel, E.T., et al. (2015). Deleterious germline BLM mutations and the risk for early-onset colorectal cancer. Sci. Rep. *5*, 14060.

Deschler, B., and Lübbert, M. (2006). Acute myeloid leukemia: Epidemiology and etiology. Cancer *107*, 2099-2107.

DiNardo, C.D., Ravandi, F., Agresta, S., Konopleva, M., Takahashi, K., Kadia, T., Routbort, M., Patel, K.P., Mark Brandt., Pierce, S., et al. (2015). Characteristics, clinical outcome, and prognostic significance of IDH mutations in AML. Am. J. Hematol. *90*, 732-736.

Efron, B. (1988). Logistic regression, survival analysis, and the kaplan-meier curve. J. Am. Stat. Assoc. *83*, 414-425.

ENCODE Project Consortium. (2004). The ENCODE (ENCyclopedia of DNA elements) project. Science *306*, 636-640.

Farré, D., Bellora, N., Mularoni, L., Messeguer, X., and Albà, M.M. (2007). Housekeeping genes tend to show reduced upstream sequence conservation. Genome Biol. *8*, R140.

Ferrara, F., and Schiffer, C.A. (2013). Acute myeloid leukaemia in adults. Lancet *381*, 484-495.

Graham, T.A., and Sottoriva, A. (2017). Measuring cancer evolution from the genome. J. Pathol. *241*, 183-191.

Greenberg, P., Cox, C., LeBeau, M.M., Fenaux, P., Morel, P., Sanz, G., Sanz, M., Vallespi, T., Hamblin, T., Oscier, D., et al. (1997). International scoring system for evaluating prognosis in myelodysplastic syndromes. Blood *89*, 2079-2088.

Heideman, M.R., Lancini, C., Proost, N., Yanover, E., Jacobs, H., & Dannenberg, J.H. (2014). Sin3a-associated Hdac1 and Hdac2 are essential for hematopoietic stem cell homeostasis and contribute differentially to hematopoiesis. Haematologica *99*, 1292-1303.

Hulegårdh, E., Nilsson, C., Lazarevic, V., Garelius, H., Antunovic, P., Rangert Derolf, Å, Möllgård, L., Uggla, B., Wennström, L., Wahlin, A., et al. (2015). Characterization and prognostic features of secondary acute myeloid leukemia in a population-based setting: A report from the swedish acute leukemia registry. Am. J. Hematol. *90*, 208-214.

Jiang, S., Willox, B., Zhou, H., Holthaus, A.M., Wang, A., Shi, T. T., Maruo, S., Kharchenko, P.V., Johannsen, E.C., Kieff, E., et al. (2014). Epstein-barr virus nuclear antigen 3C binds to BATF/IRF4 or SPI1/IRF4 composite sites and recruits Sin3A to repress CDKN2A. Proc. Natl. Acad. Sci. USA *111*, 421-426.

Jiang, L., Huang, J., Higgs, B. W., Hu, Z., Xiao, Z., Yao, X., Conley, S., Zhong, H., Liu, Z., Brohawn, P., et al. (2016). Genomic landscape survey identifies SRSF1 as a key oncodriver in small cell lung cancer. PLoS Genet. *12*, e1005895.

Juliusson, G., Antunovic, P., Derolf, A., Lehmann, S., Mollgard, L., Stockelberg, D., Tidefelt, U., Wahlin, A., and Höglund, M. (2009). Age and acute myeloid leukemia: Real world data on decision to treat and outcomes from the swedish acute leukemia registry. Blood *113*, 4179-4187.

Kandoth, C., McLellan, M.D., Vandin, F., Ye, K., Niu, B., Lu, C., Xie, M., Zhang, Q., McMichael, J.F., Wyczalkowski, M.A., et al. (2013). Mutational landscape and significance across 12 major cancer types. Nature *502*, 333-339.

Keightley, P.D., Lercher, M. J., and Eyre-Walker, A. (2005). Evidence for widespread degradation of gene control regions in hominid genomes. PLoS Biol. *3*, e42.

Kircher, M., Witten, D.M., Jain, P., O'Roak, B.J., Cooper, G.M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. Nat. Genet. *46*, 310-315.

Koboldt, D.C., Zhang, Q., Larson, D. E., Shen, D., McLellan, M.D., Lin, L., Miller, C.A., Mardis, E.R., Ding, L., and Wilson, R.K. (2012). VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. *22*, 568-576.

Lanotte, M., Martin-Thouvenin, V., Najman, S., Balerini, P., Valensi, F., and Berger, R. (1991). NB4, a maturation inducible cell line with t(15;17) marker isolated from a human acute promyelocytic leukemia (M3). Blood *77*, 1080-1086.

Lawrence, M.S., Stojanov, P., Polak, P., Kryukov, G.V., Cibulskis, K., Sivachenko, A., Carter, S.L., Stewart, C., Mermel, C.H., Roberts, S.A., et al. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature *499*, 214-218.

Lee, S., Chen, J., Zhou, G., Shi, R. Z., Bouffard, G.G., Kocherginsky, M., Ge, X., Sun, M., Jayathilaka, N., Kim, Y.C., et al. (2006). Gene expression profiles in acute myeloid leukemia with common translocations using SAGE. Proc. Natl. Acad. Sci. USA *103*, 1030-1035.

Lee, J., Lee, J., Kim, S., Kim, S., Youk, J., Park, S., An, Y., Keam, B., Kim, D.W., Heo, D.S., et al. (2017). Clonal history and genetic predictors of transformation into small-cell carcinomas from lung adenocarcinomas. J. Clin. Oncol. *35*, 3065-3074.

Lee, L.Y., Hernandez, D., Rajkhowa, T., Smith, S.C., Raman, J.R., Nguyen, B., Small, D., and Levis, M. (2017). Preclinical studies of gilteritinib, a next-generation FLT3 inhibitor. Blood *129*, 257-260.

Ley, T. J., Mardis, E. R., Ding, L., Fulton, B., McLellan, M.D., Chen, K., Dooling, D., Dunford-Shore, B.H., McGrath, S., Hickenbotham, M., et al. (2008). DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. Nature *456*, 66-72.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with burrows-wheeler transform. Bioinformatics *25*, 1754-1760.

Li, Y., Liang, M., and Zhang, Z. (2014). Regression analysis of combined gene expression regulation in acute myeloid leukemia. PLoS Comput. Biol. *10*, e1003908.

Malakar, P., Shilo, A., Mogilevsky, A., Stein, I., Pikarsky, E., Nevo, Y., Benyamini, H., Elgavish, S., Zong, X., et al. (2017). Long noncoding RNA MALAT1 promotes hepatocellular carcinoma development by SRSF1 upregulation and mTOR activation. Cancer Res. *77*, 1155-1167.

Mathelier, A., Zhao, X., Zhang, A. W., Parcy, F., Worsley-Hunt, R., Arenillas, D.J., Buchman, S., Chen, C.Y., Chou, A., Ienasescu, H., et al. (2014). JASPAR 2014: An extensively expanded and updated open-access database of transcription factor binding profiles. Nucleic Acids Res. *42*, D142-7.

Mathelier, A., Fornes, O., Arenillas, D.J., Chen, C.Y., Denay, G., Lee, J., Shi, W., Shyr, C., Tan, G., Worsley-Hunt, R., et al. (2016). JASPAR 2016: A major expansion and update of the open-access database of transcription factor binding profiles. Nucleic Acids Res. *44*, D110-5.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. *20*, 1297-1303.

McLean, C.Y., Bristor, D., Hiller, M., Clarke, S. L., Schaar, B.T., Lowe, C.B., Wenger, A.M., and Bejerano, G. (2010). GREAT improves functional interpretation of cis-regulatory regions. Nat. Biotechnol. *28*, 495-501.

Ong, C., and Corces, V.G. (2011). Enhancer function: New insights into the regulation of tissue-specific gene expression. Nat. Rev. Genet. *12*, 283-293.

Oran, B., and Weisdorf, D.J. (2012). Survival for older patients with acute myeloid leukemia: A population-based study. Haematologica *97*, 1916-1924.

Papaemmanuil, E., Gerstung, M., Bullinger, L., Gaidzik, V.I., Paschka, P., Roberts, N.D., Potter, N.E., Heuser, M., Thol, F., Bolli, N., et al. (2016). Genomic classification and prognosis in acute myeloid leukemia. N. Engl. J. Med. *374*, 2209-2221.

Patel, J.P., Gönen, M., Figueroa, M.E., Fernandez, H., Sun, Z., Racevskis, J., Van Vlierberghe, P., Dolgalev, I., Thomas, S., Aminova, O., et al. (2012). Prognostic relevance of integrated genetic profiling in acute myeloid leukemia. N. Engl. J. Med. *366*, 1079-1089.

Rapin, N., Bagger, F.O., Jendholm, J., Mora-Jensen, H., Krogh, A., Kohlmann, A., Thiede, C., Borregaard, N., Bullinger, L., Winther, O., et al. (2014). Comparing cancer vs normal gene expression profiles identifies new disease entities and common transcriptional programs in AML patients. Blood *123*, 894-904.

Royston, P., and Altman, D.G. (2013). External validation of a cox prognostic model: Principles and methods. BMC Med. Res. Methodol. *13*, 33-2288-13-33.

Rozen, S., and Skaletsky, H. (2000). Primer3 on the WWW for general users and for biologist programmers. Methods Mol. Biol. *132*, 365-386.

Stein, E.M., DiNardo, C.D., Pollyea, D.A., Fathi, A.T., Roboz, G.J., Altman, J.K., Stone, R.M., DeAngelo, D.J., Levine, R.L., Flinn, I.W., et al. (2017). Enasidenib in mutant-IDH2 relapsed or refractory acute myeloid leukemia. Blood *130*, 722-731.

Stephens, P.J., Tarpey, P.S., Davies, H., Van Loo, P., Greenman, C., Wedge, D.C., Nik-Zainal, S., Martin, S., Varela, I., Bignell, G.R., et al. (2012). The landscape of cancer genes and mutational processes in breast cancer. Nature *486*, 400-404.

Tan, G., and Lenhard, B. (2016). TFBSTools: An R/bioconductor package for transcription factor binding site analysis. Bioinformatics *32*, 1555-1556.

Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res., *38*, e164.

Ward, L.D., and Kellis, M. (2011). HaploReg: A resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. Nucleic Acids Res. *40*, D930-D934.

Ward, L.D., and Kellis, M. (2012). Interpreting noncoding genetic variation in complex traits and human disease. Nat. Biotechnol. *30*, 1095-1106.

Zhang, Q., Cheng, T., Jin, S., Guo, Y., Wu, Y., Liu, D., Xu, X., Sun, Y., Li, Z., He H., et al. (2017). Genome-wide open chromatin regions and their effects on the regulation of silk protein genes in bombyx mori. Sci. Rep. *7*, 12919.