Supplemental Data

Horizontal Meta-Analysis Identifies Common Deregulated Genes across AML Subgroups Providing a Robust Prognostic Signature

Ali Nehme[1,#], Hassan Dakik[1,#], Frédéric Picou[1,2], Meyling Cheok[3], Claude Preudhomme[3,4], Hervé Dombret[5], Juliette Lambert[6], Emmanuel Gyan[1,7], Arnaud Pigneux[8], Christian Recher[9], Marie C Béné[10], Fabrice Gouilleux[1], Kazem Zibara[11,12], Olivier Herault[1,2,13], Frédéric Mazurier[1,13]

Correspondence:

Frédéric Mazurier: EA 7501 University of Tours, CNRS ERL7001 LNOx, 10 boulevard Tonnellé, BP 3223, 37032 Tours Cedex 01, France. Tel: +33 2 47 36 60 75, email: frederic.mazurier@inserm.fr.

Ali Nehme: McGill University and Genome Quebec Innovation Centre, 740 Dr. Penfield street, Montreal, Quebec, Canada, H3T1E6., email: ali.nehme2@mcgilll.ca

Hassan Dakik: McGill University, RI-MUHC, Glen Site, 1001 Décarie Boulevard, Montreal, Quebec, Canada, H4A3J1., email: hassan.dakik@mail.mcgill.ca

Supplemental Methods

Datasets assembly

Published microarray datasets containing transcriptomic data from AML patient samples were downloaded from the Gene Expression Omnibus (GEO) database, which was queried for the following terms: "AML", "Acute Myeloid Leukemia", "Leukemia", "Bone marrow" and "hematopoietic". Affymetrix™ GeneChip Human Genome U133 Plus 2.0 Array data were used in this study. Affymetrix™ data were downloaded as raw CEL files from the GEO database. Samples were annotated using Supplemental annotation files available on the GEO database and using detailed annotation files published in corresponding articles. To increase robustness and reduce unknown covariate effects during data analysis, we excluded: (1) datasets with less than 20 samples; (2) samples with undefined tissue of origin, cell type or cytogenetic abnormality; (3) samples corresponding to sorted cells; and (4) Refractory anemia with excess blasts (RAEB) samples.

Quality control and normalization

The R/Bioconductor[1,2] *Simpleaffy* and arrayQualityMetrics packages were used to extract quality measurement of microarrays.[3,4] RNA degradation was evaluated by assessing 3' to 5' ratio of GAPDH and beta-actin transcripts, where a cutoff of 1 and 3 were set, respectively. Hybridization quality was examined using hybridization spike-in controls (BioB, BioC, BioD and Crex) and percent present values. Samples were excluded due to low quality, which was defined according to the recommendations of Affymetrix, based on different criteria including scale factor, hybridization quality (bioB), RNA degradation, Normalized Unscaled Standard Error (NUSE) and Relative Log Expression (RLE). Samples with array-intensity beyond 3-folds, as compared to the median intensity across arrays, were referred to as "technical" outliers and hence excluded [5]. High quality AML samples (N=1534), retained after quality control (Supplemental tables 1-2), were background corrected and RMA normalized using RMAexpress software (http://rmaexpress.bmbolstad.com/).

Differential gene expression and enrichment analyses

Pairwise comparisons between each of the 10 main AML karyotypes and the normal control samples were performed using Statistical Analysis of Microarrays (SAM)[6] after global batch adjustment of all samples.[7] A cutoff with log2-fold change (FC) >1.5 and *Q* value <.05 was applied for differential gene expression analysis. To identify genes with robust differential expression, the list of commonly deregulated genes (CODEG) was narrowed down to those that also passed the cutoff, 1) in the absence of batch adjustment and 2) after pairwise batch adjustment between each karyotype and control samples. Batch adjustment was performed using supervised algorithm implemented in *ComBat* R/Bioconductor package by including samples karyotype as covariate of interest in the equation.[8,9] Cytogenetic groups with less than 5 samples were eliminated from comparisons.

Enrichment analysis on gene ontology biological processes (GO BP) was conducted in R environment using the Bioconductor's *topGO* package.[10] Only genes that are mapped to Affymetrix™ plus 2 platform were used as a background reference. GO terms with less than 10 genes were removed from the analysis. Terms were considered significant when 5 or more enriched genes with weighted-Fisher *P* value below .05. Significant terms were ranked by fold-enrichment, and up to 20 terms were visualized with circos plots using *Circlize* package in R environment [11]. Protein-protein interaction (PPI) scores were extracted from STRING database.[12] PPI and GO networks were built using Cytoscape software.[13]

Normalized GSE76009, GSE65625, GSE83533 and GSE24759 datasets were downloaded from the GEO database, and the probeset with the highest average intensity was selected for each gene. For GSE76009 and GSE65625 datasets. Gene Set Enrichment Analysis (GSEA)[14,15], was performed using default settings with 1000 phenotype permutations, whereas for GSE24759 dataset, which has a small number of samples per phenotype, analysis was performed with 1000 gene set permutations. Comparisons with nominal *P* value <.05 and FDR <.05 were considered significant. Of note, among the 330 differentially expressed genes, a total of 256, 320 and 305 were detected in GSE24759, GSE76009 and GSE65625 datasets, respectively.

Methylation and gene mutation analysis

Methylation (HM450) beta-values, RNA sequencing expression levels (RNA Seq V2 RSEM, Illumina GA-IIX), and mutation data from whole exome or genome sequencing for genes of interest were downloaded from the AML TCGA dataset using cbioportal's *cgdsr* package in R environment.[16,17]

Validation datasets

Microarray analysis

The score was validated on five independent cohorts from four microarray datasets, GSE6891[18], GSE10358[19], GSE12417[20], and ALFA-0701.[21] Clinical annotations and treatment protocols are described in corresponding publications. Raw Affymetrix™ CEL files for these datasets were downloaded from the GEO database and individually normalized using RMA algorithm[22]. For each dataset, a representative probeset with the highest average intensity was selected for each gene. The CODEG22 score calculation and patient stratification were performed as described above. For GSE6891 dataset, clinical annotations for 279 patients were collected from the Leukemia-Gene-Atlas website (http://www.leukemia-gene-atlas.org/LGAtlas/).[23] For GSE10358 dataset, clinical annotations for 223 patients were collected from the GEO database and the corresponding articles.[19,24]

RT-qPCR analysis

The score was also validated on a retrospective cohort of 142 patients from the French Innovative Leukemia Organization (FILO, N° BB-0033-00073, Goelamsthèque/FILOthèque Cochin hospital, Paris). Briefly, primary leucoblasts were obtained after informed consent from BM samples of patients with hyperleucocytic AML (Supplemental Table 19-20). RNA purity was analyzed using Agilent 2100 Bioanalyzer (Agilent Technologies, Les Ulis, France). One microgram of RNA were reverse transcribed using the SuperScript® VILO™ cDNA Synthesis kit (Invitrogen, Paris, France). RT-qPCR reactions were performed on three ng of cDNA using LightCycler® 480 Probes Master (Roche). Samples were subjected to initial denaturation step (5 min, 95°C), followed by 45 PCR cycles (10 s, 95°C, then 30 s, 60°C) and a final cooling step (30 s, 40°C). Triplicates of each sample were analyzed using the Cycle threshold (Ct) values determined with the LightCycler®

480 software. The geometric Ct mean of human *GAPDH* and *EF1A* were used as endogenous control to normalize the expression of target genes: ΔCT = "Ct target" – "Ct reference geomean". ΔCT values for each patient are presented in Supplemental Table 19. The CODEG22 score was calculated for each patient from -ΔCT after gene-wise scaling and centering, as described above. The sequences of primers and probes are documented in Supplemental Table 22.

Survival analysis

Relapse-free survival (RFS) was defined as the time from complete remission (CR) until relapse, death, or last follow-up. Overall survival (OS) was defined as the time from AML diagnosis until death or last follow-up. Event-free survival (EFS) was defined as the time from diagnosis until an event occurred (induction failure, relapse or death) or last follow-up. Survival analysis was done as described previously.[25] Briefly, survival curves are visualized using Kaplan-Meier[26] plots and comparisons between categories were performed using Mantel-Cox Log-Rank test.[27] Cox proportional hazard (CPH) regression was used to perform univariate and multivariate analyses[28]. Violation of the proportional hazards assumption was examined using Schoenfeld residuals[29]. Wald's test was used to evaluate the significance of individual regression coefficients, and the Likelihood Ratio Test (LRT) was used to evaluate the global significance of multivariate models. Survival analysis was performed and visualized in R environment using *survival*[30] and *survminer*[31] packages, respectively.

Supplemental Results

The expression profile of CODEGs correlates with their methylation profile

We hypothesized that the high frequency of downregulated genes in CODEGs could be associated with CpG hypermethylation in AML. Therefore, the methylation profile of both up- and downregulated genes was investigated in the AML methylation dataset from TCGA. As expected, the majority of downregulated genes (71%) were highly methylated, with CpG methylation level above 30%, whereas most of the upregulated genes (78%) were hypomethylated in AML samples (Supplemental Figure 8A). Next, the methylation of CODEGs was examined in association with the mutational status of DNA methylation regulators that are frequently mutated in AML. Interestingly, the methylation of 25 genes, all downregulated except *PDGFC*, was increased in association with mutations in the positive demethylation effectors *IDH1/2*, *TET1/2* and *WT1* (Supplemental Table 23 and Supplemental Figure 8B). In addition, the methylation of 33 genes, all downregulated except *ATP6V0A2*, was decreased in correlation with inactivation mutations in DNA methyltransferase (DNMT) enzymes (Supplemental Table 23 and Supplemental Figure 8B). Of note, the DNA methylation of five genes (*ADGRG3*, *FAR2*, *VNN3*, *GSAP*, and *FGR*) was epigenetically associated with mutations in both groups of methylation regulators. Together, these results suggest that the expression of downregulated CODEGs may rely on epigenetic regulation.

We also examined whether CODEGs contained genes that are known to be mutated in AML. Only two such genes, *FLT3* and *DNMT3A*, were found after examining the mutational status in the TCGA AML dataset (Supplemental Figure 9). Thus, the increased expression of these two genes may play a major role, independent of their mutation status, in all AML subgroups.

High CODEG22 correlates with poor survival in AML patients of various cytogenetic groups

The prognostic value of the model was independently verified on 2 well-annotated and heterogeneous AML microarray datasets: GSE6891 (Supplemental Tables 24 N=279)[18] and GSE10358 (Supplemental Table 16, N=223).[19] Interestingly, a High CODEG22 score was associated with poor OS and EFS in both datasets (Figure 5A-B). Indeed, High score patients showed shorter median OS and EFS times compared to Low score patients, both in GSE6891 (Supplemental Table 24, OS: 16.59 vs 85.78 months, $p = .0021$; EFS: 9.43 vs 16.51, $p = .0034$) and GSE10358 (Supplemental Table 16, OS: 14.4 months vs not reached, $p < .001$; EFS: 9.7 vs

29.7 months, p <.001). Similarly, a High score was also associated with poorer survival probability in both GSE6891 (Figure 5A and univariate model in Supplemental Table 25, OS HR=1.57 a p = .007; EFS HR=1.53 with p = .007) and GSE10358 (Figure 5B and Supplemental Table 26, OS HR=2.53 with p <.001; EFS HR=1.53 with p <.001). Results also showed that CODEG22 score was neither associated with gender, karyotype, NPM1 mutations and FLT3ITD status in both datasets (Supplemental Tables 16 and 24), nor with blasts percentage and WBC in the GSE10358 dataset (Supplemental Table 16). Moreover, the CODEG22 score remained prognostic after adjustment for age and cytogenetic abnormalities, both in GSE6891 (Supplemental Table 25, OS HR=1.49, p = .018; EFS HR=1.54, p = .008) and GSE10358 (Supplemental Table 26, OS HR=2.01, p = .001; EFS HR=1.83, p = .002). Interestingly, the addition of the CODEG22 score (multivariate model 2) to the model containing age and cytogenetic abnormalities (multivariate model 1) increased the model's predictive value in GSE6891 dataset based on Likelihood-Ratio-Test (LRT) assessment (Supplemental Table 25, OS LRT p = .0351; EFS LRT p = .0154). This proves that the predictive power of the CODEG22 score is independent of age and cytogenetic abnormalities.

High CODEG22 correlates with poor survival in the Beat-AML RNA-seq data set

The Beat-AML RNA-seq dataset was downloaded from the Supplemental data of the work done by Tyner *et al.* (Tyner et al., 2018). The dataset offers whole-exome-sequencing, clinical annotation, and RNA-seq data for 451 AML samples. It is worth noting that only 277 samples are collected from AML patients at diagnosis, whereas 174 samples are either from MDS/MPN patients (n=12) or from relapsed AML specimen. Because our main objective from this data was to further validate the prognostic power of our model, we used the diagnosis subset of this dataset to test the CODEG22 signature.

Our analysis showed that patients with high CODEG22 score in the Beat-AML dataset showed shorter overall-survival (OS) (Supplemental Table 27: OS time: 10.46 vs 23.19 months, and Figures 10A-B), and a poorer survival probability (Supplemental Table 28: univariate model: HR=1.71 and p=0.004), compared to patients with low score. Interestingly, high score was also associated with higher relapse and lower complete response rates (Supplemental Table 27). The correlation of high score with poor OS outcome was maintained for patients with CA-AML (Supplemental Figure 10C: OS p<0.001), as well as for patients belonging to the ELN poor risk

group (Supplemental Figure 10D: OS p=0.02). In contrast to the other validation datasets, the score was not prognostic within the CN-AML subset (data not shown). This is probably due to the correlation between our score and the mutational status of NPM1 in this particular cohort (Supplemental Table 27).

Nevertheless, CODEG22 remained prognostic in multivariate Cox-regression-analysis of the whole cohort after adjustment for age, cytogenetic risk, NPM1 mutation, FLT3ITD, biallelic CEBPA, TP53 mutation and other recurrent mutations (Supplemental Table 28: CODEG22-High HR=1.81 and p=0.045). In addition, the inclusion of CODEG22 in the multivariate model improved its overall prognostic power (LTR p-value decreased from $1.23x10^{-5}$ to $5.18x10^{-6}$). Taken together, these data further confirm that our score offers independent prognostic information that is not captured by recurrent mutations or by other currently used prognostic factors.

It is worth noting that CODEG22 outperformed the LSC17 score in this dataset when both scores were included in the same model (Univariate analysis: HR = 1.71 and P = 0.004 vs. HR = 1.67 and P = 0.006; Multivariate analysis: HR = 1.47 and P = 0.099 vs. HR = 1.37 and P = 0.203).

Description of the up-regulated CODEGs.

ANKRD28: ankyrin repeat domain 28, also called *KIAA0379*, is putative regulatory subunit of protein phosphatase 6 (PP6) that may be involved in the recognition of phosphoprotein substrates.[32] *ANKRD28* has been reported to be upregulated in CML[33], and was identified as an *NUP98* fusion partner in a case of secondary AML.[34]

ATP6V0A2: V-type proton ATPase 116 kDa subunit a isoform 2 is part of the proton channel of V-ATPases. It is an essential component of the endosomal pH-sensing machinery that have been shown to activate prolyl hydroxylases (PHD) leading to the degradation of HIF-1alpha[35]. *ATP6V0A2* is one of 33 genes among CODEGs that we identified as hypomethylated in association with DNMTs mutations. Notedly, it has been reported to be epigenetically regulated in association with mutations in epigenome-modifying enzymes in AML.[36]

CDK6: Cyclin-dependent kinase 6 is a serine/threonine-protein kinase involved in the control of the cell cycle and differentiation. Indeed, CDK6 has been found to be required for the progression of *MLL*-rearranged AML[37] and identified as key regulator in the activation of LSCs.[38] CDK inhibitors have been used to treat a wide spectrum of cancers.[39-42]

DNM1: Dynamin-1 is a microtubule-associated force-producing protein that is required for clathrin-mediated endocytosis and mitochondrial division. *DNM1* has been shown to be abnormally expressed in lung and colorectal cancers.[43] Together with DNM2, DNM1 is proposed as potential therapeutic target in cancer.[44]

DNMT3A: DNA (cytosine-5)-methyltransferase 3A is essential for genome-wide de novo methylation. It is one of the most frequently and early mutated genes in AML in association with a loss of methylation activity and poor prognosis.[45,46]

FLT3: It is a tyrosine-protein kinase that acts as cell-surface receptor for the cytokine FLT3LG and regulates differentiation, proliferation and survival of hematopoietic progenitor cells. Interestingly, FLT3, a hallmark of high risk AML and associated with high percentages of BM blasts [47], showed the highest fold-increase in all AML samples, compared to control samples. FLT3 inhibitors have shown promising results in treating AML patients harboring *FLT3* mutations.[48] Indeed, combination therapy targeting several aberrant pathways could be used in the future to improve the response to treatment in resistant patients.[49]

MIB1: mindbomb E3 ubiquitin protein ligase 1 is an E3 ubiquitin-protein ligase that was proposed to disassemble the centriolar satellites and suppress ciliogenesis by marking fold protein pericentriolar matrix protein 1 (PCM1) for proteasomal degradation.[50,51] MIB1 also regulates all known canonical Notch ligands in the Notch signal-sending cells.[52] *MIB1*'s conditional knockout in mice models leads to myeloproliferative disease[53], however, this was attributed to defective signaling in the microenvironment rather than hematopoietic cells.[54,55] Since Notch activation mediates multilineage potential while its downregulation is associated with differentiation[56], it is thereby possible that the increased expression of *MIB1* in AML bone marrow could be linked to AML differentiation blockage.

MLLT11: MLLT11 transcription factor 7 cofactor, also called AF1Q. The overexpression of *MLLT11* is associated with poor prognosis in AML[57], and resistance to imatinib in CML[58] and is

involved in the progression of ovarian and bladder cancers[59,60]. Translocation between *KMT2A* and *MLLT11* has been reported in AML.[61]

NRXN2: Neurexin-2 is a neuronal cell surface protein that may be involved in cell recognition and cell adhesion. It is one of three genes that have been found to harbor age-related hypomethylation CpG sites in human monocytes.[62]

PDGFC: Platelet-derived growth factor C is a member of PDGF family that is essential for the regulation of a range of biological processes from embryonic development, to cell proliferation, angiogenesis and cell migration.[63] PDGFs have been proposed to promote the proliferation of AML blasts while AML-secreted PDGFs was suggested to modulate the bone marrow microenvironment.[64] PDGFs are known to mediate oncogenic signaling, and PDGFC autocrine signaling is reported to promote the progression of breast cancer[65] and fibrocarcinoma.[66] Therefore, many specific antibodies and small molecules inhibitors have been developed to target PDGF signaling in cancer.[67] It is worth noting that *PDGFC* was downregulated in the two cytogenetic groups harboring *MLL* fusion mutations in contrary to the other cytogenetic groups.

PLEKHA5: Pleckstrin homology domain containing A5. Its expression in melanoma was associated with early development of brain metastasis[68], and was thereby proposed as potential therapeutic target.[69]

RABEP2: rabaptin, RAB GTPase-binding effector protein 2, also called FRA, is a member of the rabaptin family and a component in the endosomal vesicle trafficking complex.[70] It was found to be associated with poor prognosis in AML.[71]

SOX4: SRY-box 4 is a member of the SOX transcription factors and is crucial for embryogenesis and the development of many tissues. It promotes survival, proliferation, epithelial mesenchymal transition as well as metastasis in a multitude of cancers.[72] *SOX4* is a poor prognostic marker in AML.[73] Its expression has been reported to be increased in AML samples harboring t(8;21) translocation[74], and was found to contribute to AML progression in *CEBPA* mutant AML.[75]

SINHCAF: SIN3-HDAC complex-associated factor, also called FAM60A, is a member of the SIN3A–HDAC (histone deacetylase) complex that is a master transcriptional repressor.[76] SINHCAF is required for self-renewal in embryonic stem cells[77], and is reported to act as repressor

of HIF2A.[78] It was recently reported to be transcriptionally upregulated within a population of immune-evading AML cells that is enriched in LSCs.[79]

SPINK2: is a serine protease inhibitor of the Kazal type (SPINK) that is highly expressed in HSCs[80], LSCs[81] and in most leukemia cell lines.[82] It was recently reported as poor prognostic marker in AML.[83] This gene was among the top downregulated genes in apoptotic chronic lymphocytic leukemia (CLL) cell lines after treatment with arsenic trioxide.[84] It is worth noting that although globally upregulated in AML, *SPINK2* is down-regulated in t(8;21) subtype compared to normal bone marrow.

TGIF2: TGFB induced factor homeobox 2 is a transcriptional co-repressor that represses TGFB signaling by interacting with TGFB-activated SMAD proteins.[85] TGIF2 was shown to promote colon cancer[86], osteosarcoma[87] as well as HBV-associated hepatocarcinogenesis.[88] It was also found to be upregulated in LSCs in AML.[80]

ZBTB8A: Zinc finger and BTB domain-containing protein 8A, also called BOZF1, is a member of the POZ domain and Krüppel-like zinc finger (POK) family of proteins that regulate apoptosis and cell cycle. It has been found to be upregulated in many cancers and was shown to stimulate cell proliferation through the inhibition of p53 and p21.[89]

ZBTB10: is a zinc finger and BTB domain-containing protein, also called RINZF. It has been found to be increased in LSCs.[90] However, it is also a repressor of Specificity protein (SP) family of transcription factors and is activated by reactive oxygen species (ROS) downstream a wide spectrum of ROS-inducing anticancer agents.[91-93]

References

1. Gentleman RC, Carey VJ, Bates DM, et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*. 2004;5:R80.
2. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2018.
3. Wilson CL, Miller CJ. Simpleaffy: a BioConductor package for Affymetrix Quality Control and data analysis. *Bioinformatics*. 2005;21(18):3683-3685.
4. Kauffmann A, Gentleman R, Huber W. arrayQualityMetrics--a bioconductor package for quality assessment of microarray data. *Bioinformatics (Oxford, England)*. 2009;25(3):415-416.
5. McCall MN, Murakami PN, Lukk M, Huber W, Irizarry RA. Assessing affymetrix GeneChip microarray quality. *Bmc Bioinformatics*. 2011;12:10.

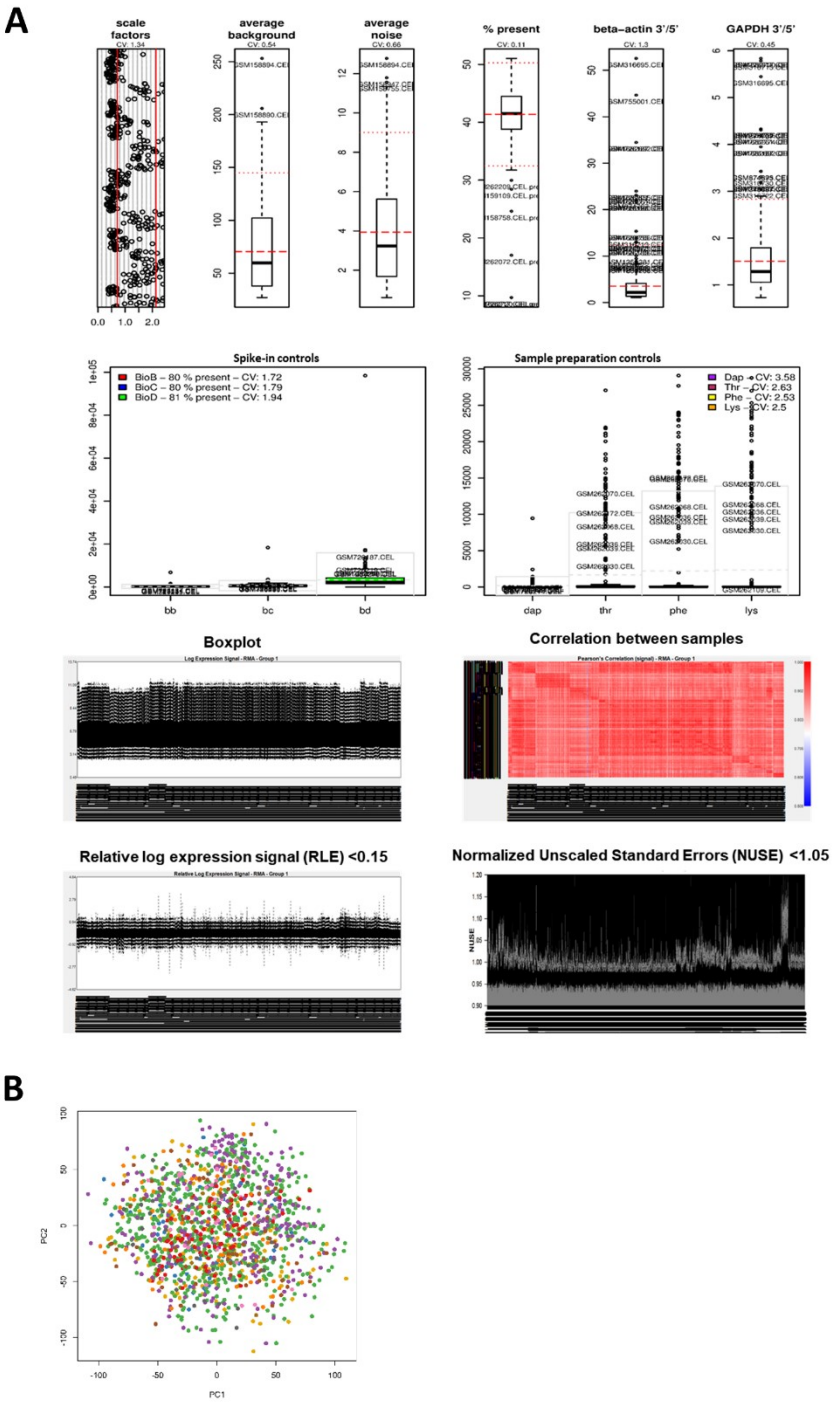
6. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America*. 2001;98(9):5116-5121.
7. Nygaard V, Rodland EA, Hovig E. Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses. *Biostatistics*. 2016;17(1):29-39.
8. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*. 2007;8(1):118-127.
9. Muller C, Schillert A, Roethemeier C, et al. Removing Batch Effects from Longitudinal Gene Expression - Quantile Normalization Plus ComBat as Best Approach for Microarray Transcriptome Data. *Plos One*. 2016;11(6):23.
10. Alexa A, Rahnenfuhrer J. topGO: Enrichment Analysis for Gene Ontology; 2016:R package.
11. Gu ZG, Gu L, Eils R, Schlesner M, Brors B. circlize implements and enhances circular visualization in R. *Bioinformatics*. 2014;30(19):2811-2812.
12. Szklarczyk D, Franceschini A, Wyder S, et al. STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Research*. 2015;43(D1):D447-D452.
13. Shannon P, Markiel A, Ozier O, et al. Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Research*. 2003;13(11):2498-2504.
14. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*. 2005;102(43):15545-15550.
15. Mootha VK, Lindgren CM, Eriksson KF, et al. PGC-1 alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genetics*. 2003;34(3):267-273.
16. Gao JJ, Aksoy BA, Dogrusoz U, et al. Integrative Analysis of Complex Cancer Genomics and Clinical Profiles Using the cBioPortal. *Science Signaling*. 2013;6(269):19.
17. Cerami E, Gao JJ, Dogrusoz U, et al. The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data. *Cancer Discovery*. 2012;2(5):401-404.
18. Verhaak RGW, Wouters BJ, Erpelinck CAJ, et al. Prediction of molecular subtypes in acute myeloid leukemia based on gene expression profiling. *Haematologica-The Hematology Journal*. 2009;94(1):131-134.
19. Tomasson MH, Xiang Z, Walgren R, et al. Somatic mutations and germline sequence variants in the expressed tyrosine kinase genes of patients with de novo acute myeloid leukemia. *Blood*. 2008;111(9):4797-4808.
20. Metzeler KH, Hummel M, Bloomfield CD, et al. An 86-probe-set gene-expression signature predicts survival in cytogenetically normal acute myeloid leukemia. *Blood*. 2008;112(10):4193-4201.
21. Castaigne S, Pautas C, Terre C, et al. Effect of gemtuzumab ozogamicin on survival of adult patients with de-novo acute myeloid leukaemia (ALFA-0701): a randomised, open-label, phase 3 study. *Lancet*. 2012;379(9825):1508-1516.
22. Lim WK, Wang K, Lefebvre C, Califano A. Comparative analysis of microarray normalization procedures: effects on reverse engineering gene networks. *Bioinformatics*. 2007;23(13):I282-I288.
23. Hebestreit K, Grottrup S, Emden D, et al. Leukemia Gene Atlas - A Public Platform for Integrative Exploration of Genome-Wide Molecular Data. *Plos One*. 2012;7(6):7.
24. Mardis ER, Ding L, Dooling DJ, et al. Recurring Mutations Found by Sequencing an Acute Myeloid Leukemia Genome. *New England Journal of Medicine*. 2009;361(11):1058-1066.
25. Ibrahim S, Dakik H, Vandier C, et al. Expression Profiling of Calcium Channels and Calcium-Activated Potassium Channels in Colorectal Cancer. *Cancers*. 2019;11(4):15.
26. Kaplan EL, Meier P. NONPARAMETRIC-ESTIMATION FROM INCOMPLETE OBSERVATIONS. *Journal of the American Statistical Association*. 1958;53(282):457-481.

27. Bland JM, Altman DG. The logrank test. *British Medical Journal*. 2004;328(7447):1073-1073.
28. Cox DR. REGRESSION MODELS AND LIFE-TABLES. *Journal of the Royal Statistical Society Series B-Statistical Methodology*. 1972;34(2):187-+.
29. Schoenfeld D. PARTIAL RESIDUALS FOR THE PROPORTIONAL HAZARDS REGRESSION-MODEL. *Biometrika*. 1982;69(1):239-241.
30. Therneau TM, Grambsch PM. Modeling Survival Data: Extending the Cox Model. New York: Springer; 2000.
31. Kassambara A, Kosinski M. survminer: Drawing Survival Curves using 'ggplot2'; 2018.
32. Stefansson B, Ohama T, Daugherty AE, Brautigan DL. Protein phosphatase 6 regulatory subunits composed of ankyrin repeat domains. *Biochemistry*. 2008;47(5):1442-1451.
33. Kaneta Y, Kagami Y, Tsunoda T, Ohno R, Nakamura Y, Katagiri T. Genome-wide analysis of gene-expression profiles in chronic myeloid leukemia cells using a cDNA microarray. *International Journal of Oncology*. 2003;23(3):681-691.
34. Ishikawa M, Yagasaki F, Okamura D, et al. A novel gene, ANKRD28 on 3p25, is fused with NUP98 on 11p15 in a cryptic 3-way translocation of t(3;5;11)(p25;q35;p15) in an adult patient with myelodysplastic syndrome/acute myelogenous leukemia. *International Journal of Hematology*. 2007;86(3):238-245.
35. Miles AL, Burr SP, Grice GL, Nathan JA. The vacuolar-ATPase complex and assembly factors, TMEM199 and CCDC115, control HIF1 alpha prolyl hydroxylation by regulating cellular on levels. *Elife*. 2017;6:28.
36. Jung N, Dai B, Gentles AJ, Majeti R, Feinberg AP. An LSC epigenetic signature is largely mutation independent and implicates the HOXA cluster in AML pathogenesis. *Nature Communications*. 2015;6:12.
37. Placke T, Faber K, Nonami A, et al. Requirement for CDK6 in MLL-rearranged acute myeloid leukemia. *Blood*. 2014;124(1):13-23.
38. Scheicher R, Hoelbl-Kovacic A, Bellutti F, et al. CDK6 as a key regulator of hematopoietic and leukemic stem cell activation. *Blood*. 2015;125(1):90-101.
39. Baghdassarian N, Ffrench M. Cyclin-dependent kinase inhibitors (CKIs) and hematological malignancies. *Hematology and Cell Therapy*. 1996;38(4):313-323.
40. Goel S, DeCristo MJ, McAllister SS, Zhao JJ. CDK4/6 Inhibition in Cancer: Beyond Cell Cycle Arrest. *Trends in Cell Biology*. 2018;28(11):911-925.
41. Asghar U, Witkiewicz AK, Turner NC, Knudsen ES. The history and future of targeting cyclin-dependent kinases in cancer therapy. *Nature Reviews Drug Discovery*. 2015;14(2):130-146.
42. Bose P, Simmons GL, Grant S. Cyclin-dependent kinase inhibitor therapy for hematologic malignancies. *Expert Opinion on Investigational Drugs*. 2013;22(6):723-738.
43. Reis CR, Chen PH, Bendris N, Schmid SL. TRAIL-death receptor endocytosis and apoptosis are selectively regulated by dynamin-1 activation. *Proceedings of the National Academy of Sciences of the United States of America*. 2017;114(3):504-509.
44. Meng JH. Distinct functions of dynamin isoforms in tumorigenesis and their potential as therapeutic targets in cancer. *Oncotarget*. 2017;8(25):41701-41716.
45. Ribeiro AFT, Pratcorona M, Erpelinck-Verschueren C, et al. Mutant DNMT3A: a marker of poor prognosis in acute myeloid leukemia. *Blood*. 2012;119(24):5824-5831.
46. Ley TJ, Ding L, Walter MJ, et al. DNMT3A Mutations in Acute Myeloid Leukemia. *New England Journal of Medicine*. 2010;363(25):2424-2433.
47. Kuchenbauer F, Kern W, Schoch C, et al. Detailed analysis of FLT3 expression levels in acute myeloid leukemia. *Haematologica-the Hematology Journal*. 2005;90(12):1617-1625.
48. Daver N, Schlenk RF, Russell NH, Levis MJ. Targeting FLT3 mutations in AML: review of current knowledge and evidence. *Leukemia*. 2019;33(2):299-312.
49. Stein EM, Tallman MS. Emerging therapeutic drugs for AML. *Blood*. 2016;127(1):71-78.

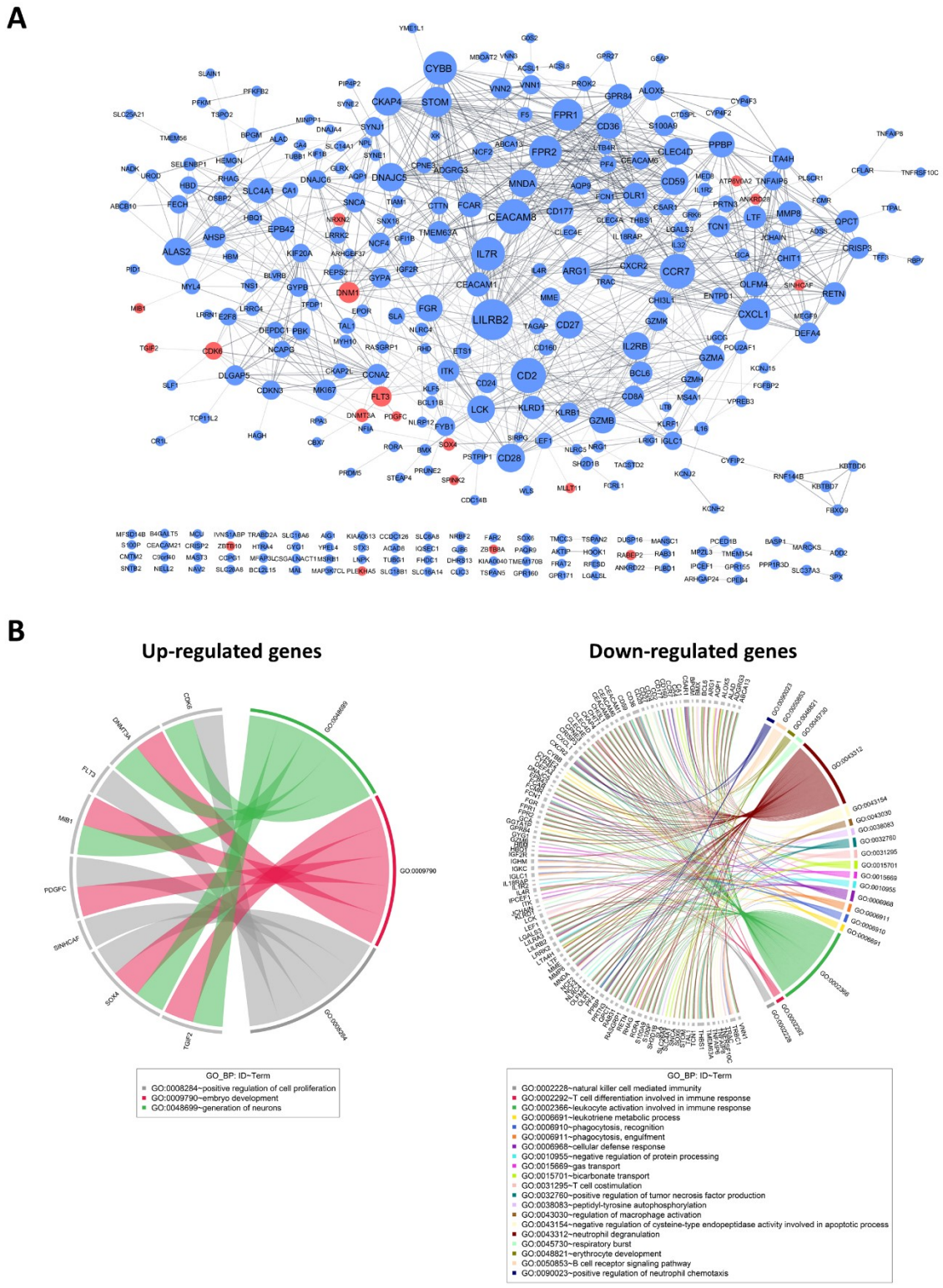
50. Wang L, Lee K, Malonis R, Sanchez I, Dynlacht BD. Tethering of an E3 ligase by PCM1 regulates the abundance of centrosomal KIAA0586/Talpid3 and promotes ciliogenesis. *Elife*. 2016;5:18.
51. Douanne T, Andre-Gregoire G, Thys A, Trillet K, Gavard J, Bidere N. CYLD Regulates Centriolar Satellites Proteostasis by Counteracting the E3 Ligase MIB1. *Cell Reports*. 2019;27(6):1657-+.
52. Koo BK, Lim HS, Song R, et al. Mind bomb 1 is essential for generating functional Notch ligands to activate Notch. *Development*. 2005;132(15):3459-3470.
53. Kim YW, Koo BK, Jeong HW, et al. Defective Notch activation in microenvironment leads to myeloproliferative disease. *Blood*. 2008;112(12):4628-4638.
54. Lampreia FP, Carmelo JG, Anjos-Afonso F. Notch Signaling in the Regulation of Hematopoietic Stem Cell. *Current Stem Cell Reports*. 2017;3(3):202-209.
55. Korn C, Mendez-Ferrer S. Myeloid malignancies and the microenvironment. *Blood*. 2017;129(7):811-822.
56. Heidel FH, Mar BG, Armstrong SA. Self-renewal related signaling in myeloid leukemia stem cells. *International Journal of Hematology*. 2011;94(2):109-117.
57. Akhter A, Farooq F, Elyamany G, et al. Acute Myeloid Leukemia (AML): Upregulation of BAALC/MN1/MLLT11/EVI1 Gene Cluster Relate With Poor Overall Survival and a Possible Linkage With Coexpression of MYC/BCL2 Proteins. *Applied Immunohistochemistry & Molecular Morphology*. 2018;26(7):483-488.
58. Li W, Ji M, Lu F, et al. Novel AF1q/MLLT11 favorably affects imatinib resistance and cell survival in chronic myeloid leukemia. *Cell Death & Disease*. 2018;9:13.
59. Tiberio P, Lozneanu L, Angeloni V, et al. Involvement of AF1q/MLLT11 in the progression of ovarian cancer. *Oncotarget*. 2017;8(14):23246-23264.
60. Jin HL, Sun WR, Zhang YM, et al. MicroRNA-411 Downregulation Enhances Tumor Growth by Upregulating MLLT11 Expression in Human Bladder Cancer. *Molecular Therapy-Nucleic Acids*. 2018;11:312-322.
61. Lee SG, Park TS, Yang JJ, et al. Molecular Identification of a New Splicing Variant of the MLL-MLLT11 Fusion Transcript in an Adult with Acute Myeloid Leukemia and t(1;11)(q21;q23). *Acta Haematologica*. 2012;128(3):131-138.
62. Tserel L, Limbach M, Saare M, et al. CpG sites associated with NRP1, NRXN2 and miR-29b-2 are hypomethylated in monocytes during ageing. *Immunity & Ageing*. 2014;11:5.
63. Bartoschek M, Pietras K. PDGF family function and prognostic value in tumor biology. *Biochemical and Biophysical Research Communications*. 2018;503(2):984-990.
64. Foss B, Ulvestad E, Bruserud O. Platelet-derived growth factor (PDGF) in human acute myelogenous leukemia: PDGF receptor expression, endogenous PDGF release and responsiveness to exogenous PDGF isoforms by in vitro cultured acute myelogenous leukemia blasts (vol 67, pg 267, 2001). *European Journal of Haematology*. 2001;67(4):267-278.
65. Hurst NJ, Najy AJ, Ustach CV, Movilla L, Kim HRC. Platelet-derived growth factor-C (PDGF-C) activation by serine proteases: implications for breast cancer progression. *Biochemical Journal*. 2012;441:909-918.
66. Kinjo T, Sun CH, Ikeda T, et al. Platelet-derived growth factor-C functions as a growth factor in mouse embryonic stem cells and human fibrosarcoma cells. *Cellular & Molecular Biology Letters*. 2018;23:11.
67. Papadopoulos N, Lennartsson J. The PDGF/PDGFR pathway as a drug target. *Molecular Aspects of Medicine*. 2018;62:75-88.
68. Jilaveanu LB, Parisi F, Barr ML, et al. PLEKHA5 as a Biomarker and Potential Mediator of Melanoma Brain Metastasis. *Clinical Cancer Research*. 2015;21(9):2138-2147.
69. Eisele SC, Gill CM, Shankar GM, Brastianos PK. PLEKHA5: A Key to Unlock the Blood-Brain Barrier? *Clinical Cancer Research*. 2015;21(9):1978-1980.

70. Airik R, Schueler M, Airik M, et al. SDCCAG8 Interacts with RAB Effector Proteins RABEP2 and ERC1 and Is Required for Hedgehog Signaling. *Plos One*. 2016;11(5):19.
71. Guo ZY, Wang AL, Zhang WD, et al. PIM inhibitors target CD25-positive AML cells through concomitant suppression of STAT5 activation and degradation of MYC oncogene. *Blood*. 2014;124(11):1777-1789.
72. Vervoort SJ, van Boxtel R, Coffer PJ. The role of SRY-related HMG box transcription factor 4 (SOX4) in tumorigenesis and metastasis: friend or foe? *Oncogene*. 2013;32(29):3397-3409.
73. Lu JW, Hsieh MS, Hou HA, Chen CY, Tien HF, Lin LI. Overexpression of SOX4 correlates with poor prognosis of acute myeloid leukemia and is leukemogenic in zebrafish. *Blood Cancer Journal*. 2017;7:9.
74. Tonks A, Pearn L, Musson M, et al. Transcriptional dysregulation mediated by RUNX1-RUNX1T1 in normal human progenitor cells and in acute myeloid leukaemia. *Leukemia*. 2007;21(12):2495-2505.
75. Zhang H, Alberich-Jorda M, Amabile G, et al. Sox4 Is a Key Oncogenic Target in C/EBP alpha Mutant Acute Myeloid Leukemia. *Cancer Cell*. 2013;24(5):575-588.
76. Munoz IM, MacArtney T, Sanchez-Pulido L, Ponting CP, Rocha S, Rouse J. Family with Sequence Similarity 60A (FAM60A) Protein Is a Cell Cycle-fluctuating Regulator of the SIN3-HDAC1 Histone Deacetylase Complex. *Journal of Biological Chemistry*. 2012;287(39):32346-32353.
77. Streubel G, Fitzpatrick DJ, Oliviero G, et al. Fam60a defines a variant Sin3a-Hdac complex in embryonic stem cells required for self-renewal. *Embo Journal*. 2017;36(15):2216-2232.
78. Biddlestone J, Batie M, Bandarra D, Munoz I, Rocha S. SINHCAF/FAM60A and SIN3A specifically repress HIF-2 alpha expression. *Biochemical Journal*. 2018;475:2073-2090.
79. Paczulla AM, Rothfelder K, Raffel S, et al. Absence of NKG2D ligands defines leukaemia stem cells and mediates their immune evasion. *Nature*. 2019;572(7768):254-+.
80. Eppert K, Takenaka K, Lechman ER, et al. Stem cell gene expression programs influence clinical outcome in human leukemia. *Nature Medicine*. 2011;17(9):1086-U1091.
81. Ng SWK, Mitchell A, Kennedy JA, et al. A 17-gene stemness score for rapid determination of risk in acute leukaemia. *Nature*. 2016;540(7633):433-+.
82. Chen T, Lee TR, Liang WG, Chang WSW, Lyu PC. Identification of trypsin-inhibitory site and structure determination of human SPINK2 serine proteinase inhibitor. *Proteins-Structure Function and Bioinformatics*. 2009;77(1):209-219.
83. Xue C, Zhang J, Zhang G, Xue Y, Zhang G, Wu X. Elevated SPINK2 gene expression is a predictor of poor prognosis in acute myeloid leukemia. *Oncology Letters*. 2019;18(3):2877-2884.
84. Amigo-Jimenez I, Bailon E, Aguilera-Montilla N, Garcia-Marco JA, Garcia-Pardo A. Gene expression profile induced by arsenic trioxide in chronic lymphocytic leukemia cells reveals a central role for heme oxygenase-1 in apoptosis and regulation of matrix metalloproteinase-9. *Oncotarget*. 2016;7(50):83359-83377.
85. Wotton D, Lo RS, Lee S, Massague J. A Smad transcriptional corepressor. *Cell*. 1999;97(1):29-39.
86. Liu WW, Huang MX, Lin WY, Zou QQ, Chen JF, Gao YX. microRNA-34c Suppressed Epithelial to Mesenchymal Transition via TGIF2 in Colon Cancer. *Journal of Biomaterials and Tissue Engineering*. 2018;8(8):1168-1174.
87. Xi L, Zhang YF, Kong SN, Liang W. miR-34 inhibits growth and promotes apoptosis of osteosarcoma in nude mice through targetly regulating TGIF2 expression. *Bioscience Reports*. 2018;38:10.
88. Wang Y, Wang CM, Jiang ZZ, et al. MicroRNA-34c targets TGFB-induced factor homeobox 2, represses cell proliferation and induces apoptosis in hepatitis B virus-related hepatocellular carcinoma. *Oncology Letters*. 2015;10(5):3095-3102.
89. Kim MK, Jeon BN, Koh DI, et al. Regulation of the Cyclin-dependent Kinase Inhibitor 1A Gene (CDKN1A) by the Repressor BOZF1 through Inhibition of p53 Acetylation and Transcription Factor Sp1 Binding. *Journal of Biological Chemistry*. 2013;288(10):7053-7064.

90. Yang XH, Li MY, Wang B, et al. Systematic computation with functional gene-sets among leukemic and hematopoietic stem cells reveals a favorable prognostic signature for acute myeloid leukemia. *Bmc Bioinformatics*. 2015;16:21.
91. Kasiappan R, Jutooru I, Mohankumar K, Karki K, Lacey A, Safe S. Reactive Oxygen Species (ROS)-Inducing Triterpenoid Inhibits Rhabdomyosarcoma Cell and Tumor Growth through Targeting Sp Transcription Factors. *Molecular Cancer Research*. 2019;17(3):794-805.
92. Li X, Pathi SS, Safe S. Sulindac sulfide inhibits colon cancer cell growth and downregulates specificity protein transcription factors. *Bmc Cancer*. 2015;15:11.
93. Hedrick E, Li X, Safe S. Penfluridol Represses Integrin Expression in Breast Cancer through Induction of Reactive Oxygen Species and Downregulation of Sp Transcription Factors. *Molecular Cancer Therapeutics*. 2017;16(1):205-216.

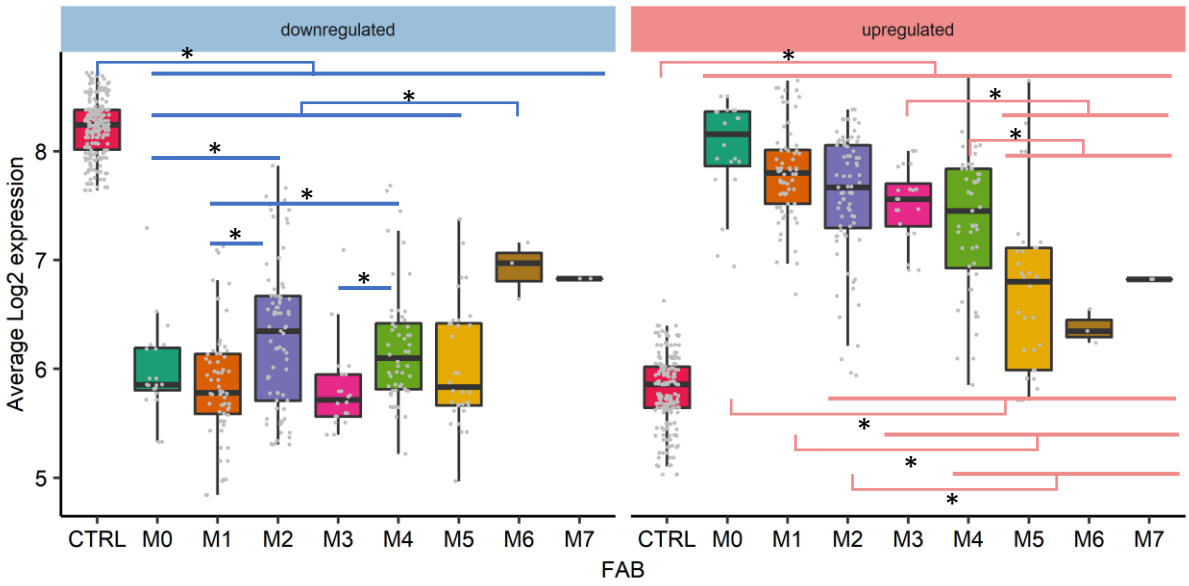


Supplemental Figure 1. Quality control assessment of the samples. (A) Representative figure showing quality control assessment of 500 samples. (B) principal component analysis on batch-adjusted bone marrow samples. colors represent different batches.

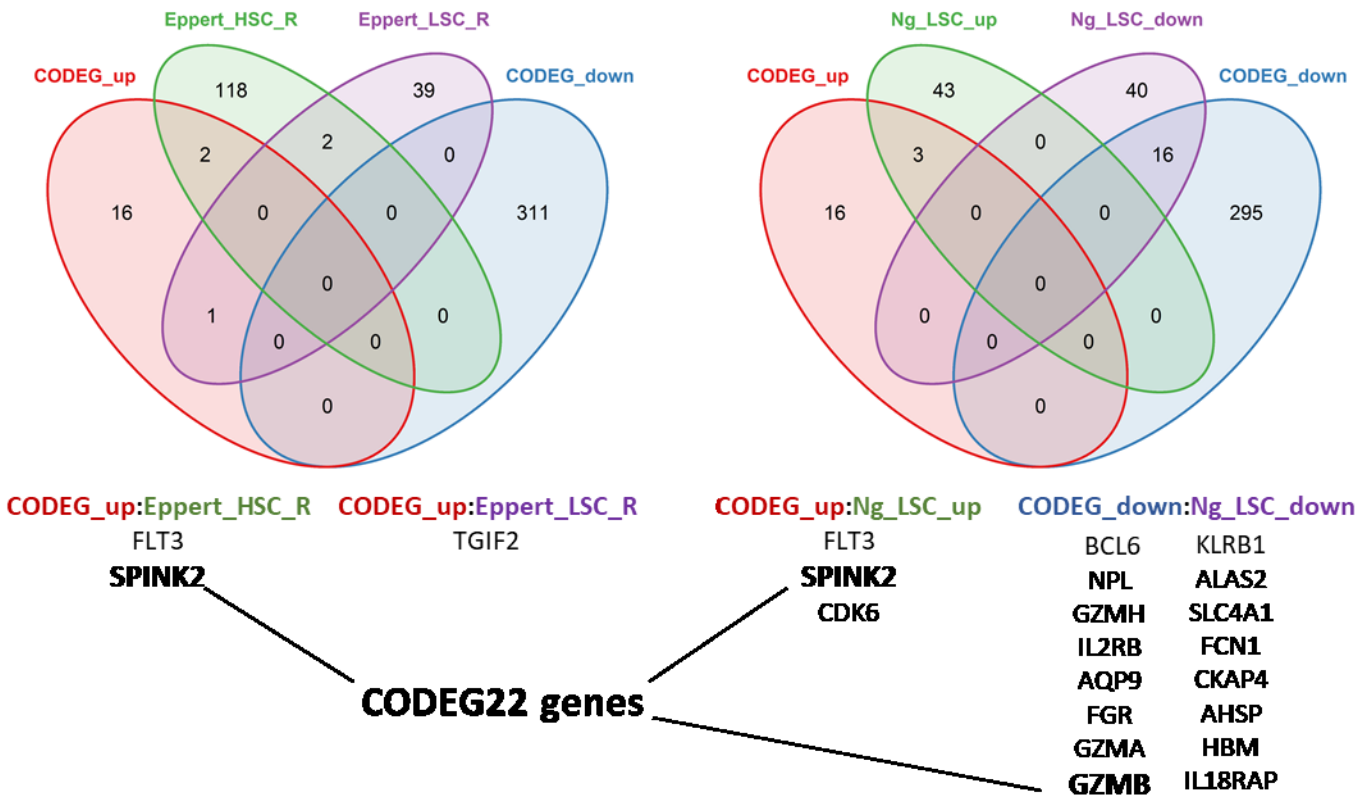


Supplemental Figure 2. Interaction and Gene ontology (GO) enrichment analysis of CODEG genes. (A) Protein-protein interaction network analysis of CODEG genes based on STRINGdb. Node size is proportional to number of undirected edges while edge size and transparency are

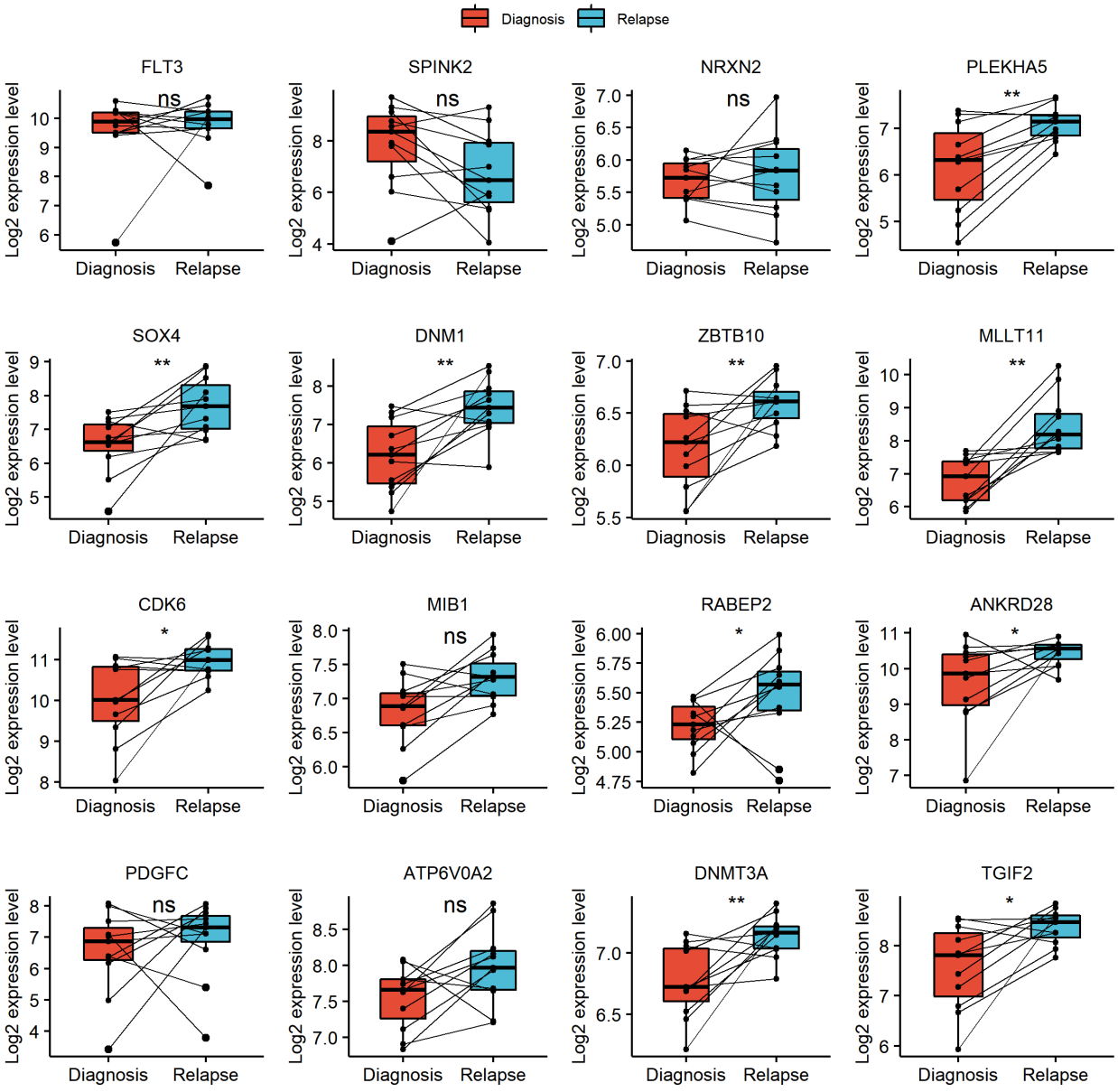
proportional to interaction scores. Red and blue colors are used to label up- and down-regulated genes, respectively. (B) Circos plots visualizing the most significantly enriched “Biological process” GO terms alongside their corresponding genes for both up- and down-regulated subsets.



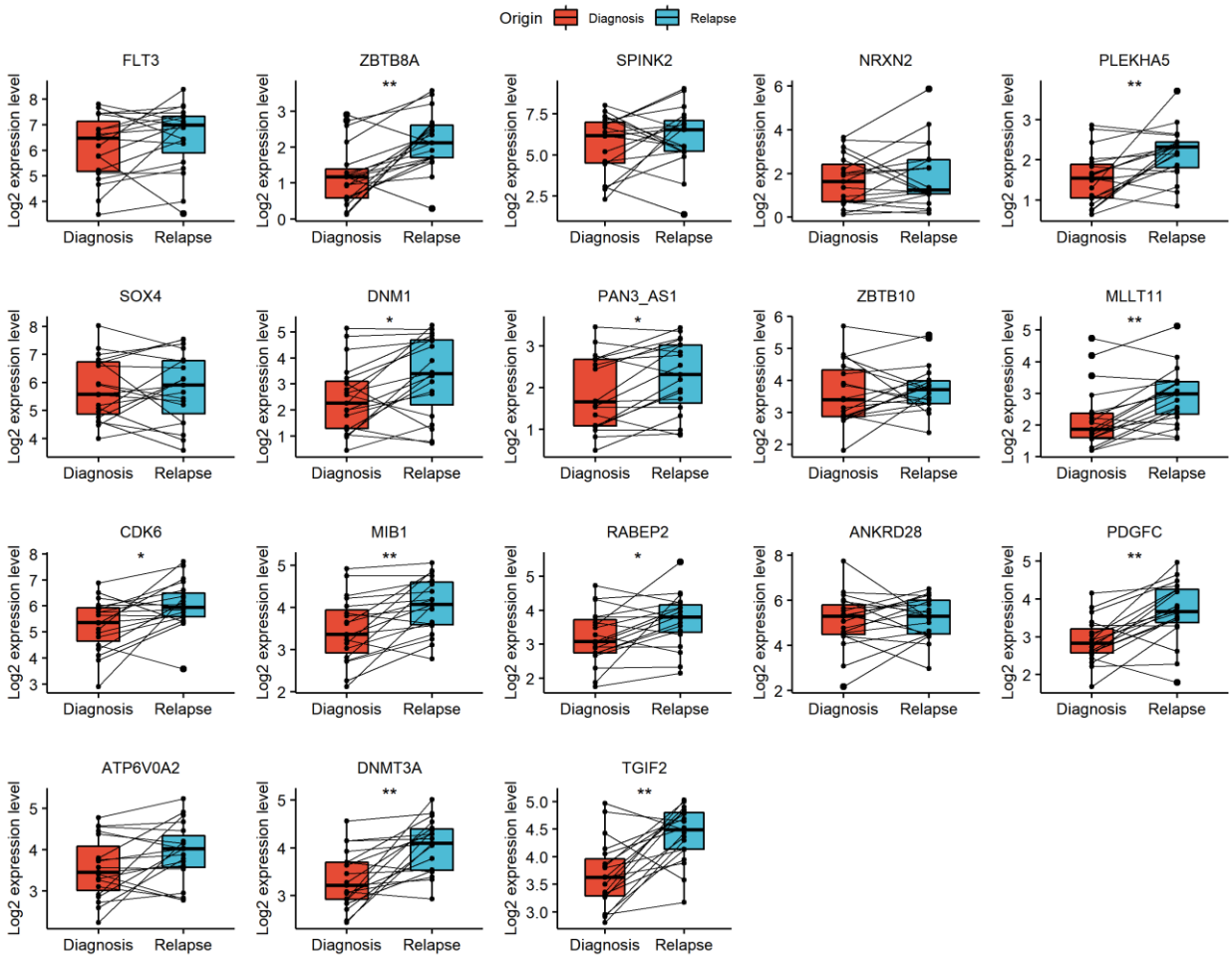
Supplemental Figure 3. Average expression profile of up- and downregulated CODEGs throughout AML maturation. * Wilcoxon test, $p<0.05$.



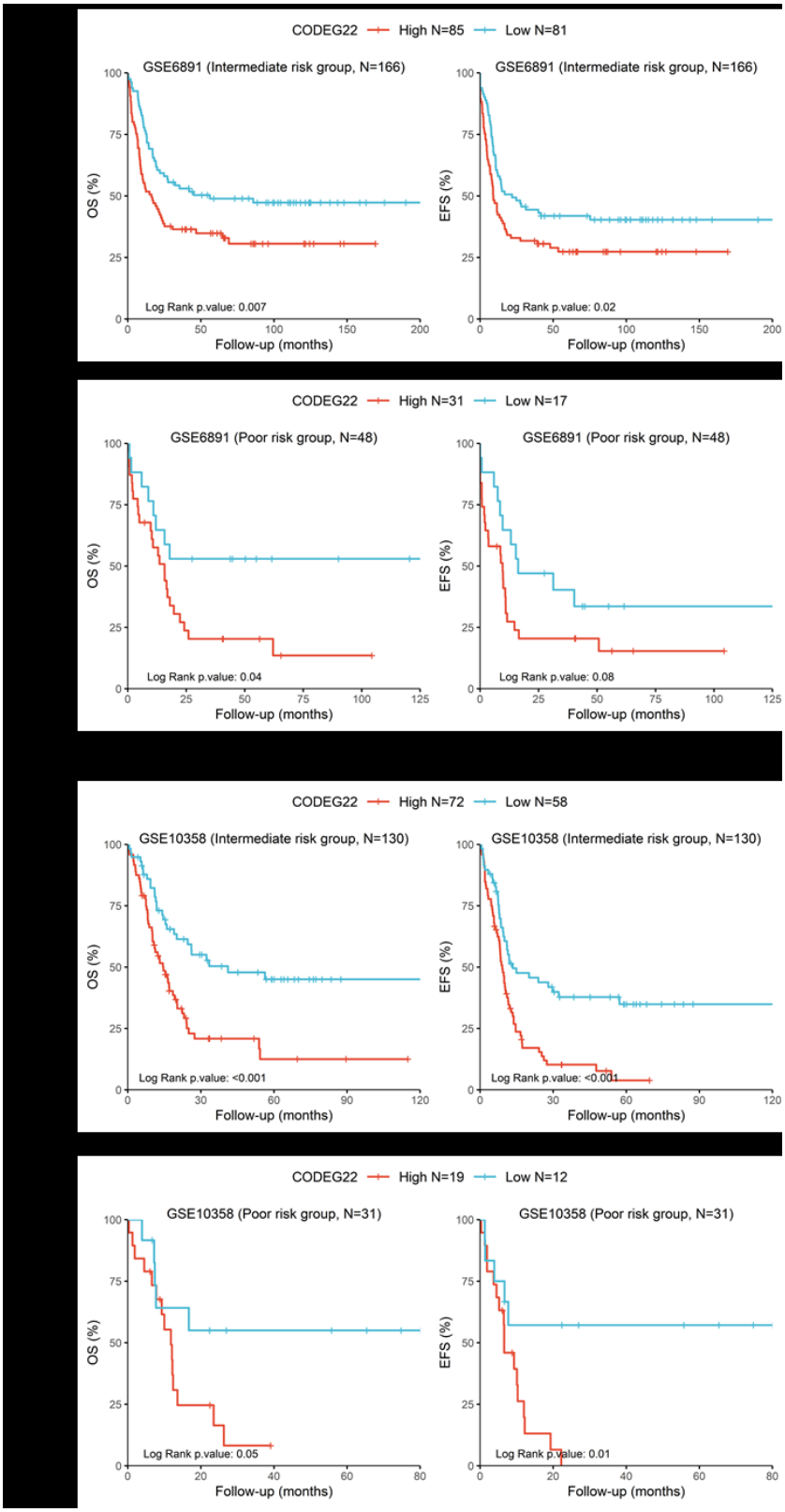
Supplemental Figure 4. Venn diagrams comparing CODEGs to previously reported HSC and LSC signatures.



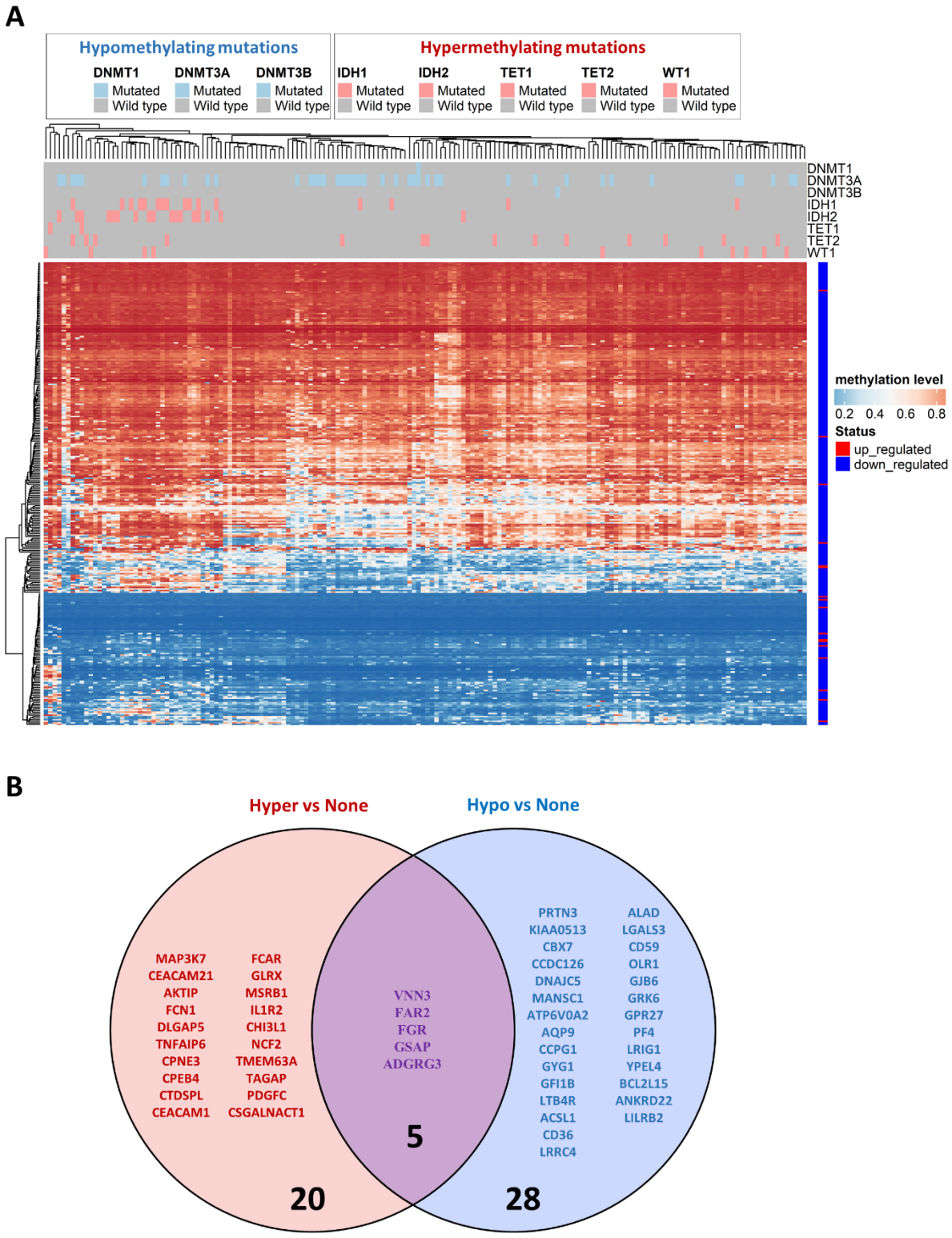
Supplemental Figure 5. Boxplots showing the expression of upregulated CODEG22 genes in paired diagnosis and relapse AML samples from GSE66525. Wilcoxon test: * $p < 0.05$; ** $p < 0.01$; ns, not significant.



Supplemental Figure 6. Boxplots showing the expression of upregulated CODEG22 genes in paired diagnosis and relpase AML samples from GSE83533 (RNA-seq). Wilcoxon test: * $p < 0.05$; ** $p < 0.01$.

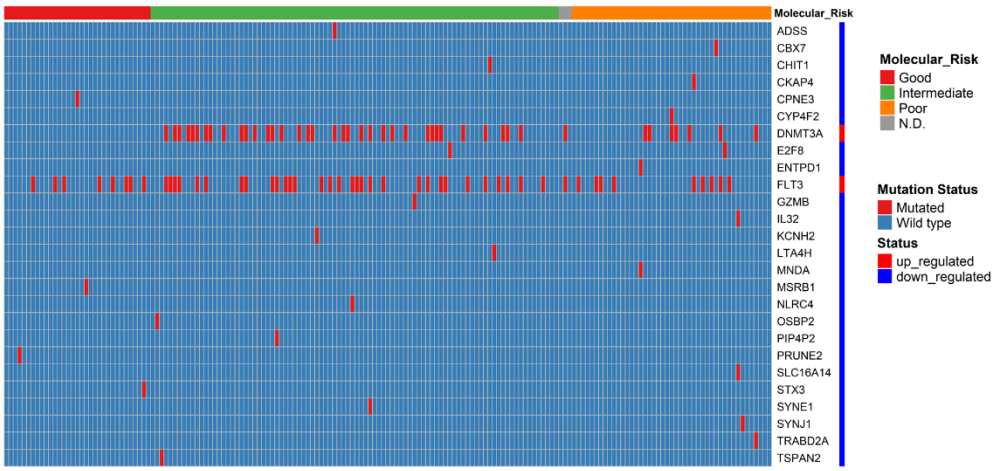


Supplemental Figure 7. OS and EFS analysis of CODEG22 score in intermediate and poor risk groups from A) GSE6891 and B) GSE10358 data sets.

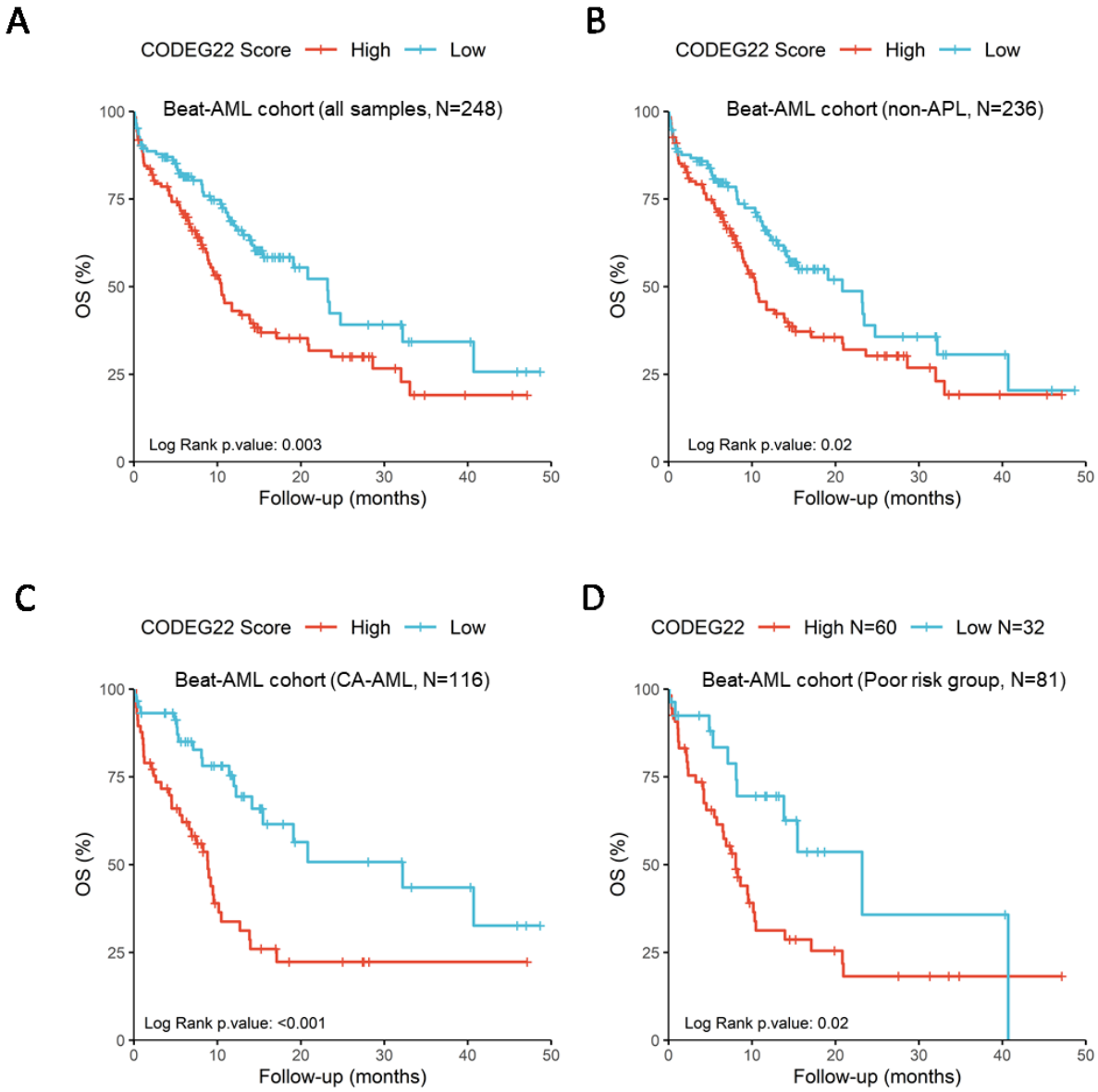


Supplemental Figure 8. Methylation profile of deregulated genes. A) Methylation profile of 271 deregulated genes and correlation with gene expression level in the TCGA AML dataset

(n = 170 samples). The heatmap columns and rows are clustered using Euclidean distance and average method. The mutational profile of many genes known to regulate DNA methylation is presented on top of the heatmap. B) Venn diagram highlighting deregulated genes, which DNA methylation level was associated with the mutation of methylation regulators. Hyper: group of patients harbouring inactivation mutations in the DNA demethylation effectors: IDH1/2, TET1/2 or WT1 with no mutations in the DNMTs. Hypo: group of patients harbouring inactivation mutations in the DNA methyltransferases DNMT1, DNMT3A or DNMT3B with no mutation in the DNA demethylation effectors. None: group of patients without mutations in any of the DNA methylation regulators. Genes in the "Hyper" vs "None" comparison circle (blue) were hyper methylated in "Hyper" compared to "None" group. Genes in the Hypo vs None circle (red) were hypomethylated in "Hypo" compared to "None" group. An alteration in DNA methylation level by 10% with adjusted *P* value <.05 was considered significant.



Supplemental Figure 9. Mutational profile of deregulated genes. The mutational profile of genes that showed at least one missense mutation in the AML dataset from TCGA (N=173) is presented. B) Methylation profile of 271 deregulated genes and correlation with gene expression level in the TCGA AML dataset (n = 170 samples). The heatmap columns and rows are clustered using Euclidean distance and average method. The mutational profile of many genes known to regulate DNA methylation is presented on top of the heatmap.



Supplemental Figure 10. Stratification of patients from the Beat-AML cohort based on high and low CODEG22 score. (A) Overall survival (OS) curves of patients including all cytogenetic abnormalities (n=248). (B) OS curves of non-APL patients (n=236). (C) OS of patients with cytogenetically abnormal AML (CA-AML, n=116). (D) OS of patients from the poor cytogenetic risk group (n=81). CODEG22 scores above and below the median are labelled High (in red) and Low score (in blue), respectively. Log-rank test was used to compare the survival curves of High and Low score subsets.