Regular Article

blood

MYELOID NEOPLASIA

Gene expression and risk of leukemic transformation in myelodysplasia

Yusuke Shiozawa,[1,2,*] Luca Malcovati,[3,4,*] Anna Gallì,[4] Andrea Pellagatti,[5] Mohsen Karimi,[6] Aiko Sato-Otsubo,[2] Yusuke Sato,[2,7] Hiromichi Suzuki,[2] Tetsuichi Yoshizato,[2] Kenichi Yoshida,[2] Yuichi Shiraishi,[8] Kenichi Chiba,[8] Hideki Makishima,[2] Jacqueline Boultwood,[5] Eva Hellström-Lindberg,[6] Satoru Miyano,[8,9] Mario Cazzola,[3,4,†] and Seishi Ogawa[2,†]

[1]Department of Pediatrics, The University of Tokyo, Tokyo, Japan; [2]Department of Pathology and Tumor Biology, Kyoto University, Kyoto, Japan; [3]Department of Molecular Medicine, University of Pavia, Pavia, Italy; [4]Department of Hematology Oncology, Fondazione IRCCS Policlinico San Matteo, Pavia, Italy; [5]Bloodwise Molecular Haematology Unit, Nuffield Division of Clinical Laboratory Sciences, Radcliffe Department of Medicine, University of Oxford, and Oxford BRC Haematology Theme, Oxford, United Kingdom; [6]Department of Medicine, Center for Hematology and Regenerative Medicine, Karolinska Institutet, Stockholm, Sweden; [7]Department of Urology, The University of Tokyo, Tokyo, Japan; and [8]Laboratory of DNA Information Analysis, and [9]Laboratory of Sequence Analysis, Human Genome Center, The Institute of Medical Science, The University of Tokyo, Tokyo, Japan

Key Points

- Through a comprehensive transcriptomic analysis, we discovered 2 major subgroups of myelodysplasia defined by gene expression profiles.
- The gene expression–based subgroups had independent prognostic value, which was validated in an external cohort.

Myelodysplastic syndromes (MDSs) are a heterogeneous group of clonal hematopoietic disorders with a highly variable prognosis. To identify a gene expression–based classification of myelodysplasia with biological and clinical relevance, we performed a comprehensive transcriptomic analysis of myeloid neoplasms with dysplasia using transcriptome sequencing. Unsupervised clustering of gene expression data of bone marrow CD34$^+$ cells from 100 patients identified 2 subgroups. The first subtype was characterized by increased expression of genes related to erythroid/megakaryocytic (EMK) lineages, whereas the second subtype showed upregulation of genes related to immature progenitor (IMP) cells. Compared with the first so-called EMK subtype, the IMP subtype showed upregulation of many signaling pathways and downregulation of several pathways related to metabolism and DNA repair. The IMP subgroup was associated with a significantly shorter survival in both univariate (hazard ratio [HR], 5.0; 95% confidence interval [CI], 1.8-14; $P < .001$) and multivariate analysis (HR, 4.9; 95% CI, 1.3-19; $P = .02$). Leukemic transformation was limited to the IMP subgroup. The prognostic significance of our classification was validated in an independent cohort of 183 patients. We also constructed a model to predict the subgroups using gene expression profiles of unfractionated bone marrow mononuclear cells (BMMNCs). The model successfully predicted clinical outcomes in a test set of 114 patients with BMMNC samples. The addition of our classification to the clinical model improved prediction of patient outcomes. These results indicated biological and clinical relevance of our gene expression–based classification, which will improve risk prediction and treatment stratification of MDS. (*Blood.* 2017; 130(24):2642-2653)

Introduction

Myelodysplastic syndrome (MDS) and related myeloid disorders (myelodysplasia) are a heterogeneous group of clonal hematopoietic disorders that are characterized by peripheral blood cytopenias with dysplastic marrow morphology and an increased risk of transformation to acute myeloid leukemia (AML; secondary AML).[1,2] Their prognostic profile is highly variable, with survival ranging from a few months to >10 years,[3] underscoring the importance of predicting clinical outcomes for treatment stratification. Several scoring systems have been developed on the basis of known prognostic factors, including percentage of marrow blasts, degree of cytopenias, cytogenetic abnormalities, and transfusion requirement.[4-8] The effects of gene mutations on clinical outcomes have recently been investigated in large cohorts of myelodysplasia patients and incorporated into a prognostic model.[9-11]

Gene expression profiling provides a systematic approach for identifying tumor subtypes with prognostic significance and have successfully been applied to the identification of the *BCR-ABL1*-like subtype of acute lymphoblastic leukemia, activated B-cell–like diffuse large B-cell lymphoma, and basal-like breast cancer.[12-15] Recently, several groups performed gene expression profiling of myelodysplasia using microarray platforms, based on which new prognostic models have been proposed.[16-19] However, unlike the case with other cancers, these models were not constructed in an unbiased manner and do not represent biologically distinct subsets of patients. Gene expression

Submitted 4 May 2017; accepted 11 October 2017. Prepublished online as *Blood* First Edition paper, 2 November 2017; DOI 10.1182/blood-2017-05-783050.

*Y. Shiozawa and L.M. contributed equally to this study.

†M.C. and S.O. contributed equally to this study.

The online version of this article contains a data supplement.

The publication costs of this article were defrayed in part by page charge payment. Therefore, and solely to indicate this fact, this article is hereby marked "advertisement" in accordance with 18 USC section 1734.

profiles similar to de novo AML were shown to be associated with leukemic transformation of MDS.[19] This raises the possibility that, according to gene expression profiles, myelodysplasia may be subdivided into a subtype with an indolent clinical course and that are at high risk of clonal evolution.

In this study, we performed comprehensive genomic and transcriptomic analyses of 214 patients with myelodysplasia using transcriptome sequencing and targeted-capture sequencing of myelodysplasia-related genes. Through unsupervised class discovery, we have revealed 2 discrete subtypes of myelodysplasia that were characterized by unique gene expression signatures and mutation patterns and, moreover, distinct prognostic profiles.

Methods

Patients and samples

This study was approved by the ethics committees of the Fondazione IRCCS Policlinico San Matteo (Pavia), Karolinska Institutet (Stockholm), and Kyoto University (Kyoto). We enrolled 214 patients with MDS (n = 152), myelodysplastic/myeloproliferative neoplasm (n = 44), and AML with myelodysplasia-related changes (AML-MDS; n = 18) who had been followed at the Department of Hematology, University of Pavia and Fondazione IRCCS Policlinico San Matteo, Pavia (Table 1). Bone marrow mononuclear cells (BMMNCs; n = 165) and/or bone marrow $CD34^+$ cells (n = 100) were obtained from 214 patients, 51 of whom were analyzed for both cell fractions (Figure 1A). $CD34^+$ cells were isolated using magnetic-activated cell sorting separation columns (Miltenyi Biotec, Bergisch Gladbach, Germany).[20] Samples were collected 0 to 281 months after diagnosis, between April 2004 and June 2013. No treatment other than supportive care was given prior to sample collection. All patients were reclassified according to the 2016 revision to the World Health Organization (WHO) classification of myeloid neoplasms and acute leukemia.[21] BMMNCs (StemCell Technologies, Vancouver, Canada) and $CD34^+$ cells (Lonza, Basel, Switzerland) of 3 healthy adults each were used as controls.

RNA sequencing

RNA was extracted with TRIzol reagent (Life Technologies, Carlsbad, CA) and treated with DNase I (Qiagen, Hilden, Germany). RNA integrity numbers were confirmed to be >7 using the TapeStation (Agilent Technologies, Palo Alto, CA). The RNA-sequencing libraries were prepared from poly(A)-selected RNA using the NEBNext Ultra RNA Library Prep kit for Illumina (New England BioLabs, Ipswich, MA). Libraries were sequenced using the HiSeq 2000 or 2500 platform according to a standard 100-bp paired-end read protocol (Illumina, San Diego, CA).

Sequencing reads were aligned to the human reference genome (hg19) using RUM version 2.0.4.[22] Fusion transcripts were detected by Genomonfusion (http://genomon.hgc.jp/rna/),[23] followed by reverse transcription polymerase chain reaction and direct sequencing of the polymerase chain reaction products. Differential expression analysis was conducted using edgeR version 3.6.8.[24] The analysis was confined to those genes expressed at >1 counts per million (CPM) in >5 samples. Generalized linear models were used to compare gene expression data. Correction for multiple testing was done by the Benjamini-Hochberg method, in which q-value < 0.01 was considered significant. High *EVI1* expression was defined as CPM >100 and >20 in $CD34^+$ cells and BMMNCs, respectively. ROAST implemented in the R package limma version 3.22.4 was used to find the significantly upregulated pathways, where the pathways defined by the Kyoto Encyclopedia of Genes and Genomes were tested.[25] Genes specifically expressed in each hematopoietic stem/progenitor population were adopted from a previous report.[26] Genome data have been deposited in the European Genome-phenome Archive (http://www.ebi.ac.uk/ega/) under accession number EGAS00001002346.

Clustering of gene expression data

Consensus clustering was performed based on $\log_2$(CPM) values of genes that showed larger dispersions than genes with similar mean expression levels,[27] where K-means clustering based on the Ward and the Euclidean distance was repeated 1000 times by randomly sampling 80% of the entire sample set using the R package ConsensusClusterPlus. The number of clusters was determined from the relative change in the area under the cumulative distribution function curve. Clustering was also performed based on a nonnegative matrix factorization algorithm using the R package NMF.

A classifier of the gene expression–based subgroups was constructed using the 100 $CD34^+$ cell samples as a training set (Figure 1A). For applicability to gene expression microarray data, the model only included moderately expressed genes showing 50 to 75 percentile of mean signals among all the genes in 183 MDS patients analyzed on GeneChip Human Genome U133 Plus 2.0 array (Affymetrix).[28] We used elastic net logistic regression in which the parameters were determined as $a = 0.45$ and $\lambda = 0.111$ using 10-fold cross-validation on the training set. The mean classification error was 0.05. Another classifier was constructed using BMMNC samples from 51 patients, who had been assigned to the subgroups by the gene expression data of their $CD34^+$ cells (Figure 1A; supplemental Methods, available on the *Blood* Web site). Since unfractionated BMMNCs contain mature hematopoietic cells and contaminating peripheral blood, variables in the model were selected from highly upregulated genes in the second subgroup of this training cohort. We first performed rigorous differential expression analysis using the generalized linear model likelihood ratio test.[24] Elastic net logistic regression was then applied to the 30 most significantly upregulated genes in the second subgroup as compared with the first subgroup and the healthy adults. The parameters of the elastic net were determined as $a = 0.7$ and $\lambda = 0.0176$ using 10-fold cross-validation, with a mean classification error of 0.10.

Targeted DNA sequencing

Genomic DNA was available for 211 of the 214 patients (99%) and was extracted from peripheral blood granulocytes (n = 111), BMMNCs (n = 56), bone marrow polymorphonuclear cells (n = 43), or bone marrow $CD34^+$ cells (n = 1). Nine DNA samples were subjected to whole-genome amplification. Sequencing libraries were prepared from 200 to 1000 ng DNA. Target capture was performed using a SureSelect custom kit (Agilent Technologies). RNA baits were designed using SureDesign (Agilent Technologies) to capture coding exons from 89 known or putative driver genes in myelodysplasia (supplemental Table 1, available on the *Blood* Web site) and 1674 single-nucleotide polymorphisms.

Libraries were sequenced using the HiSeq 2000 or 2500 platform with a standard 100-bp paired-end read protocol (Illumina). Sequencing reads were aligned to the human reference genome (hg19) using BWA version 0.7.10. Oncogenic variants were identified as previously described.[10,29] Genomic copy-number analysis was performed based on the sequencing depths of the target regions compared with those of pooled controls.

Gene expression microarray

Gene expression profiling using the GeneChip Human Genome U133 Plus 2.0 array (Affymetrix) was performed for an independent cohort of 183 MDS patients (Figure 1A; supplemental Table 2).[28] Gene expression levels of each gene were obtained through preprocessing by GeneChip Robust Multiarray Analysis (GC-RMA)[30] and selection of a representative probe set by the JetSet annotations,[31] as described previously.[32]

Statistical analysis

Numerical and categorical variables were compared using the Mann-Whitney test and Fisher's exact test, respectively. Overall survival analyses were performed with the Kaplan-Meier method and log-rank test. Patients who had already developed leukemia at the time of sampling were removed from the analysis of leukemia-free survival, with nonleukemic death treated as a competing risk.[33] Multivariate analyses were performed using Cox proportional hazards regression for overall survival and competing risk regression for leukemia-free survival, both with stepwise selection based on the Akaike information criterion (AIC) score. Tested variables were the 2 gene expression–based

Table 1. Patient characteristics

	Patients with CD34$^+$ cell samples (n = 100)			Patients with BMMNC samples only (n = 114)		
	First (EMK) subgroup	Second (IMP) subgroup	*P*	First (EMK) subgroup	Second (IMP) subgroup	*P*
Number of cases	61	39	—	71	43	—
Age (y), median (range)	69 (32–87)	62 (30–83)	.01	67 (30-83)	67 (39-91)	.46
Sex (male/female)	38/23	27/12	.52	42/27	29/18	1
Diagnosis, n (%)						
MDS	37 (61)	22 (56)	—	57 (80)	36 (84)	—
MDS-SLD	4 (6.6)	1 (2.6)	.65	5 (7.0)	1 (2.3)	.41
MDS-RS-SLD	13 (21)	1 (2.6)	.008	19 (27)	1 (2.3)	<.001
MDS-MLD	5 (8.2)	4 (10)	.73	13 (18)	17 (40)	.02
MDS-RS-MLD	5 (8.2)	2 (5.1)	.70	9 (13)	1 (2.3)	.09
MDS with isolated del(5q)	2 (3.3)	0 (0)	.52	1 (1.4)	1 (2.3)	1
MDS-EB-1	4 (6.6)	5 (13)	.31	7 (9.9)	6 (14)	.55
MDS-EB-2	4 (6.6)	9 (23)	.03	3 (4.2)	9 (22)	.009
MDS/MPN	24 (39)	6 (15)	—	13 (18)	1 (2.3)	—
CMML-1	19 (31)	5 (13)	.05	8 (11)	0 (0)	.02
CMML-2	1 (1.6)	0 (0)	1	0 (0)	0 (0)	1
MDS/MPN-RS-T	4 (6.6)	0 (0)	.15	3 (4.2)	1 (2.3)	1
MDS/MPN-U	0 (0)	1 (2.6)	.39	2 (2.8)	0 (0)	.53
AML-MDS	0 (0)	11 (28)	<.001	1 (1.4)	6 (14)	.01
Hemoglobin (g/dL), median (range)	10.4 (7.0-14.7)	10.2 (7.0-14.4)	.92	9.8 (6.3-15.5)	9.7 (6.8-13.9)	.43
ANC ($\times 10^9$/L), median (range)	3.0 (0.35-13)	1.7 (0.40-32)	.06	2.1 (0.20-22)	1.3 (0.20-11)	.002
Platelet count ($\times 10^9$/L), median (range)	173 (17.5-895)	76.0 (15.0-697)	<.001	175 (16.0-849)	102 (12.5-939)	.03
Myeloid/erythroid ratio, median (range)	2.0 (0.50–10)	2.0 (0.13–10)	.86	2.0 (0.25-10)	1.5 (0.20-10)	.71
Bone marrow blasts (%), median (range)	2 (0–15)	11 (1–90)	<.001	2 (0-86)	4 (0-63)	.003
Bone marrow ring sideroblasts (%), median (range)	9 (0–94)	5 (0–78)	.20	13 (0–94)	0.3 (0–81)	.08
IPSS-R in the patients with MDS, n (%)						
Very low	0 (0)	0 (0)	1	0 (0)	0 (0)	1
Low	15 (25)	3 (7.7)	.04	26 (37)	10 (23)	.15
Intermediate	12 (20)	7 (18)	1	17 (24)	8 (19)	.64
High	7 (11)	5 (13)	1	9 (13)	10 (23)	.19
Very high	3 (4.9)	7 (18)	.04	3 (4.2)	6 (14)	.08
Missing	0 (0)	0 (0)	1	2 (2.8)	2 (4.7)	.63
Treatment during the follow-up period,* n (%)						
Allogeneic stem cell transplantation	3 (4.9)	7 (18)	.04	4 (5.6)	4 (9.3)	.47
Cytotoxic chemotherapy	0 (0)	7 (18)	<.001	2 (2.8)	6 (14)	.05
Azacitidine	3 (4.9)	2 (11)	1	3 (4.2)	4 (9.3)	.42
Response	1 (33)	0 (0)	1	0 (0)	1 (25)	1
Only supportive care	56 (92)	26 (67)	.003	63 (89)	31 (72)	.04

AML-MDS, AML with myelodysplasia-related changes; ANC, absolute neutrophil count; CMML, chronic myelomonocytic leukemia; MDS-EB, MDS with excess of blasts; MDS-MLD, MDS with multilineage dysplasia; MDS/MPN, myelodysplastic/myeloproliferative neoplasm; MDS/MPN-RS-T, MDS/MPN with ring sideroblasts and thrombocytosis; MDS/MPN-U, MDS/MPN, unclassifiable; MDS-RS-SLD, MDS with ring sideroblasts with single-lineage dysplasia; MDS-SLD, MDS with single-lineage dysplasia; MDS-RS-MLD, MDS with ring sideroblasts with multilineage dysplasia.

*Some patients received various treatments (eg, cytotoxic chemotherapy followed by allogeneic stem cell transplantation). This led to the sums of patients exceeding 100%.

subgroups, age, sex, and the prognostic variables in the revised international prognostic scoring system for MDS (IPSS-R): percentage of marrow blasts, cytogenetic abnormalities, hemoglobin, absolute neutrophil count, and platelet levels. Cytogenetic abnormalities were classified according to the MDS Cytogenetic Scoring System.[5] Comparison between models was performed by means of the AIC score and c-index. Analyses were performed using R versions 2.15.3 (model construction using elastic net logistic regression) and 3.0.1 (the other analyses). The R scripts are shown in supplemental Methods.

Results

Identification of unique expression clusters in myelodysplasia

To identify discrete subtypes of myelodysplasia, RNA-sequencing data were analyzed using consensus clustering. Clustering analysis of the data from bone marrow CD34+ cell samples (n = 100) revealed 2 stable clusters (Figure 1B); the clustering stability was decreased for >2 clusters (supplemental Figure 1A-B), which indicated that further subdivision of the 2 clusters was not feasible. Nonnegative matrix factorization, another algorithm for class discovery, also supported a 2-class split of the gene expression data (supplemental Figure 1C-D). By contrast, no stable clusters were detected in the analysis of unfractionated BMMNC samples (n = 165; supplemental Figure 1E-F).

The 2 subgroups had a distinct hematological picture (Table 1). WHO subtypes with increased ring sideroblasts were more enriched in the first subgroup (odds ratio, 6.8; 95% confidence interval [CI], 5.5-8.1; $P = .002$). In contrast, patients with advanced disease (ie, MDS-EB-1/2 and AML-MDS) were more frequently clustered into the second subgroup (odds ratio, 12; 95% CI, 4.4-32; $P < .001$). They also had lower platelet counts (median 76.0×10^9/L vs 173×10^9/L; $P < .001$) and

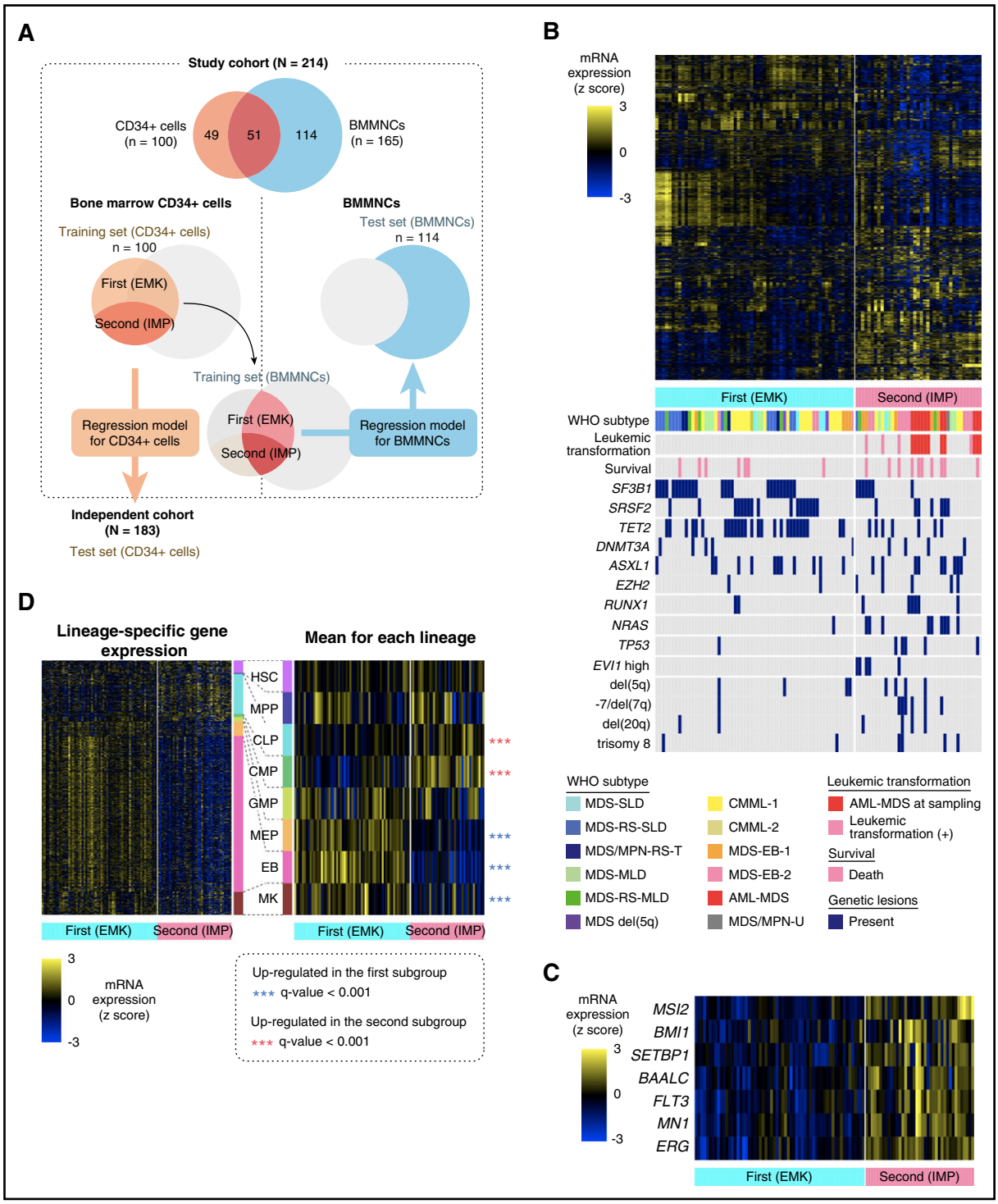


Figure 1. Study design and characterization of the gene expression–based subgroups of myelodysplasia. (A) Schematic depicting the study design. Venn diagrams show sources of RNA. The left and right halves indicate outlines of the analyses for bone marrow CD34$^+$ cells and BMMNCs, respectively. Unsupervised clustering was performed on gene expression data of bone marrow CD34$^+$ cells from 100 patients (training set of CD34$^+$ cell samples), revealing 2 distinct subgroups. A regression model was constructed from the training set, followed by validation in an independent cohort (lower left). A regression model to predict the subgroups using BMMNC samples was also constructed from gene expression data of 51 patients with both CD34$^+$ cell and BMMNC samples (training set of BMMNC samples). Prognostic significance of the model was tested in 114 patients with only BMMNC samples. (B) A heatmap shows expression levels of 3141 genes with high variability in 100 CD34$^+$ cell samples. Each row represents 1 gene, and each column represents 1 sample. Gene expression–based subgroups, WHO subtypes, genetic lesions, and patients' prognosis are shown below the heatmap. AML-MDS, AML with myelodysplasia-related changes; CMML, chronic myelomonocytic leukemia; MDS-EB, MDS with excess blasts; MDS-MLD, MDS with multilineage dysplasia; MDS/MPN-RS-T, MDS/MPN with ring sideroblasts and thrombocytosis; MDS/MPN-U, MDS/MPN, unclassifiable; MDS-RS-SLD, MDS with ring sideroblasts with single lineage dysplasia; MDS-SLD, MDS with single-lineage dysplasia; MDS-RS-MLD, MDS with ring sideroblasts with multilineage dysplasia. (C) A heatmap of expression levels of 7 genes of known prognostic significance in 100 CD34$^+$ cell samples. (D) Expression levels of genes related to specific hematopoietic lineages. The left panel is a heatmap of gene expression levels in 100 CD34$^+$ cell samples. Rows represent genes sorted according to hematopoietic lineages in which they are specifically expressed. Columns represent samples along with their gene expression–based subgroups and WHO subtypes. The middle panel represents mean z scores for each hematopoietic lineage. CLP, common lymphoid progenitor; CMP, common myeloid progenitor; EB, erythroblast; GMP, granulocyte monocyte progenitor; HSC, hematopoietic stem cell; MPP, multipotent progenitor; MEP, megakaryocyte/erythrocyte progenitor; MK, megakaryocyte.

higher percentages of bone marrow blasts (median 11% vs 2%, $P < .001$) as compared with the first subgroup.

A marked difference between the 2 subgroups was found at the transcriptome level. The patients in the second subgroup showed higher expression of the genes known to be associated with poor prognosis in myeloid malignancies (*MSI2*, *BMI1*, *SETBP1*, *BAALC*, *FLT3*, *MN1*, and *ERG*; q-value < 0.01) (Figure 1C).[34-39] Pathway analysis of differentially expressed genes between both subgroups further characterized their expression profiles. Compared with the first subgroup, 19 and 75 pathways were up- and downregulated in the second subgroup, respectively (supplemental Table 3). Cell signaling accounted for 12 of 19 upregulated pathways (63%) in the second subgroup, including MAPK, phosphatidylinositol 3-kinase, Notch, and JAK/STAT pathways (q-value < 0.01). Of 75 downregulated pathways, 46 (61%) were related to metabolism and 5 (6.7%) were to DNA repair. The 2 subgroups also exhibited contrasting gene expression profiles in terms of the commitment to specific hematopoietic lineages.[26] The first subgroup showed increased expression of the sets of genes specifically expressed in erythroblasts and megakaryocyte/erythrocyte progenitors (Figure 1D). Mean expression levels of the erythroid genes were not significantly correlated with bone marrow myeloid/erythroid ratios ($r = -0.16$; 95% CI, −0.36 to 0.05; $P = .13$; supplemental Figure 2), suggesting that a strong erythroid signature in the $CD34^+$ fraction did not simply reflect erythroid hyperplasia. In the second subgroup, by contrast, expression of these genes was significantly decreased, even compared with that in the normal individuals (supplemental Figures 3 and 4), whereas genes characteristic of more immature hematopoietic lineages were upregulated (Figure 1D). We hereafter refer to the 2 subgroups as the subgroup enriched with erythroid/megakaryocytic (EMK) signatures and that with immature progenitor (IMP) signatures.

Prognostic significance of the gene expression–based classification

As expected from the association with many adverse prognostic features, the IMP subtype was characterized by poor clinical outcomes; in univariate analysis, the IMP subgroup was significantly associated with an inferior overall survival compared with the EMK subgroup (HR, 5.0; 95% CI, 1.8-14; $P < .001$) (Figure 2A). In addition to the patients who had already developed leukemia at the time of sampling, the patients who later progressed to leukemia were also limited to the IMP subgroup (Figures 1B and 2B). The negative impact on the survival of the IMP subgroup was still observed after adjustment for the effects of known risk factors using multivariate analysis (HR, 4.9; 95% CI, 1.3-19; $P = .02$) (Table 2).[5]

The strong effects of the gene expression subtype on survival were validated in an independent cohort.[17] Because gene expression levels were analyzed using a different platform (the Affymetrix GeneChip Human Genome U133 Plus 2.0 array), in this external cohort, we developed a classifier of the 2 expression subgroups relying on the expression of a small number of genes so that it can be used across different assay platforms. Elastic net logistic regression was first performed on gene expression data of $CD34^+$ cells in our cohort as a training set (Figure 1A). Elastic net parameters were determined by 10-fold cross-validation, yielding a logistic regression model that predicts the subgroups based on expression levels of 68 genes (Figure 3A; supplemental Table 4). The regression model was then applied to the microarray-based expression data of bone marrow $CD34^+$ cells from the external cohort of 183 patients with MDS.[28] According to this model, 116 (63%) and 67 (37%) patients were classified into the EMK and IMP subgroups, respectively (Figure 3B). The gene expression signatures of hematopoietic lineages in the 2 subgroups were largely recapitulated in this validation data set: upregulation of the genes related to erythroid lineages in the EMK subgroup and that of the progenitor signatures in IMP (supplemental Figure 5). Patients assigned to the IMP subtype were shown to have significantly shorter survival than those in the EMK subgroup (HR, 3.0; 95% CI, 1.9-5.1; $P < .001$) (Figure 3C). The negative impact of the IMP expression profile became more prominent, when the association was tested for leukemic transformation (HR, 6.8; 95% CI, 3.2-14; $P < .001$) (Figure 3D). The difference in survival and leukemic transformation remained significant even after the effect of bone marrow blasts was adjusted (supplemental Figure 6). These results confirmed the reproducibility of the prognostic value of the expression profiles and the robustness of this set of classifiers, regardless of the assay used for gene expression profiling.

Prediction of the gene expression–based subgroups using BMMNC samples

To facilitate the clinical use of this molecular classification without relying on $CD34^+$ cell selection, we constructed a classifier for the 2 expression subtypes using the data from unfractionated BMMNCs. Among the 100 patients who had been assigned to the EMK or IMP subgroups on the basis of the gene expression data of $CD34^+$ cells, 51 were also analyzed by RNA sequencing for BMMNCs and thus used as a training cohort (Figures 1A and 4A). Through 10-fold cross-validation on this training set, we developed a logistic regression model to predict the subgroups based on the expression levels of 9 genes (Figure 4B). The model was applied to the gene expression data of BMMNCs in the remaining 114 cases, of whom 71 (62%) and 43 (38%) were predicted to be the EMK and IMP subgroup, respectively (supplemental Figure 7). Compared with the predicted EMK subgroup, the IMP subgroup was associated with a significantly shorter survival in univariate analysis (HR, 4.5; 95% CI, 2.0-10; $P < .001$) (Figure 4C). Again, association was more pronounced for leukemic transformation (HR, 7.3; 95% CI, 2.0-26; $P = .002$) (Figure 4D) than for overall survival. Multivariate analysis also demonstrated that the predicted IMP subgroup was independently associated with overall survival (HR, 2.9; 95% CI, 1.1-7.6; $P = .03$) (Table 2) and leukemic transformation (HR, 5.9; 95% CI, 1.6-22; $P = .008$) (Table 2). These results indicated the prognostic value of the classification based on the gene expression profiles of BMMNCs.

Genetic alterations in the gene expression subgroups

The 2 expression subgroups also had unique profile of genetic alterations. We interrogated myelodysplasia-related mutations as well as chromosomal lesions identified by targeted-capture sequencing with a mean coverage of 1009× (range, 207-2306×). In total, we identified 313 nonsynonymous single-nucleotide variants, 147 small insertion–deletions in common targets of myeloid neoplasms, and 170 copy-number abnormalities and/or allelic imbalances (supplemental Figure 8). In addition, RNA sequencing detected aberrant gene fusions in 4 patients, of whom 3 had *EVI1*-containing fusions, including *NRIP1-EVI1* (n = 2) and *RUNX1-EVI1* (n = 1). Elevated *EVI1* expression was also found in an additional 3 patients with no accompanying gene fusions, of whom 1 harbored a 3q26 abnormality (supplemental Figure 9). Of interest, most of these *EVI1*-overexpressed cases (5 out of 6) were accompanied by mutated *SF3B1*. All combined, 1 or more genetic lesions were identified in 194 patients with a median of 3 (0-12) lesions per case.

The number of genetic lesions was significantly higher in the IMP subgroup (median, 3 [range, 0-12] vs 2 [0-7], $P = .003$). The prevalence of individual genetic lesions also showed substantial variation between the 2 expression subtypes. *SF3B1* mutation was more frequent

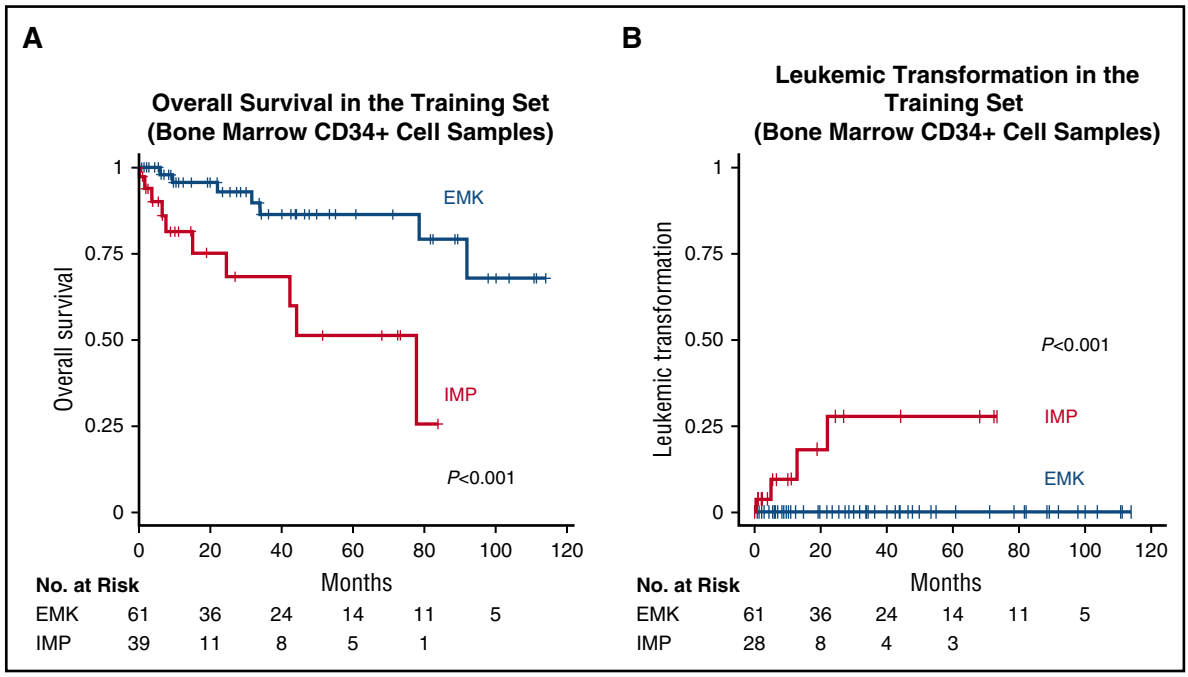


Figure 2. Prognostic significance of the subgroups based on gene expression profiles of bone marrow CD34$^+$ cells. (A) Rates of overall survival in the 2 subgroups for the training set of bone marrow CD34$^+$ cell samples. (B) Cumulative incidence of leukemic transformation in the 2 subgroups for the training set of bone marrow CD34$^+$ cell samples. The patients who had already developed leukemia at the time of sampling were removed from the analysis.

in the EMK subgroup (q-value < 0.05; Figure 5A). Enrichment of *SF3B1* mutations in the EMK subgroup was observed when the analysis was limited to the patients with bone marrow ring sideroblasts of ≥15% (n = 89) (odds ratio, 5.2; 95% CI, 1.9-14; $P = .003$). Notable exceptions for this were the *SF3B1*-mutated patients with *EVI1* overexpression, who were all grouped into the IMP subgroup, which is in line with a strong adverse effect of *EVI1* overexpression on survival (Figure 5A; supplemental Figure 8).[40,41] Effects of genetic alterations on transcriptome profiles were also suggested by higher frequencies of genetic lesions known to be associated with worse prognosis in the IMP subgroup (−7/del(7q) and *NRAS* and *RUNX1* mutations; q-value < 0.05) (Figure 5A).[9] In accordance with their higher rate of leukemic transformation, patients in the IMP subgroup were enriched for the type I mutations, including those in *FLT3*, *PTPN11*, *WT1*, *IDH1/IDH2*, *NPM1*, and *NRAS*, which were recently reported to be associated with transformation to secondary AML (supplemental Figure 10).[42]

Comparison between our gene expression–based classification and other prognostic models

In view of the enhanced progenitor signatures in the IMP subgroup, it would be of interest to correlate our classifier with the LSC17 score, which has recently been proposed to predict a subset of poor-risk AML based on the expression of 17 genes related to a leukemic stem cell signature.[43] None of the genes in the LSC17 score overlapped with our regression model. We first investigated the prognostic impact of the LSC17 score in our test cohort of patients, in which expression profiling was performed using BMMNCs (n = 114). As shown in supplemental Figure 11, patients with higher LSC17 scores (above the median) had a significantly shorter overall survival than those with lower scores (HR, 3.2; 95% CI, 1.4-7.1; $P = .003$). However, the LSC17 score was outperformed by our classification; the IMP subgroup was associated with a larger HR and a higher accuracy to predict clinical outcomes (Figure 4B). Conversely, when applied to the cohort of 179 AML patients from The Cancer Genome Atlas,[44] the LSC17 score was a better predictor of survival than our model (supplemental Figure 12). These results suggest that the IMP subgroup and LSC17-associated expression signatures seem to reflect different aspects of leukemia pathogenesis, including alterations in cell populations (myelodysplasia with low blast counts and full-blown leukemia with increased marrow blasts). Logistic regression odds ratio of being classified as the IMP subgroup (IMP score) dramatically increased during clonal evolution of myelodysplasia, from healthy controls to patients with myelodysplasia and then to those who had experienced leukemic transformation (Figure 5B). Most of the patients with de novo AML had increased IMP scores. Thus, the dramatic increase in the IMP score during leukemic

Table 2. HR for death or leukemic transformation in a multivariate model

	HR (95% CI)	*P*
Patients with CD34$^+$ cell samples (n = 100)		
Death		
Hemoglobin (g/dL)	0.54 (0.37-0.77)	<.001
Bone marrow blast (%)	1.05 (1.02-1.09)	.003
Age (y)	1.09 (1.02-1.16)	.009
IMP vs EMK subgroup	4.87 (1.25-19.0)	.02
Leukemic transformation*		
Patients with BMMNC samples only (n = 114)		
Death		
Bone marrow blast (%)	1.11 (1.06-1.16)	<.001
Hemoglobin (g/dL)	0.67 (0.52-0.86)	.001
IMP vs EMK subgroup	2.90 (1.10-7.62)	.03
Absolute neutrophil count (×10^9/L)	1.12 (1.00-1.26)	0.05
Cytogenetic abnormalities	2.16 (0.92-5.07)	0.08
Platelet (×10^9/L)	0.996 (0.992-1.001)	.12
Leukemic transformation		
Bone marrow blast (%)	1.14 (1.08-1.21)	<.001
IMP vs EMK subgroup	5.86 (1.59-21.6)	.008
Hemoglobin (g/dL)	0.80 (0.66-0.98)	.03

*Failure in convergence due to no leukemic transformation in the EMK subgroup.

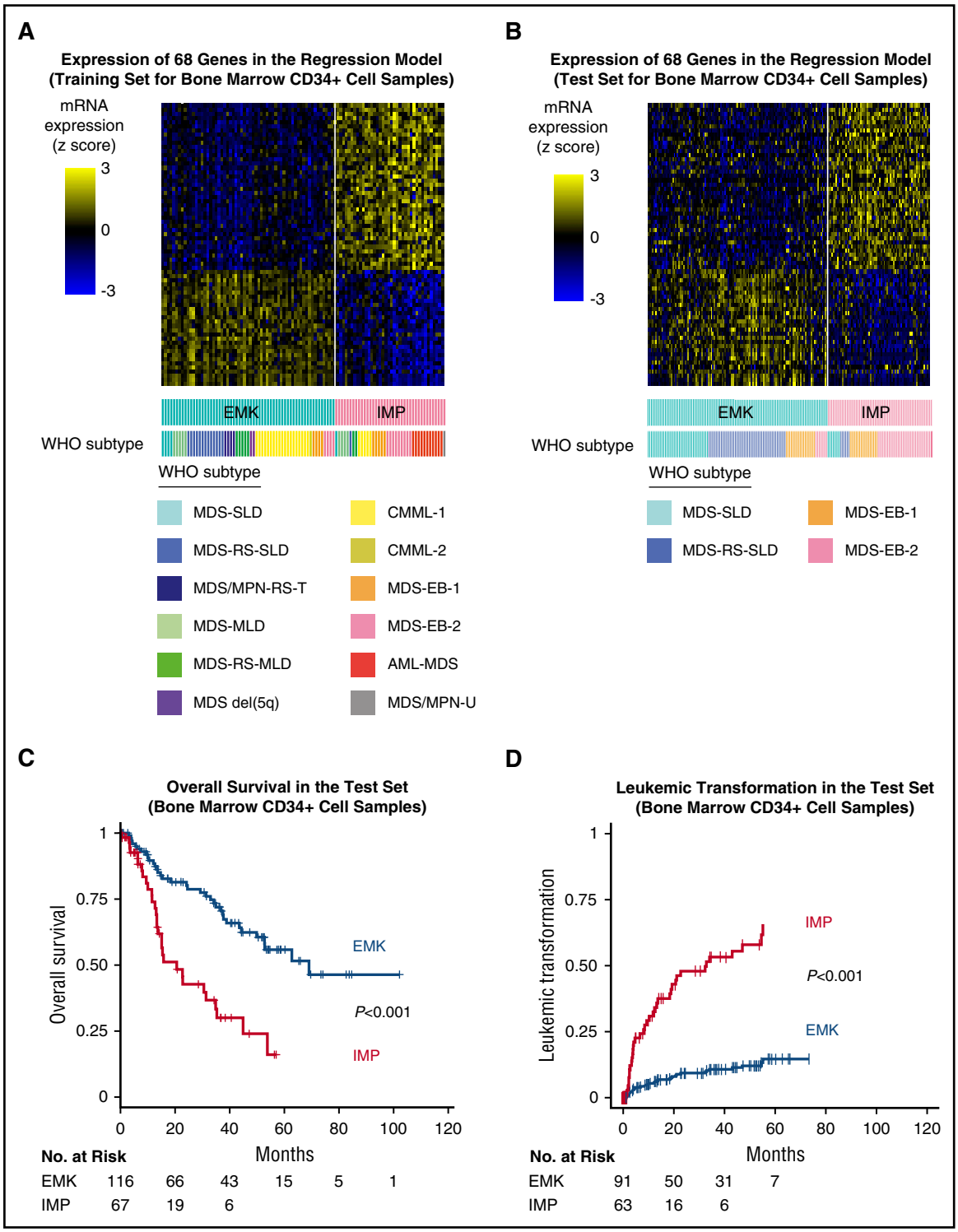


Figure 3. Validation of the prognostic significance of the subgroups in the test set for bone marrow CD34$^+$ cell samples. (A) Heatmap showing expression levels of 68 genes in the regression model for 100 patients with bone marrow CD34$^+$ cell samples. Each row represents 1 gene, and each column represents 1 sample. Gene expression–based subgroups, WHO subtypes, and patients' prognosis are shown below the heatmap. (B) A heatmap shows expression levels of 68 genes in the regression model for an independent cohort of 183 patients with MDS. (C) Rates of overall survival in the 2 subgroups for the test set of bone marrow CD34$^+$ cell samples. (D) Cumulative incidence of leukemic transformation in the 2 subgroups for the test cohort of bone marrow CD34$^+$ cell samples.

transformation might be the basis of a strong prognostic impact of the IMP subgroup in myelodysplasia. By contrast, the LSC17 score showed a modest, if not statistically significant, increase during clonal evolution of myelodysplasia (supplemental Figure 13), suggesting that unlike the IMP score, the LSC17 score is thought to more reflect an aggressive phenotype of AML than the pathogenesis of AML itself.

Finally, we assessed whether prediction of prognosis is improved by incorporating our gene expression–based classification into the IPSS-R. The analysis was confined to 148 patients with

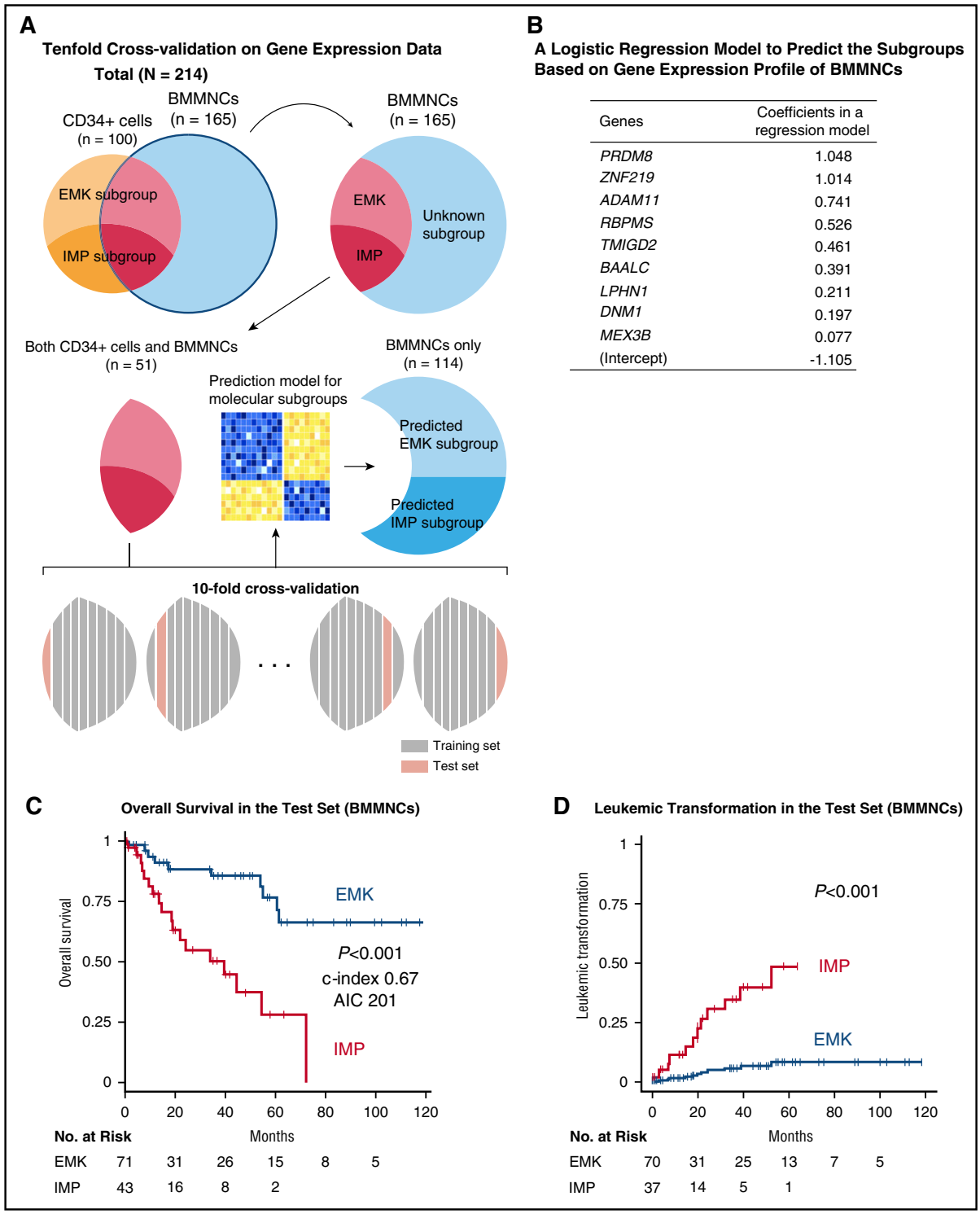
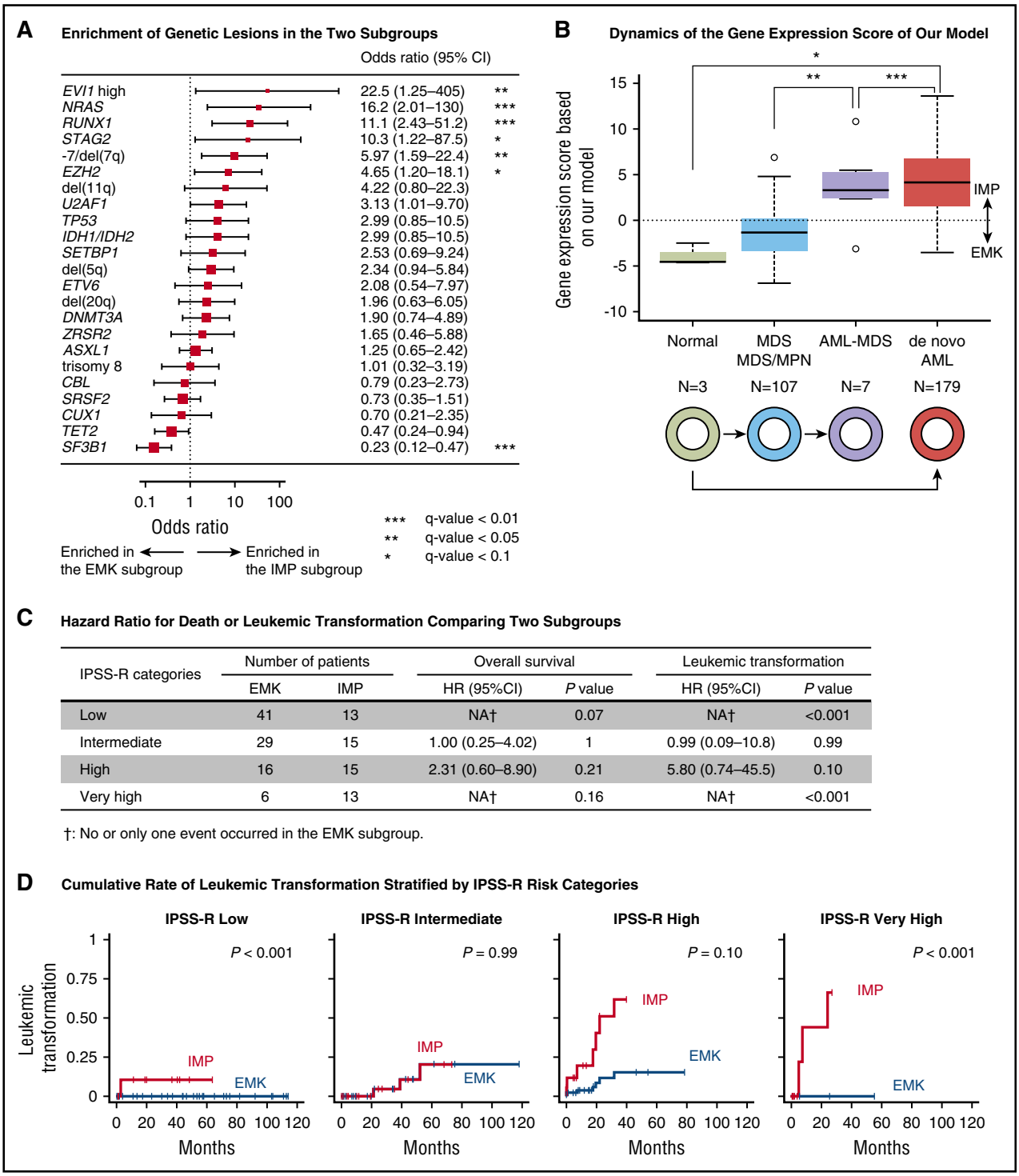


Figure 4. Prognostic significance of the subgroups based on gene expression profiles of BMMNCs. (A) A schematic depicts how to construct a logistic regression model to predict the subgroups using 10-fold cross-validation on gene expression data. (B) A table shows gene names and their coefficients in a logistic regression model to predict the subgroups based on gene expression profile of BMMNCs. (C) Rates of overall survival in the 2 subgroups for the test set of BMMNC samples. (D) Cumulative incidence of leukemic transformation in the 2 subgroups for the test set of BMMNC samples. The patients who had already developed leukemia at the time of sampling were removed from the analysis.

MDS and complete clinical data (Table 1). Although not all samples were obtained at diagnosis, no patients received treatment other than supportive care prior to sample collection. We thus tentatively defined IPSS-R categories based on clinical information available at the time of sampling. Clinical outcomes were first compared between the EMK and IMP subgroups stratified by IPSS-R categories. As compared with the patients in the EMK subgroup, those in the IMP subgroup had a significantly higher risk of leukemic transformation in the IPSS-R low- and very-high-risk groups (Figure 5C-D; supplemental Figure 14). We next made a prognostic



IPSS-R categories	Number of patients		Overall survival		Leukemic transformation	
	EMK	IMP	HR (95%CI)	*P* value	HR (95%CI)	*P* value
Low	41	13	NA†	0.07	NA†	<0.001
Intermediate	29	15	1.00 (0.25–4.02)	1	0.99 (0.09–10.8)	0.99
High	16	15	2.31 (0.60–8.90)	0.21	5.80 (0.74–45.5)	0.10
Very high	6	13	NA†	0.16	NA†	<0.001

†: No or only one event occurred in the EMK subgroup.

Figure 5. Comparison between our gene expression-based classification and other prognostic models. (A) A forest plot of odds ratios for detecting genetic lesions in the IMP subgroup. *P* values are based on Fisher's exact test. The square sizes were inversely proportional to the confidence intervals of the estimated odds ratio. (B) A boxplot shows distribution of log odds ratio for the predicted IMP subgroup that was calculated based on our regression model using expression levels of 9 genes. Boxes correspond to healthy adults (n = 3), patients with MDS or myelodysplastic/myeloproliferative neoplasm (MDS/MPN; n = 107), those with AML with myelodysplasia-related changes (AML-MDS; n = 7), and those in a published external AML cohort from The Cancer Genome Atlas (n = 179). *$P < .05$ (Bonferroni adjusted); **$P < .01$ (Bonferroni adjusted); ***$P < .001$ (Bonferroni adjusted). (C) HR of death or leukemic transformation comparing the EMK and IMP subgroups. Number of patients in the EMK and IMP subgroup, HR, its 95% CI, and *P* values are shown for each IPSS-R risk category. (D) Cumulative incidence of leukemic transformation in MDS patients stratified by categories of IPSS-R.

model that incorporated our gene expression–based classification into the IPSS-R. This combined model outperformed the IPSS-R only model in predicting overall survival and leukemic transformation (supplemental Figure 15). This indicated an additive effect of the gene expression–based classification in patient prognostication.

Discussion

Myelodysplasia is a highly heterogeneous group of myeloid neoplasms in terms of clinical, pathological, and mutational features, as separately classified into a number of distinct subtypes in the WHO classification system.[10,11,21,45] It is rather unexpected that patients were separated into 2 major subgroups with enhanced EMK and IMP signatures based on gene expression profiles. It should be underscored that unlike previous studies,[16-18] these 2 myelodysplasia subgroups were identified in a totally unbiased manner based solely on gene expression and without relying on the clinical outcome. Nonetheless, they had strong prognostic value, which was independent of clinical and demographic variables, including percentages of bone marrow blasts. Altered transcriptional programs may precede and predict overt morphological changes and clonal expansion. This may also be consistent with the simple distinction of myelodysplasia between 2 categories: clonal cytopenia before clonal evolution and oligoblastic myelogenous leukemias.[46] Importantly, these expression subgroups were largely reproduced in an independent external cohort of patients, even though a different platform was used for gene expression profiling, indicating the robustness of the above classification.

Each subgroup had distinct clinical, genomic, and transcriptomic profiles. The EMK subgroup was associated with the ring sideroblast phenotype, *SF3B1* mutation, and the strong erythroid signature. The IMP subgroup was characterized by lower platelet counts, increased marrow blasts, high-risk genomic lesions, and enhanced immature progenitor signatures. Features of the IMP subgroup as compared with the EMK subgroup provide insights into the molecular pathogenesis of disease progression in myelodysplasia. The deregulation of genes related to hematopoietic lineages suggests a severe differentiation block and/or stem cell proliferation in the IMP subgroup. The IMP subgroup also showed upregulation of various cell signaling pathways. Deregulated pathways included Notch, MAPK, phosphatidylinositol 3-kinase, and JAK-STAT signaling, of which activation is known to have a role in hematopoietic differentiation and self-renewal of stem cells.[47-50] The importance of cell signaling activation is also supported by the enrichment of *NRAS* mutation in the IMP subgroup and by the frequent acquisition of mutations in genes encoding signaling molecules, such as *NRAS*, *KRAS*, *PTPN11*, and *FLT3*, during disease progression of myelodysplasia.[42] In contrast, many pathways related to DNA repair and metabolism were downregulated in the IMP subgroup. Impaired DNA repair and damage response in high-risk MDS is consistent with a previous report[28] and can accelerate leukemia development. Levels of various metabolites and genes related to metabolism have been shown to change dramatically during differentiation from quiescent hematopoietic stem cells.[51,52] Global repression of genes related to metabolism in the IMP subgroup might reflect enrichment of metabolically quiescent stem/progenitor cells. Alternatively, rapid cell proliferation associated with ineffective hematopoiesis in low-risk MDS can be a basis for active metabolism in the EMK subgroup.[53,54] Deregulation of several metabolic pathways at the transcriptome level was previously reported in refractory anemia.[28] The expression of immunodeficiency, apoptosis, and chemokine signaling pathways, which were previously reported to be significantly deregulated in refractory anemia,[28] did not differ significantly between the EMK and IMP subgroups.

These gene expression subtypes were identified from the analysis of purified $CD34^+$ cells, but not whole BMMNCs, suggesting that their characteristic expression profiles represent intrinsic features of leukemic progenitors. Difficulty in identifying stable clusters from global gene expression profiles of unfractionated BMMNCs might be due to mature hematopoietic cells or contaminating peripheral blood cells in BMMNC samples that obscured characteristic gene expression of immature progenitors. The importance of relying on the gene expression of stem/progenitor fractions for better characterization is also underscored by the successful use of the leukemia stem cell signature to predict a prognostic model for AML.[43,55] Nevertheless, the 2 subgroups could be successfully predicted on the basis of the expression levels of a small number of genes (n = 9) selected from highly upregulated genes in BMMNC samples of the IMP subgroup. The transcriptomic differences that were consistently detected in more differentiated cell populations suggested the presence of intrinsic biological processes underlying phenotypic differences. The regression model obviates the need for purification of $CD34^+$ progenitor cells, further enhancing the clinical utility of our classification.

In conclusion, we demonstrated that myelodysplasia patients can be classified into 2 distinct clusters, the EMK and IMP subgroups, which have unique genomic, transcriptomic, and clinical features. Our results support the integration of gene expression data into prognostic models of myelodysplasia. Further studies are warranted to assess the relevance of our gene expression–based classification for newly diagnosed myelodysplasia in a prospective setting.

Acknowledgments

The authors thank all patients for their participation in this study and Hiroshi Kobayashi for fruitful discussions. They deeply appreciate Elsa Bernard, Elli Papaemmanuil, and Yasuhito Nannya for their help with statistics and computation.

This work was supported by the Project for Development of Innovative Research on Cancer Therapeutics (P-DIRECT, 16cm0106501h0001), Practical Research for Innovative Cancer Control (15Ack0106014h0002, 16ck0106073h0003), and Project for Cancer Research and Therapeutic Evolution (P-CREATE, 16cm0106501h0001) from the Japan Agency for Medical Research and Development (S.O.). This work was also supported by the Japan Society for the Promotion of Science KAKENHI (15J02911 [Y. Shiozawa]; 15H05668 [K.Y.]; 22134006, 26221308, and 15H05909 [S.O.]) and by research grants from Takeda Science Foundation (S.O.), Associazione Italiana per la Ricerca sul Cancro (Special Program Molecular Clinical Oncology 5 per Mille, project 1005 [M.C.] and investigator grant 15356 [L.M.]), Fondazione Regionale Ricerca Biomedica (project 2015-0042 [M.C.]), Bloodwise, UK (A.P. and J.B.), the Swedish Cancer Society, the Research Council of Sweden (E.H.-L.), and the Uehara Memorial Foundation (K.Y.). This research used computational resources of the Human Genome Center, the Institute of Medical Science, the University of Tokyo, Japan, and the K computer provided by the RIKEN Advanced Institute for Computational Science through the High Performance Computing Infrastructure System Research project (hp160219).

Authorship

Contribution: L.M., A.G., A.P., M.K., J.B., E.H.-L., and M.C. collected and provided the specimens; Y. Shiozawa, Y. Sato, and K.Y. performed sequencing experiments; Y. Shiozawa, H.S., T.Y., Y. Shiraishi, K.C., and S.M. developed bioinformatics pipelines; Y. Shiozawa and A.S.-O. performed sequencing data analyses; A.P. and J.B. performed microarray data analyses; Y. Shiozawa, L.M., and H.M. performed clinical analysis; Y. Shiozawa, L.M., S.O., and

M.C. generated the figures and tables and wrote the manuscript; M.C. and S.O. co-led the entire project; and all authors participated in discussions and interpretation of the data and results.

Conflict-of-interest disclosure: The authors declare no competing financial interests.

ORCID profiles: Y. Shiozawa, 0000-0001-9814-9230; L.M., 0000-0002-1460-1611; A.G., 0000-0002-7912-6428; A.P., 0000-0002-6122-0221; M.K., 0000-0002-8946-6944; Y. Sato, 0000-0003-0913-7815; H.S., 0000-0003-1434-7104; K.Y., 0000-0001-5189-5105; Y. Shiraishi, 0000-0001-6144-5845; J.B., 0000-0002-4330-2928; E.H.-L., 0000-0002-0602-3815; S.M., 0000-0002-1753-6616; M.C., 0000-0001-6984-8817; S.O., 0000-0002-7778-5374.

Correspondence: Mario Cazzola, Department of Molecular Medicine, University of Pavia, Pavia 27100, Italy; e-mail: mario.cazzola@unipv.it; and Seishi Ogawa, Department of Pathology and Tumor Biology, Kyoto University, Kyoto 606-8501, Japan; e-mail: sogawa-tky@umin.ac.jp.

References

1. Cazzola M, Della Porta MG, Malcovati L. The genetic basis of myelodysplasia and its clinical relevance. *Blood*. 2013;122(25):4021-4034.
2. Lindsley RC, Ebert BL. Molecular pathophysiology of myelodysplastic syndromes. *Annu Rev Pathol*. 2013;8(1):21-47.
3. Malcovati L, Porta MG, Pascutto C, et al. Prognostic factors and life expectancy in myelodysplastic syndromes classified according to WHO criteria: a basis for clinical decision making. *J Clin Oncol*. 2005;23(30):7594-7603.
4. Greenberg P, Cox C, LeBeau MM, et al. International scoring system for evaluating prognosis in myelodysplastic syndromes. *Blood*. 1997;89(6):2079-2088.
5. Greenberg PL, Tuechler H, Schanz J, et al. Revised international prognostic scoring system for myelodysplastic syndromes. *Blood*. 2012; 120(12):2454-2465.
6. Malcovati L, Germing U, Kuendgen A, et al. Time-dependent prognostic scoring system for predicting survival and leukemic evolution in myelodysplastic syndromes. *J Clin Oncol*. 2007; 25(23):3503-3510.
7. Kantarjian H, O'Brien S, Ravandi F, et al. Proposal for a new risk model in myelodysplastic syndrome that accounts for events not considered in the original International Prognostic Scoring System. *Cancer*. 2008;113(6):1351-1361.
8. Garcia-Manero G, Shan J, Faderl S, et al. A prognostic score for patients with lower risk myelodysplastic syndrome. *Leukemia*. 2008; 22(3):538-543.
9. Bejar R, Stevenson K, Abdel-Wahab O, et al. Clinical effect of point mutations in myelodysplastic syndromes. *N Engl J Med*. 2011; 364(26):2496-2506.
10. Papaemmanuil E, Gerstung M, Malcovati L, et al; Chronic Myeloid Disorders Working Group of the International Cancer Genome Consortium. Clinical and biological implications of driver mutations in myelodysplastic syndromes. *Blood*. 2013;122(22):3616-3627, quiz 3699.
11. Haferlach T, Nagata Y, Grossmann V, et al. Landscape of genetic lesions in 944 patients with myelodysplastic syndromes. *Leukemia*. 2014; 28(2):241-247.
12. Den Boer ML, van Slegtenhorst M, De Menezes RX, et al. A subtype of childhood acute lymphoblastic leukaemia with poor treatment outcome: a genome-wide classification study. *Lancet Oncol*. 2009;10(2):125-134.
13. Alizadeh AA, Eisen MB, Davis RE, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*. 2000; 403(6769):503-511.
14. Perou CM, Sørlie T, Eisen MB, et al. Molecular portraits of human breast tumours. *Nature*. 2000; 406(6797):747-752.
15. Sørlie T, Perou CM, Tibshirani R, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci USA*. 2001; 98(19):10869-10874.
16. Sridhar K, Ross DT, Tibshirani R, Butte AJ, Greenberg PL. Relationship of differential gene expression profiles in CD34+ myelodysplastic syndrome marrow cells to disease subtype and progression. *Blood*. 2009;114(23):4847-4858.
17. Pellagatti A, Benner A, Mills KI, et al. Identification of gene expression-based prognostic markers in the hematopoietic stem cells of patients with myelodysplastic syndromes. *J Clin Oncol*. 2013; 31(28):3557-3564.
18. Gerstung M, Pellagatti A, Malcovati L, et al. Combining gene mutation with gene expression data improves outcome prediction in myelodysplastic syndromes. *Nat Commun*. 2015; 6:5901.
19. Mills KI, Kohlmann A, Williams PM, et al. Microarray-based classifiers and prognosis models identify subgroups with distinct clinical outcomes and high risk of AML transformation of myelodysplastic syndrome. *Blood*. 2009;114(5): 1063-1072.
20. Malcovati L, Della Porta MG, Pietra D, et al. Molecular and clinical features of refractory anemia with ringed sideroblasts associated with marked thrombocytosis. *Blood*. 2009;114(17): 3538-3545.
21. Arber DA, Orazi A, Hasserjian R, et al. The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia. *Blood*. 2016;127(20):2391-2405.
22. Grant GR, Farkas MH, Pizarro AD, et al. Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM). *Bioinformatics*. 2011;27(18):2518-2528.
23. Sato Y, Yoshizato T, Shiraishi Y, et al. Integrated molecular analysis of clear-cell renal cell carcinoma. *Nat Genet*. 2013;45(8):860-867.
24. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*. 2010;26(1):139-140.
25. Wu D, Lim E, Vaillant F, Asselin-Labat ML, Visvader JE, Smyth GK. ROAST: rotation gene set tests for complex microarray experiments. *Bioinformatics*. 2010;26(17):2176-2182.
26. Chen L, Kostadima M, Martens JHA, et al. Transcriptional diversity during lineage commitment of human blood progenitors. *Science*. 2014;345(6204):1251033.
27. Verhaak RG, Hoadley KA, Purdom E, et al; Cancer Genome Atlas Research Network. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell*. 2010;17(1):98-110.
28. Pellagatti A, Cazzola M, Giagounidis A, et al. Deregulated gene expression pathways in myelodysplastic syndrome hematopoietic stem cells. *Leukemia*. 2010;24(4):756-764.
29. Yoshida K, Sanada M, Shiraishi Y, et al. Frequent pathway mutations of splicing machinery in myelodysplasia. *Nature*. 2011;478(7367):64-69.
30. Wu ZJ, Irizarry RA, Gentleman R, Martinez-Murillo F, Spencer F. A model-based background adjustment for oligonucleotide expression arrays. *J Am Stat Assoc*. 2004;99(468):909-917.
31. Li Q, Birkbak NJ, Gyorffy B, Szallasi Z, Eklund AC. Jetset: selecting the optimal microarray probe set to represent a gene. *BMC Bioinformatics*. 2011;12(1):474.
32. Radujkovic A, Dietrich S, Andrulis M, et al. Expression of CDKN1C in the bone marrow of patients with myelodysplastic syndrome and secondary acute myeloid leukemia is associated with poor survival after conventional chemotherapy. *Int J Cancer*. 2016;139(6): 1402-1413.
33. Fine JP, Gray RJ. A proportional hazards model for the subdistribution of a competing risk. *J Am Stat Assoc*. 1999;94(446):496-509.
34. Ozeki K, Kiyoi H, Hirose Y, et al. Biologic and clinical significance of the FLT3 transcript level in acute myeloid leukemia. *Blood*. 2004;103(5): 1901-1908.
35. Mohty M, Yong AS, Szydlo RM, Apperley JF, Melo JV. The polycomb group BMI1 gene is a molecular marker for predicting prognosis of chronic myeloid leukemia. *Blood*. 2007;110(1): 380-383.
36. Langer C, Marcucci G, Holland KB, et al. Prognostic importance of MN1 transcript levels, and biologic insights from MN1-associated gene and microRNA expression signatures in cytogenetically normal acute myeloid leukemia: a cancer and leukemia group B study. *J Clin Oncol*. 2009;27(19):3198-3204.
37. Schwind S, Marcucci G, Maharry K, et al. BAALC and ERG expression levels are associated with outcome and distinct gene and microRNA expression profiles in older patients with de novo cytogenetically normal acute myeloid leukemia: a Cancer and Leukemia Group B study. *Blood*. 2010;116(25):5660-5669.
38. Cristóbal I, Blanco FJ, Garcia-Orti L, et al. SETBP1 overexpression is a novel leukemogenic mechanism that predicts adverse outcome in elderly patients with acute myeloid leukemia. *Blood*. 2010;115(3):615-625.
39. Byers RJ, Currie T, Tholouli E, Rodig SJ, Kutok JL. MSI2 protein expression predicts unfavorable outcome in acute myeloid leukemia. *Blood*. 2011; 118(10):2857-2867.
40. Thol F, Yun H, Sonntag AK, et al. Prognostic significance of combined MN1, ERG, BAALC, and EVI1 (MEBE) expression in patients with myelodysplastic syndromes. *Ann Hematol*. 2012; 91(8):1221-1233.
41. Barjesteh van Waalwijk van Doorn-Khosrovani S, Erpelinck C, van Putten WL, et al. High EVI1 expression predicts poor survival in acute myeloid leukemia: a study of 319 de novo AML patients. *Blood*. 2003;101(3):837-845.

42. Makishima H, Yoshizato T, Yoshida K, et al. Dynamics of clonal evolution in myelodysplastic syndromes. *Nat Genet*. 2017;49(2):204-212.

43. Ng SW, Mitchell A, Kennedy JA, et al. A 17-gene stemness score for rapid determination of risk in acute leukaemia. *Nature*. 2016;540(7633): 433-437.

44. Ley TJ, Miller C, Ding L, et al; Cancer Genome Atlas Research Network. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N Engl J Med*. 2013;368(22): 2059-2074.

45. Walter MJ, Shen D, Shao J, et al. Clonal diversity of recurrently mutated genes in myelodysplastic syndromes. *Leukemia*. 2013;27(6):1275-1282.

46. Lichtman MA. Does a diagnosis of myelogenous leukemia require 20% marrow myeloblasts, and does <5% marrow myeloblasts represent a remission? The history and ambiguity of arbitrary diagnostic boundaries in the understanding of myelodysplasia. *Oncologist*. 2013;18(9):973-980.

47. Rossi L, Lin KK, Boles NC, et al. Less is more: unveiling the functional core of hematopoietic stem cells through knockout mice. *Cell Stem Cell*. 2012;11(3):302-317.

48. McCubrey JA, Steelman LS, Abrams SL, et al. Targeting survival cascades induced by activation of Ras/Raf/MEK/ERK, PI3K/PTEN/Akt/mTOR and Jak/STAT pathways for effective leukemia therapy. *Leukemia*. 2008;22(4): 708-722.

49. Rizo A, Vellenga E, de Haan G, Schuringa JJ. Signaling pathways in self-renewing hematopoietic and leukemic stem cells: do all stem cells need a niche? *Hum Mol Genet*. 2006; 15(Spec No 2 suppl_2):R210-R219.

50. Dreesen O, Brivanlou AH. Signaling pathways in cancer and embryonic stem cells. *Stem Cell Rev*. 2007;3(1):7-17.

51. Cabezas-Wallscheid N, Klimmeck D, Hansson J, et al. Identification of regulatory networks in HSCs and their immediate progeny via integrated proteome, transcriptome, and DNA methylome analysis. *Cell Stem Cell*. 2014;15(4):507-522.

52. Agathocleous M, Meacham CE, Burgess RJ, et al. Ascorbate regulates haematopoietic stem cell function and leukaemogenesis. *Nature*. 2017; 549(7673):476-481.

53. Raza A, Gezer S, Mundle S, et al. Apoptosis in bone marrow biopsy samples involving stromal and hematopoietic cells in 50 patients with myelodysplastic syndromes. *Blood*. 1995;86(1): 268-276.

54. Raza A, Mundle S, Iftikhar A, et al. Simultaneous assessment of cell kinetics and programmed cell death in bone marrow biopsies of myelodysplastics reveals extensive apoptosis as the probable basis for ineffective hematopoiesis. *Am J Hematol*. 1995;48(3): 143-154.

55. Eppert K, Takenaka K, Lechman ER, et al. Stem cell gene expression programs influence clinical outcome in human leukemia. *Nat Med*. 2011; 17(9):1086-1093.