Microarray-based classifiers and prognosis models identify subgroups with distinct clinical outcomes and high risk of AML transformation of myelodysplastic syndrome

Ken I. Mills,[1-3] Alexander Kohlmann,[4,5] P. Mickey Williams,[4] Lothar Wieczorek,[4] Wei-min Liu,[4] Rachel Li,[4] Wen Wei,[4] David T. Bowen,[6] Helmut Loeffler,[7] Jesus M. Hernandez,[3,8] Wolf-Karsten Hofmann,[3,9,10] and Torsten Haferlach[3,5]

[1]Centre for Cancer Research and Cell Biology, Queen's University Belfast, Belfast, United Kingdom; [2]Department of Haematology, Cardiff University, Cardiff, United Kingdom; [3]MILE Study (WP13), on behalf of the European LeukemiaNet, Mannheim, Germany; [4]Genomics and Oncology, Roche Molecular Systems Inc, Pleasanton, CA; [5]MLL, Munich Leukemia Laboratory, Munich, Germany; [6]Department of Haematology, St James's Institute of Oncology, Leeds, United Kingdom; [7]St Peter, Germany; [8]Servicio de Hematologia, Hospital Universitario de Salamanca and Instituto de Biologia Molecular y Celular del Cancer, Centro de Investigación del Cáncer, Universidad de Salamanca-CSIC, Salamanca, Spain; [9]Charité, University Hospital Benjamin Franklin, Berlin, Germany; and [10]Medizinischen Klinik III, Hämatologie und Onkologie, Universitätsmedizin Mannheim, Mannheim, Germany

The diagnosis of myelodysplastic syndrome (MDS) currently relies primarily on the morphologic assessment of the patient's bone marrow and peripheral blood cells. Moreover, prognostic scoring systems rely on observer-dependent assessments of blast percentage and dysplasia. Gene expression profiling could enhance current diagnostic and prognostic systems by providing a set of standardized, objective gene signatures. Within the Microarray Innovations in LEukemia study, a diagnostic classification model was investigated to distinguish the distinct subclasses of pediatric and adult leukemia, as well as MDS. Overall, the accuracy of the diagnostic classification model for subtyping leukemia was approximately 93%, but this was not reflected for the MDS samples giving only approximately 50% accuracy. Discordant samples of MDS were classified either into acute myeloid leukemia (AML) or "none-of-the-targets" (neither leukemia nor MDS) categories. To clarify the discordant results, all submitted 174 MDS samples were externally reviewed, although this did not improve the molecular classification results. However, a significant correlation was noted between the AML and "none-of-the-targets" categories and prognosis, leading to a prognostic classification model to predict for time-dependent probability of leukemic transformation. The prognostic classification model accurately discriminated patients with a rapid transformation to AML within 18 months from those with more indolent disease. (Blood. 2009;114:1063-1072)

Introduction

The myelodysplastic syndromes (MDSs) are clonal hematopoietic disorders that are characterized by ineffective hematopoiesis and a variable propensity to evolve to acute myeloid leukemia (AML).[1] Standard diagnostic criteria for MDS and its various subtypes, using either the older French-American-British classification[2] or the more recent World Health Organization (WHO) classification,[3-6] rely heavily on the subjective morphologic evaluation of bone marrow cells.

Given the variable course of individual cases of MDS, several prognostic scoring systems have been proposed to predict survival and probability of leukemic evolution. The 2 most widely used systems, the International Prognostic Scoring System (IPSS)[7] and the WHO classification-based prognostic scoring system,[8] have both been shown to have prognostic value. However, both systems include observer-dependent criteria,[9] such as blast percentage, degree of lineage dysplasia, and presence of ringed sideroblasts. The diagnostic and prognostic challenges in MDS are compounded by the technical complexity of cytogenetic analysis. Thus, a standardized objective molecularly based classification strategy, such as gene expression profiling (GEP), could provide an improved method for diagnosis and prognostication in this group of disorders. From a technical perspective, intraplatform consistency across multiple laboratories as well as a high level of interplatform concordance in terms of genes identified as differentially expressed by microarrays have been demonstrated.[10,11]

4The first report of a microarray-based GEP schema for hematologic malignancies used an unsupervised, class discovery approach to uncover the molecular distinctions between AML and acute lymphoblastic leukemia,[12] and demonstrated that a GEP strategy could accurately subdivide acute leukemias.[13-15] Subsequently, several microarray-based GEP studies of MDS, mostly using purified $CD34^+$ or $AC133^+$ cell populations, have been published. Several studies compared gene signatures between MDS and the healthy persons, between different risk groups of MDS, or between MDS-derived AML and de novo AML.[16-19] Although these studies have provided important molecular insights into the pathophysiology of MDS, they were not designed to test the diagnostic or prognostic capabilities of GEP in this group of diseases.

The international multi-institutional Microarray Innovations in LEukemia (MILE) research program, centered on the European LeukemiaNet (ELN, www.leukemia-net.org), assessed the clinical utility of a microarray-based GEP assay in the diagnosis and subclassification of 16 clinically recognized subtypes of acute and chronic leukemia.

Submitted October 31, 2008; accepted May 14, 2009. Prepublished online as *Blood* First Edition paper, May 14, 2009; DOI 10.1182/blood-2008-10-187203.

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.

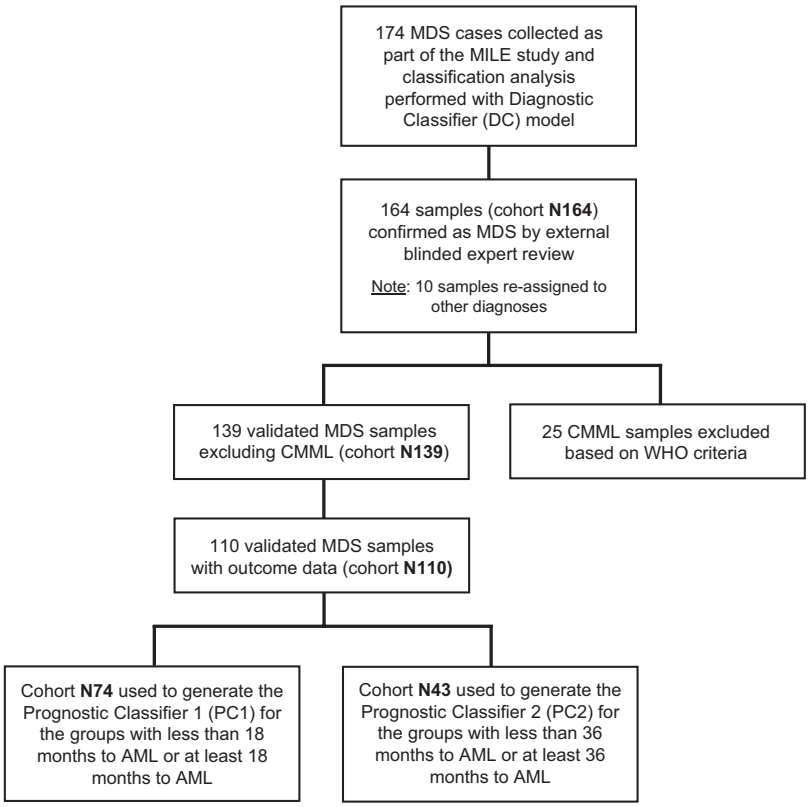

Figure 1. Flow chart showing the relationship of datasets used in the study. The chart explains the selection of patient cohorts and filtering processes in the DC model analysis and the development of the PC model risk scores.

A diagnostic classification (DC) model, developed for and evaluated during the MILE study, was also designed to distinguish leukemia from MDS and from nonleukemic conditions.[20]

Although the DC model proved to be very accurate in the classification of leukemia, it failed to confirm the clinical diagnosis of MDS in half of the MDS specimens submitted to the study.[21] In the discordant cases, the DC model classification was either AML or a nonleukemic condition, referred to as "none-of-the-targets." However, a blinded external pathologic review confirmed the initial clinical diagnosis of MDS in 94% of cases. We observed that cases of MDS classified as AML by the DC model had more aggressive disease and more rapid progression to AML, whereas MDS cases classified as "none-of-the-targets" had a more indolent clinical course. Based on this observation, we developed an improved prognostic classification (PC) model. The essence of the PC model was to provide a score related to transformation to AML and overall survival for MDS patients, which was based on the microarray data and clinical observations of time to AML of the training dataset.

Methods

Patient samples

The MILE study was approved by the relevant ethical committees in each country, and each patient sample was taken at de novo presentation of the disease with ethical informed consent for research purposes in each center in accordance with the Declaration of Helsinki. Bone marrow mononuclear cells were separated by Ficoll-Hypaque technique at each center, and total RNA was extracted according to the study protocol.

There are several patient datasets referred to in this manuscript (Figure 1). The original cohort of MDS samples submitted to the stage I of the MILE study consisted of 174 patients. After external review, 10 samples were excluded, leaving a cohort of 164 samples (dataset N164). Further subsets of N164 were defined as follows: (1) dataset N139 included the validated MDS specimens but excluded the chronic myelomonocytic leukemia (CMML) cases; (2) dataset N110 included patients in N139 for which data on survival and time to AML transformation were available; (3) dataset N74 was a subset of N110 and included patients with less than 18 months to AML or at least 18 months to AML; and (4) dataset N43 was a subset of N74 and included patients with less than 36 months to AML or at least 36 months to AML. Clinical characteristics of the N164, N139, or N110 MDS patient groups are shown in Table 1. Further annotation of the samples listing individual IPSS parameters, such as blast score, cytogenetic risk categories, and cytopenias, is available online.

RNA extraction

The methods used for RNA isolation, cRNA preparation and labeling, microarray analysis, quality control, and normalization of microarray data were as previously described.[22]

For each specimen (ie, Ficoll-banded bone marrow mononuclear cells), total RNA was converted into double-stranded cDNA by reverse transcription using a cDNA Synthesis System kit, including an oligo$(dT)_{24}$-T7 primer (Roche Applied Science) and the Poly-A control transcripts (Affymetrix). The generated cDNA was purified using the GeneChip Sample Cleanup Module (Affymetrix). Labeled cRNA was generated using the Microarray RNA target synthesis kit (Roche Applied Science) and in vitro transcription labeling nucleotide mix (Affymetrix). The generated cRNA was purified using the GeneChip Sample Cleanup Module (Affymetrix) and quantified using the NanoDrop ND-1000 spectrophotometer (NanoDrop Technologies). In each preparation, an amount of 11.0 μg cRNA was fragmented with 5× Fragmentation Buffer (Affymetrix) in a final reaction volume of 25 μL. The incubation steps during cDNA synthesis, in vitro transcription reaction, and target fragmentation were performed using the Hybex Microarray Incubation System (SciGene) and Eppendorf ThermoStat plus instruments (Eppendorf). Hybridization, washing, staining, and scanning protocols, respectively, were performed on Affymetrix GeneChip instruments (Hybridization Oven 640, Fluidics Station FS450, Scanner GCS3000) as

Table 1. Clinical characteristics of MDS data

	N164, n (%)	N139, n (%)	N110, n (%)
Age at diagnosis, y			
Younger than 50	8 (5)	7 (5)	6 (5)
51-60	21 (13)	19 (14)	17 (15)
61-70	30 (18)	25 (18)	19 (17)
71-80	50 (30)	43 (31)	35 (32)
Older than 81	52 (32)	42 (30)	33 (30)
Unknown	3 (2)	3 (2)	0 (0)
Sex			
Male	66 (40)	55 (40)	47 (43)
Female	98 (60)	84 (60)	63 (57)
Disease classification			
5q− syndrome	11 (7)	11 (8)	10 (9)
CMML	25 (15)		
RA	39 (24)	39 (28)	35 (32)
RAEB1	26 (16)	26 (19)	19 (17)
RAEB2	22 (13)	22 (16)	19 (17)
RARS	29 (18)	29 (21)	20 (18)
RCMD	12 (7)	12 (9)	7 (6)
IPSS scores			
Low	72 (44)	59 (42)	53 (48)
Intermediate-1	62 (38)	51 (37)	33 (30)
Intermediate-2	25 (15)	24 (17)	19 (17)
High	5 (3)	5 (4)	5 (5)
DC model classification			
AML	37 (23)	31 (22)	25 (23)
MDS	82 (50)	70 (50)	55 (50)
"None-of-the-targets"	39 (24)	34 (24)	27 (25)
Tied between class calls	6 (4)	4 (3)	3 (3)
Blast cell count			
Less than 5%	114 (70)	92 (66)	72 (65)
5%-10%	29 (18)	26 (19)	20 (18)
11%-20%	21 (13)	21 (15)	18 (16)
Cytogenetics			
Normal	117 (71)	97 (70)	78 (71)
Good (1 of: del(5q), del(20q), −Y)	18 (11)	18 (13)	15 (14)
Intermediate (any abnormalities)	19 (12)	14 (10)	12 (11)
Poor (complex: > 3 abnormalities)	8 (5)	8 (6)	4 (4)
Poor (chromosome 7 abnormalities)	2 (1)	2 (1)	1 (1)
Cytopenia			
0 or 1 cytopenias	98 (60)	79 (57)	65 (59)
2 or 3 cytopenias	66 (40)	60 (43)	45 (41)

recommended by the manufacturer. This procedure was well documented, and all laboratories were specifically trained in precise applications of this procedure and were required to demonstrate proficiency before commencement of this study.[22]

Image analysis and data processing

Microarray image files (DAT files) and cell intensity files (CEL files) were generated using default Affymetrix microarray analysis parameters (GCOS 1.2 software). The data preprocessing included generating probe set level signals, DS, or DQN1 algorithms, as described elsewhere.[23] Data visualization and exploratory analysis, such as box plots, principal component analysis (PCA), and hierarchical clustering, were performed with R software (http://www.R-project.org) and Partek Genomics Suite (http://www.partek.com).[24] Pathway analysis was done using the Ingenuity Pathway Analysis software (www.ingenuity.com). All microarray raw data (CEL files) and probe set signals are available at the National Center for Biotechnology Information Gene Expression Omnibus database (GEO, http://www.ncbi.nlm.nih.gov/geo/),[25] series accession number GSE15061.

Leukemia diagnostic classification model

The DC model is based on all-pairwise linear classifiers for 18 distinct classes. This model was intended to be evaluated in the MILE study and discriminated between 16 leukemia classes, MDS, and a "none-of-the-targets" control group (supplemental Table 1, available on the *Blood* website; see the Supplemental Materials link at the top of the online article). It consists of $18 \times (18-1)/2 = 153$ binary classifiers for all class pairs. For every class pair, a linear binary classifier was a support vector machine[26,27] using DQN signals.[23] The final call is based on the majority vote of the binary calls. The DC model used 534 probe sets on the HG-U133 Plus 2.0 microarray and a training dataset consisting of 1627 clinical specimens and 5 nonleukemia cell lines. A total of 1094 patients were analyzed using microarray pairs of HG-U133A and HG-U133B, and the remaining 538 GEPs were generated based on the HG-U133 Plus 2.0 microarray profiles (Affymetrix). A majority of specimens have been described elsewhere.[14] Detailed information on the 534 probe sets is available online in supplemental data.

Gene selection based on Cox proportional hazards for a prognostic classifier

In the N110 dataset, there were 19 cases with time to AML transformation after diagnosis of MDS and 91 cases with censored time to AML, including death from other causes. There were 55 observed deaths with time after diagnosis of MDS and 55 cases with censored survival time. The PC model was applied to every probe set to calculate the *P* value and recorded the top-200 probe sets with the smallest *P* values. The gene selection was not only done for the whole dataset N110 but also done in the leave-one-out

(LOO) cross-validation manner (ie, by generation of 110 sets of top-200 probe sets selected for all possible subsets containing 109 specimen; LOO of N110).

Risk scores based on classification of groups of time to AML

The 1-nearest neighbor classification was used to build risk scores for a prognostic model with 2 prognostic classifiers (PC1 and PC2). Two sub-datasets of N110 were used. The dataset N74 was used to generate the classifier for the groups with less than 18 months to AML or at least 18 months to AML. The dataset N43 was used to generate a second classifier for the groups with less than 36 months to AML or at least 36 months to AML. The top-30 probe sets were used for classification of the dataset N74, and the top-70 probe sets were used for classification of the dataset N43. Detailed information on the PC probe sets is available online (supplemental data). The accuracy of LOO cross-validation was 64 of 74 = 86% for classification with the 18-month cut-off, and that for classification with the 36-month cut-off was 35 of 43 = 81%. The process of selecting probe sets was done separately for each reiteration of cross-validation. The resubstitution of the training data back into the risk score resulted in a highly significant *P* value of the log-rank test; but to avoid an overestimate of the significance, the LOO cross-validation was used for an estimate of accuracy for the risk score.

The risk score is defined by the classification results. For a given GEP, we first applied the 1-nearest neighbor classifier of N74 in the Euclidean space of 30 probe sets to determine whether the subject will transfer to AML in less than 18 months. If yes, the risk score was 2. Otherwise, we applied the 1-nearest neighbor classifier of N43 in the Euclidean space of 70 probe sets to determine whether the subject will transform to AML in less than 36 months. If yes, the risk score was 1; otherwise, the risk score was 0. For the LOO approach, the sizes of the training datasets were 73 and 42, instead of 74 and 43, respectively. The significance of differences in time of transformation to AML and overall survival was assessed by the log-rank test for the Kaplan-Meier plots. Various covariates were compared with the hazard ratios. The multigene classification models rely on all genes in the model, with each gene weighted for its contribution to the model's output based on training data.

Results

MDS disease classification using the DC model

Within the MILE study, 174 GEPs from samples obtained at diagnosis were originally included with a clinical diagnosis of MDS. These specimens were submitted and processed in Berlin, Cardiff, Munich, or Salamanca. The diagnosis of the samples was assessed by specialists at each submission site using their individual expertise and standard diagnostic procedures. The age distribution of patients was representative of a typical nonselected MDS population, as was the distribution of MDS subtypes with a majority of samples from "low-risk" patients and cases with a normal karyotype (supplemental data).

Analysis of these 174 specimens using the DC model resulted in only 49% of them being correctly called as MDS[20] from their underlying GEPs. The remainder of the submitted MDS specimens was evenly split between a call into "none-of-the-targets" (24%) and AML (25%) categories, with a further 2% reporting a tie between classes (ie, being considered as samples with low signature confidence). There was no correlation between the classification call and the site of analysis of these specimens.

Validation of MDS samples

To confirm the clinical diagnosis of the submitted MDS specimens, bone marrow smears were sent for blinded review to 2 external experts from different institutions, who confirmed the diagnosis of MDS in 164 (94%) of the specimens. Ten discordant specimens were removed from the subsequent analysis as a result of this review process: 6 cases were reclassified as AML, which interestingly was consistent with the original class call from the DC model; and 4 cases were excluded from this MDS study: one CML, one CLL, one myeloma, and an incomplete slide set that did not permit external review. There was 82% concordance between the submitted and the external MDS subtype assignment of the final 164 samples. Four cases were submitted with a generic diagnosis of "MDS" and the validated diagnoses of these were that one had refractory anemia (RA), 2 had refractory anemia with excess blasts 1 (RAEB1), and one was RAEB2. A total of 84% (n = 21) of submitted CMML cases were confirmed; the remaining 4 patients (16%) were reclassified as RA, RAEB1, RAEB1, and refractory cytopenia with multilineage dysplasia (RCMD). Approximately one-third (17 of 52) of the submitted RA samples were reclassified to either RAEB1 (3 cases), refractory anemia with ring sideroblasts (RARS) (4 cases), RCMD (6 cases), or 5q− syndrome (4 cases). Only one of the submitted RAEB1 or RAEB2 cases (38 in total) was not validated, and in this case it was reclassified from RAEB1 to RAEB2.

In the final cohort of 164 confirmed MDS specimens, 50% were called MDS by the DC model, 24% were called "none-of-the-targets," 23% were called as AML, and the remaining 4% had mixed calls, indicating low classification confidence (ie, a classification tie of 2 or 3 classes). The clinical characteristics of the final 164 submitted and reviewed specimens are shown in Table 1. The 2001 WHO classification of the myeloid neoplasm scheme[6] reassigned CMML into a separate MDS/MPD disease group; hence, this group of 25 patients was considered a distinct group during the subsequent analysis. This left 139 validated MDS specimens with 4 specimens returning a tie with the DC model: 3 had a 3-way tie between MDS/AML/"none-of-the-targets," and one had a tie between categories AML and "none-of-the-targets."

Comparison of validated MDS samples with de novo AML and "none-of-the-targets" classes

For a comparison analysis, exemplary MILE study data from patients submitted as AML and "none-of-the-targets" by the 4 contributing centers were used. These were 202 AML specimens and 69 "none-of-the-targets" GEPs. These specimens were combined with the 135 validated MDS specimens, excluding those with tied calls, for a PCA using the probe sets from the DC model (Figure 2A). In the 3-dimensional plot, a partial overlap was observed for MDS and AML, as well as for MDS and "none-of-the-targets" samples (not highlighted). The hierarchical clustering analyses (Figure 2B) of the same data are consistent with MDS gene expression illustrating a biologic continuum from AML to a nonleukemic disease. MDS samples do not form clear and distinct clusters but rather are interspersed among AML or "none-of-the-targets" specimens, whereas many AML specimens form clear clusters. Notably, there was no clustering associated with processing center, age, or sex. Equally, no distinct separation on the basis of IPSS was seen (Figure 3).

Trends were seen between MDS and DC model calls: none of the 5q− MDS samples had received an AML call (63% were called MDS; 37% were called "none-of-the-targets"), 86% of RARS specimens had received an MDS call, 68% of RAEB2 specimens had received an AML call, 32% had received MDS, and none had received a "none-of-the-targets" call, whereas only 11% of RA specimens had an AML call. In addition, the influence of the blast cell count at diagnosis was investigated: 66% of the samples had less than 5% blasts, 19% had between 5% and 10% blasts, whereas 15% had more than 10% blasts. Approximately 19% of those samples called AML by the DC classifier had less than 5% blasts, whereas 9% of samples called "none-of-the-targets" had more than 5% blasts (supplemental Figure 1). Approximately 7% of

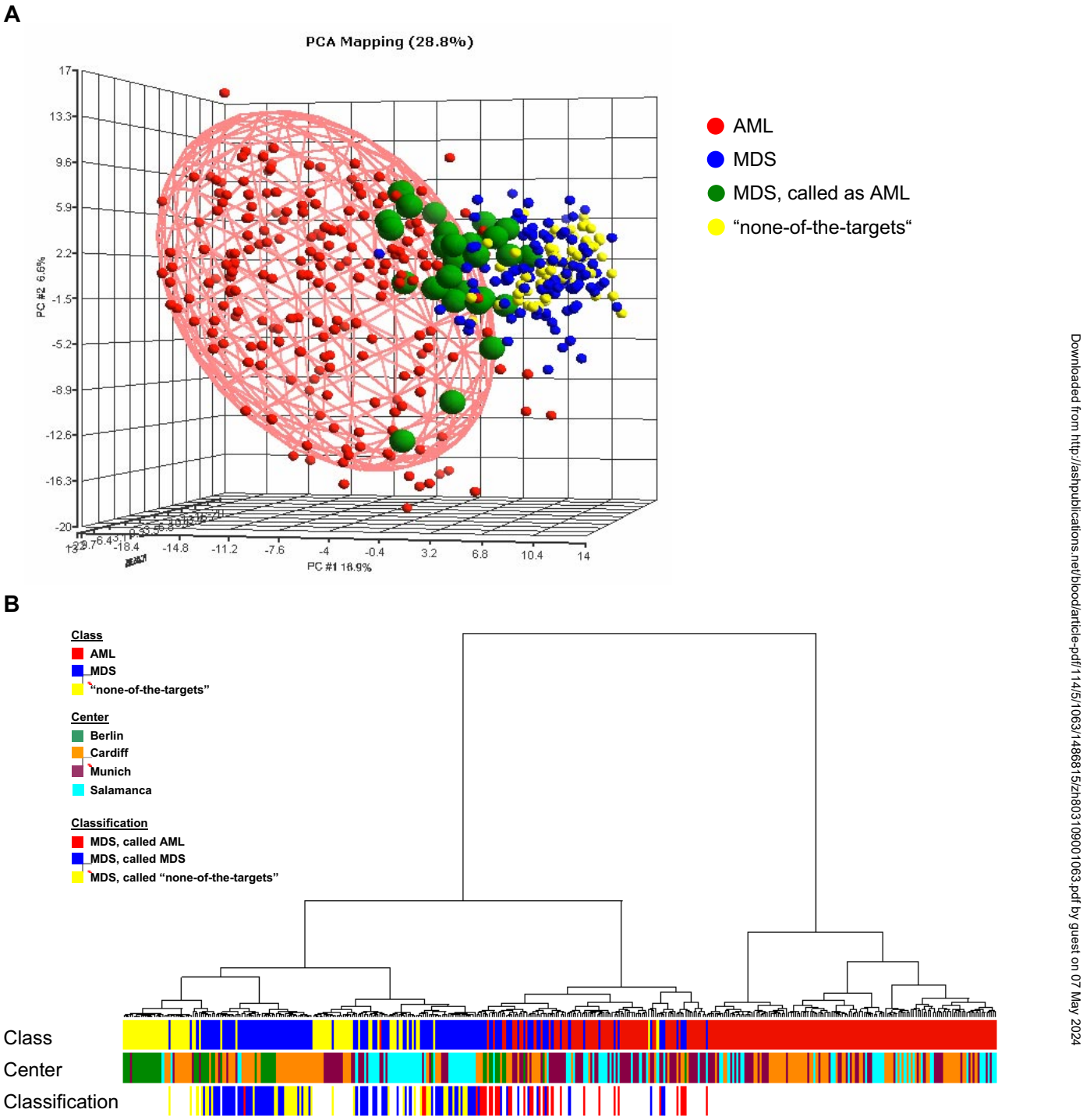

Figure 2. GEPs of MDS samples, de novo AML, and "none-of-the-targets" cases. (A) In the PCA, 406 cases are displayed based on the 534-probe set signature from the DC model. The first 3 principal components accounted for 28.8% of variation of the data (component 1 = 16.9%; component 2 = 6.6%; component 3 = 5.3%). Each sphere represents a single GEP. The AML (n = 202) and "none-of-the-targets" (n = 69) samples are colored according to the initial diagnosis. The shape of the AML ellipsoid was determined by the variability within the AML samples, and the ellipsoid was drawn to surround the samples within 2-fold SD. In the MDS group (n = 135), cases called by the diagnostic classifier as AML (n = 31) are further distinguished. Detailed information on the classifier probe sets is available online. (B) The agglomerative hierarchical clustering yields an entire hierarchy of clusters for all samples in the dataset. Euclidean distance was used to measure the dissimilarity between AML (n = 202), MDS (n = 135), and "none-of-the-targets" (n = 69) samples. Ward's minimum-variance method was used to determine the hierarchy and to define the groups. The average width of the clustering structure was represented as dendrogram in the clustering tree. The samples are annotated according to diagnostic category (Class), laboratory where the microarray analyses were performed (Center), and results of the DC model (Classification).

patients with an AML or MDS call and less than 5% blasts transformed. However, 17% of the patients who did transform had less than 5% blasts. These results would indicate that the blast cell count was not the only parameter contributing to either AML transformation or the molecular classifications based on the gene expression microarray analysis.

A similar analysis for the 25 CMML specimens showed that 12 (48%) were called by the DC model as MDS, 6 (24%) were called

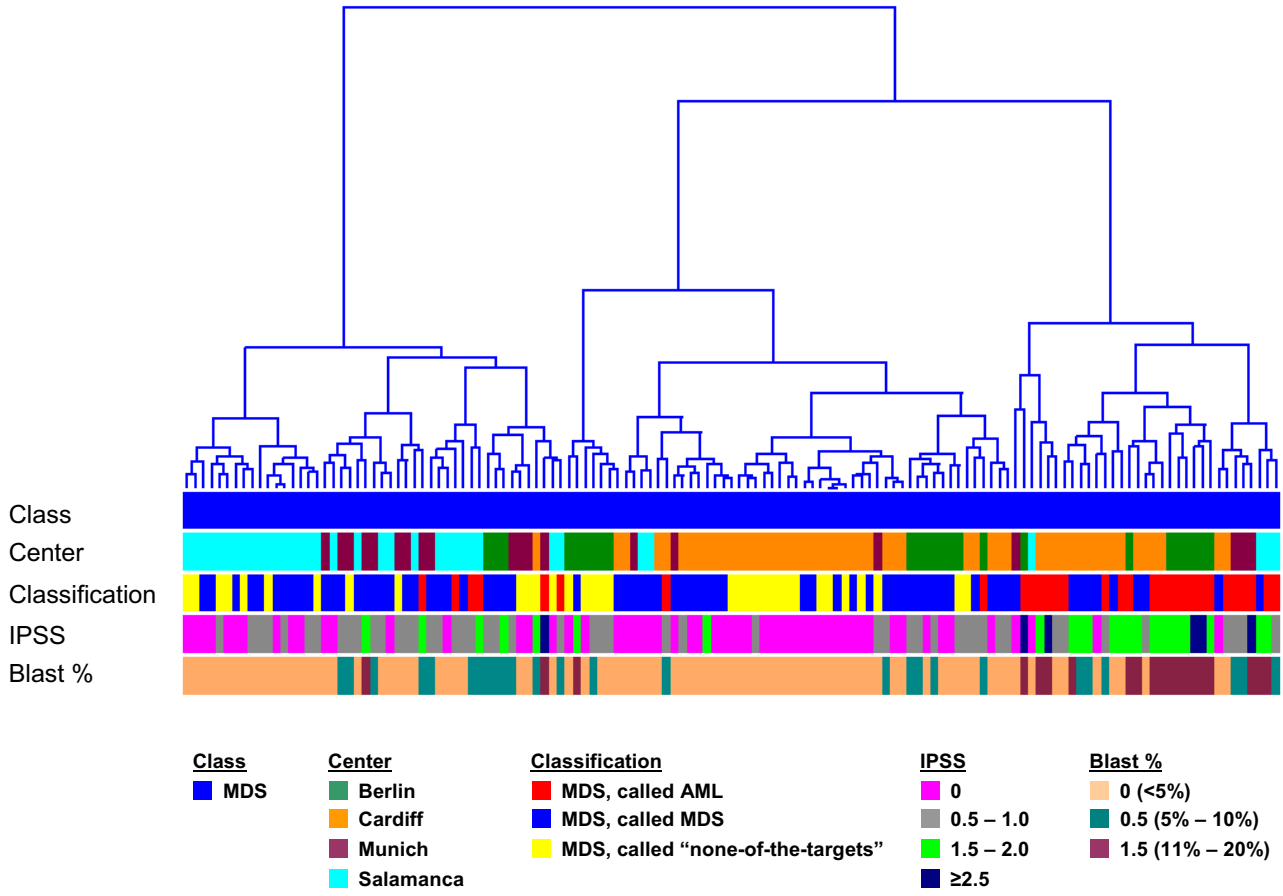

Figure 3. Hierarchical clustering of MDS samples. In the analysis, 135 MDS cases are displayed based on the 534-probe set signature from the diagnostic classifier. Euclidean distance was used to measure the dissimilarity between the MDS cases, and Ward's minimum-variance method was used to determine the hierarchy and to define the groups. The average width of the clustering structure was represented as dendrogram in the clustering tree. The samples are annotated according to diagnostic category (Class), laboratory where the microarray analyses were performed (Center), results of the DC model (Classification), IPSS score (IPSS), and percentages of blast cells (Blast %).

AML, 5 (20%) were called as "none-of-the-targets," with the remaining 2 (8%) having tied calls between AML, MDS, and "none-of-the-targets." These proportions were similar to those seen for the MDS cohort. With respect to the blast cell count, the CMML specimens with AML calls had a median of 5% blasts, those with MDS calls had a median of 2% blasts, and those with "none-of-the-targets" calls had a median of 3% blasts.

Prognostic outcome grouped by the DC model

A total of 107 of 110 validated MDS specimens with outcome data available were uniquely classified by the DC model. The median follow-up period for these patients was 27 months. A Kaplan-Meier analysis was applied to both overall survival and time to AML transformation from diagnosis. In Figure 4A, overall survival grouped by the DC model calls showed a nonsignificant P value (.167) of the log-rank test. The 5-year survival rates of AML, MDS, and "none-of-the-targets" were 15% (n = 25), 24% (n = 55), and 50% (n = 27). The median survival times were 26, 35, and 50 months, respectively.

However, there were significant differences (P value of log-rank test, $< 8 \times 10^{-5}$) in time to AML transformation among the 3 molecular classification groups: none (0%) of 27 MDS patients with a "none-of-the-targets" call, 8 (14.5%) of 55 patients with MDS call, and 11 (44%) of 25 patients with an AML call (Figure 4B). All patients with an AML call that transformed did so within 18 months; only one patient in this group had a censor date more than 72 months. The 18-month AML transformation rate for MDS patients with a MDS call was 8%. No patient with a "none-of-the-targets" call by the gene expression microarray algorithm transformed to AML within 5 years.

Prognostic classifier for MDS

The DC model had been designed to function as a diagnostic classifier for leukemia and MDS, not as a prognostic classifier. However, in light of the prognostic implications of the MDS "miscalls" using the DC, the expression data were reevaluated with the aim of producing a prognostic risk score aimed at the prediction for 3 groups of MDS patients with respect to time to AML transformation: group A, less than 18 months; group B, at least 18 months and less than 36 months; and group C, more than 36 months. Two datasets were used in the development of a hierarchical microarray-based risk score: dataset N74 was used to generate a classifier for the groups A (given a risk score of 2) and B, whereas the dataset N43 was used to generate the classifier for the groups B (risk score 1) and C (risk score 0). The Kaplan-Meier curves for overall survival and time to AML transformation by the risk score and the corresponding LOO risk scores for the dataset N110 are shown in Figure 5. Two variations of risk scores were calculated: one using resubstitution (Figure 5B,D) and a second method of using LOO cross-validation (Figure 5C,E). Both analyses showed highly significant differences between the groups for overall survival (Figure 5B-C) and time to AML transformation (Figure 5D-E). A nonhierarchical method, in which the resulting score of 0 or 1 from each of the 2 classifiers were combined, also

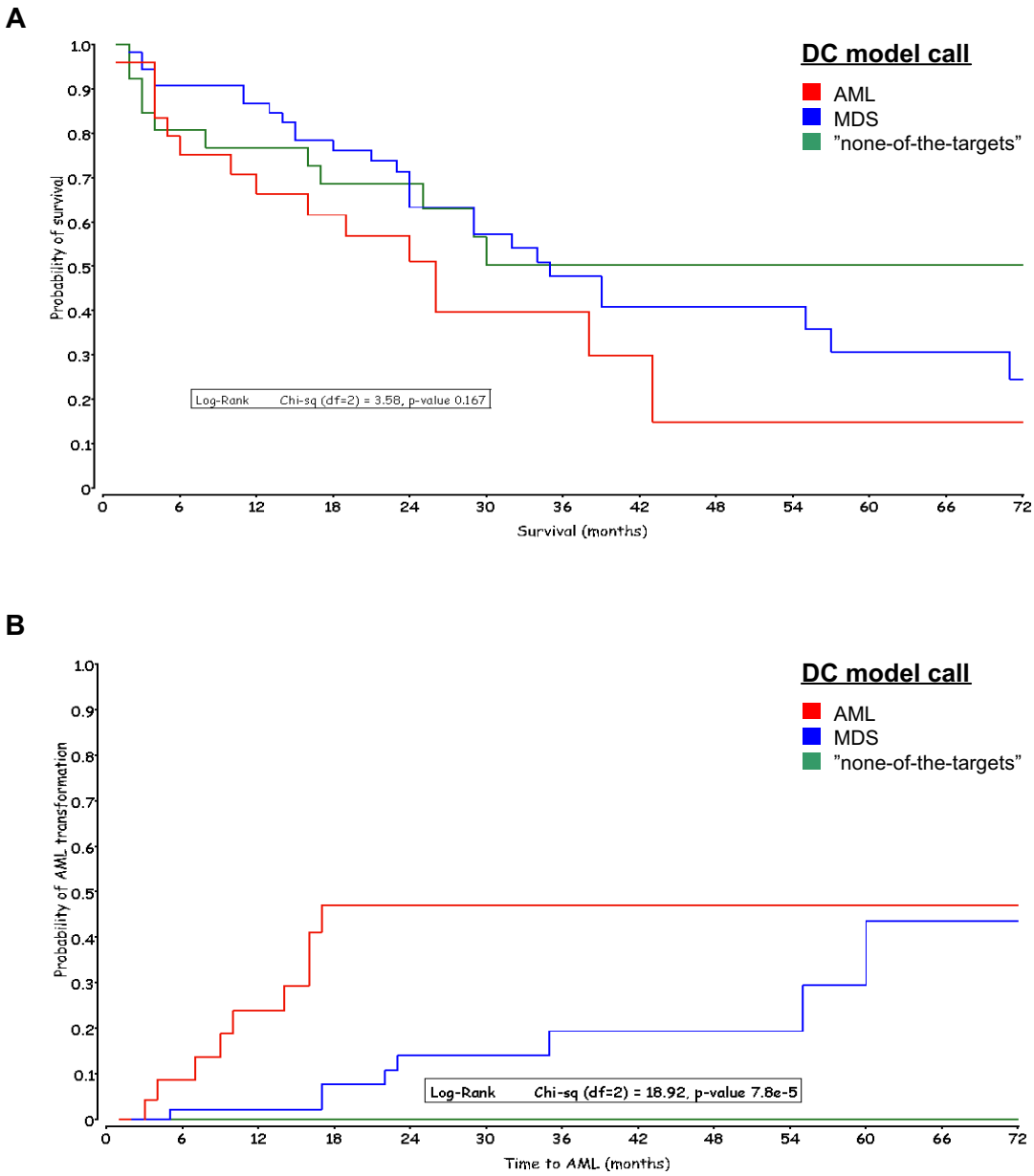

Figure 4. Kaplan-Meier curves grouped by DC model calls. The MDS samples are colored according to the DC model call. (A) Overall survival in months after diagnosis of MDS. (B) Time to AML transformation in months after diagnosis of MDS.

resulted in significant separation of the 3 patient groups (supplemental Figure 2).

Covariate analysis

To further assess the effects of covariates of interest, we calculated the hazard ratios in Cox proportional hazards models. Table 2 lists the univariate and multivariate hazard ratios of the IPSS, diagnostic groups (5q−, RA, RARS, RAEB1, RAEB2, and RCMD), and the DC and PC microarray models. For time to AML transformation, the diagnostic WHO group was not significant ($P = .099$). Similarly, for overall survival, the diagnostic group and the DC model call were not significant. The multicovariate hazard ratios showed that the PC model risk scores were more significant than IPSS for both time to AML transformation ($P = .005$) and overall survival ($P = .009$). Similarly, both the DC model calls and PC model scores were more significant than the individual IPSS components: blast count score, karyotype score, and cytopenia score (Table 2).

Molecular pathway analysis

The genes (probe sets) used by the DC and PC microarray models were further studied with a pathway analysis application. Gene-by-gene interactions for the model classes is shown in supplemental Figures 3 and 4.

Discussion

The Gene Expression Profiling working group (WP13) of the ELN initiated the international MILE study program in 7 ELN centers, 3 centers from the United States, and one center in Singapore.[20] The aim of the MILE study was to compare the concordance of classification of

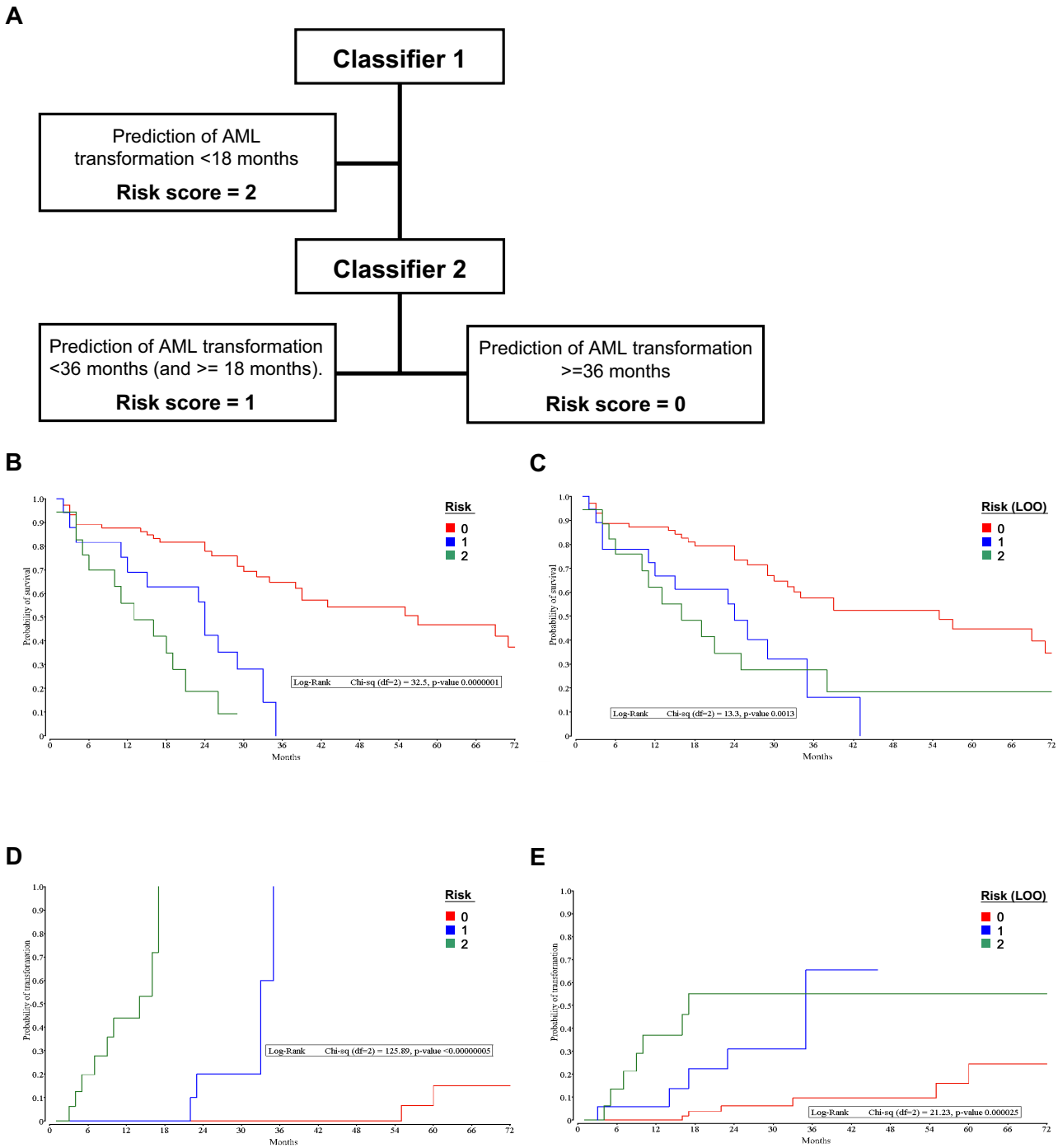

Figure 5. PC model analysis. (A) Flow diagram for the calculation of the MDS risk score. Kaplan-Meier curves grouped by PC model scores. (B) Overall survival after diagnosis of MDS, calculated by the resubstitution classifiers. (C) Overall survival after diagnosis of MDS, calculated by the hierarchical LOO classifiers. (D) Time to AML transformation after diagnosis of MDS, calculated by the resubstitution classifiers. (E) Time to AML transformation after diagnosis of MDS, calculated by the hierarchical LOO classifiers.

16 acute and chronic leukemia subclasses, MDS, and a "none-of-the-targets" group (as a control or normal class) derived from GEPs, with conventional routine diagnostic methods. The conventional approaches included a routine diagnostic workup based on all the currently used methods combined as needed, such as cytomorphology, flow cytometry, cytogenetics, fluorescence in situ hybridization, or reverse-transcribed polymerase chain reaction assays. There was no attempt made to standardize the traditional approach, and each participating center applied methodologies routinely applied at the corresponding laboratory. Within the first phase of the MILE study (stage I), standardized gene expression microarray protocols were used to study 2143 patients.[20,22] As part of the study design, designated laboratory operators were trained during a prephase on the corresponding sample preparation protocol and microarray workflow to ensure standardized assay conditions and robust laboratory proficiency. The overall classification accuracy for all the MILE leukemia classification categories (ie, excluding MDS) demonstrated approximately 95% concordance in cross-validation runs using stage I and pre-MILE study data.[20] It was an unexpected observation that the DC model made a correct diagnostic call in only 50% of MDS patients with a clinical diagnosis of MDS; the classifications for the remaining specimens split roughly equally

Table 2. Univariate and multivariate analysis for the DC model and leave-one-out risk scores for the PC model

	Time to AML transformation			Overall survival		
Covariate	***P***	**Hazard ratio**	**Confidence interval**	***P***	**Hazard ratio**	**Confidence interval**
Univariate hazard ratios						
IPSS score	.002*	1.98	1.29-3.05	.034	1.33	1.02-1.74
Disease classification	.099	1.29	0.95-1.73	.258	0.90	0.74-1.08
DC model	.001*	4.29*	1.94-9.49	.092	1.39	0.95-2.05
PC model	< .001*	2.90*	1.73-4.87	.001*	1.72	1.25-2.35
Multicovariate hazard ratios						
IPSS score	.604	1.15	0.69-1.91	.584	1.10	0.79-1.53
DC model	.023*	2.65*	1.15-6.12	.628	1.11	0.73-1.69
PC model	.005*	2.33*	1.30-4.18	.009*	1.60	1.13-2.27
Multicovariate hazard ratios						
Blast score	.472	1.43	0.54-3.79	.504	0.80	0.42-1.53
Karyotype score	.806	1.21	0.26-5.58	.344	1.59	0.61-4.16
Cytopenia score	.853	0.79	0.07-9.24	.211	2.30*	0.62-8.48
DC model	.015*	3.58*	1.28-9.99	.287	1.29	0.81-2.05
Multicovariate hazard ratios						
Blast score	.130	1.90	0.83-4.35	.526	0.84	0.48-1.46
Karyotype score	.612	1.49	0.32-6.95	.313	1.65	0.63-4.33
Cytopenia score	.828	1.28	0.14-11.8	.462	1.63	0.44-6.05
PC model	.002*	2.46*	1.39-4.36	.004*	1.66	1.17-2.35

*$P < .05$ or hazard ratio > 2.0.

between AML and "none-of-the-targets." An external morphology review confirmed that this discordance was not the result of errors in the original diagnoses. A total of 164 of the 174 submitted MDS specimens were indeed MDS, with 6 of the 10 other specimens being reclassified as AML, a diagnosis that was consistent with the microarray DC model results.

A correlation was noted for MDS disease subtype and IPSS score[7] with a higher proportion of high IPSS scores also having an AML call. The IPSS score is derived from weighted contributions of the number of blasts, cytogenetic aberrations, and number of cytopenic lineages. No correlation with cytopenia score or cytogenetic abnormalities was seen. The correlation of blast count and the DC model call was not exact and showed that a proportion of patients with an AML call had less than 5% blasts, whereas some patients with a "none-of-the-targets" call had greater than 10% blasts. This would suggest that those patients with an AML call, but low blast count, had molecular features apparent in GEPs of AML without the corresponding morphologic blast appearance.

The DC model had been designed to be a diagnostic classifier for 16 classes of leukemia, MDS, and "none-of-the-targets" across chronic and acute classes of lymphoid or myeloid malignancy, and the observed correlation with time to AML transformation or overall survival was not part of the original hypothesis. Disease classification by microarray technology has been reported in several AML and MDS studies[13,15,28] in addition to the high accuracy seen in the MILE study.[20] The DC model calls of the MDS patients included in the MILE study showed a significant association with time to AML transformation, but not to overall survival. However, it should be noted that the lack of correlation with IPSS[7] may have been expected as this scoring system was based on only 25% of patients, in each risk category, undergoing evolution to acute myeloid leukemia. Interestingly, those patients with an AML classification call that did transform did so within 18 months from diagnosis, whereas none of the patients with the "none-of-the-targets" call transformed. The majority of patients with an MDS classification result, who did transform to AML, did so more slowly, over a 5-year period. This discrepancy was exploited for the development of a time to AML transformation risk score by subdividing patients into early and late transformers. The Kaplan-Meier analysis and univariate and multivariate hazard ratios all showed high significance using the risk score calculated with the LOO approach.

Pathway analysis of the genes contributing to the discrimination between the MDS molecular subgroups identified several networks that are involved in the progression from "none-of-the-targets" through MDS to AML, with several pathways indicating that these are involved in both steps of disease progression. A similar analysis of the pathways and interacting genes from the LOO risk scores highlights several genes known to be actively involved in acute myeloid leukemia, including *HOX* cluster genes, *FLT3*, *KIT*, *RUNX1*, and *WT1*. *HOX* genes often form the basis of GEP lists in studies on AML.[29-32] The other candidate genes are also often mutated in AML and have prognostic significance, particularly in AML patients with a normal karyotype.[33-37] Furthermore, some of the genes are associated with therapy-related progression from MDS to AML.[38]

In conclusion, the DC model, evaluated as part of the MILE study program, was designed and built for improving clinical diagnosis but did not show the same high accuracy for MDS compared with the other 16 lymphoid or myeloid acute leukemia diagnostic subgroups. However, an expected, and significant, correlation with the time to AML transformation was observed, which led to the development of a prognostic algorithm that can identify MDS patients with high, intermediate, and low risk of progression to AML, based only on the respective GEPs from a microarray analysis. Thus, the molecular signatures may go beyond morphology, phenotype, and cytogenetics by removing any subjective assessment with an objective assessment based on a series of measurements of specific RNA levels using microarrays independent of any of these parameters. In addition, the genes involved in the predictive scores may also allow the development of targeted therapies for MDS patients with poor prognosis.

Acknowledgments

The authors thank Amanda Gilkes, Eva Lumbreras, Verena Nowak, and Sonja Schindela for excellent technical assistance; Tayside Cancer Tissue Bank, Dundee, which contributed to the U.K. samples; Michael Groves and Norene Keenan who processed samples and clinical data, respectively; Li Qiu, James Sun, and Yan Li for data management;

Sunhee K. Ro and Xiaoying Chen for support in data analysis; Julie Tsai and Nang Tan for assay development; Andrea Johnson, Brian Rhees, and Jing Wang for helpful discussions; Philip X. Xiang for supporting the GEO submission of microarray data; and all clinicians who provided data and samples.

Authorship

Contribution: K.I.M. contributed to the research, analyzed data, and wrote the manuscript; A.K. was responsible for the study conduct and wrote the manuscript; P.M.W. contributed to the research and analyzed data; L.W. contributed to the study conduct and supported the development of the classifier algorithms; W.-m.L. and R.L. performed data analysis; W.W. contributed vital reagents and supported the development of the classifier algorithms; D.T.B. contributed data and was involved in the external morphology review; H.L. was involved in the external morphology review; J.M.H. and W.-K.H. contributed to the research and edited the manuscript; and T.H. contributed to the study funding, research, and data analysis and edited the manuscript.

Conflict-of-interest disclosure: This study was part of the MILE study (Microarray Innovations in LEukemia) program, a collaborative effort headed by the ELN, sponsored in part by Roche Molecular Systems Inc, and investigating gene expression signatures in acute and chronic leukemia. This study further supported the Roche Molecular Systems AmpliChip Leukemia Test research program, a gene expression microarray test for the subclassification of leukemia and myelodysplastic syndromes.

For a complete list of MILE Study WP-13 participants, see the online supplemental Appendix.

Correspondence: Ken I. Mills, Centre for Cancer Research & Cell Biology, Queen's University Belfast, 97 Lisburn Rd, Belfast, BT9 7BL, United Kingdom; e-mail: k.mills@qub.ac.uk.

References

1. Nimer SD. Myelodysplastic syndromes. *Blood*. 2008;111:4841-4851.
2. Bennett JM, Catovsky D, Daniel MT, et al. Proposals for the classification of the myelodysplastic syndromes. *Br J Haematol*. 1982;51:189-199.
3. Malcovati L, Nimer SD. Myelodysplastic syndromes: diagnosis and staging. *Cancer Control*. 2008;15[Suppl]:4-13.
4. Mufti GJ, Bennett JM, Goasguen J, et al. Diagnosis and classification of myelodysplastic syndrome: International Working Group on Morphology of Myelodysplastic Syndrome (IWGM-MDS) consensus proposals for the definition and enumeration of myeloblasts and ring sideroblasts. *Haematologica*. 2008;93:1712-1717.
5. Vardiman JW, Thiele J, Arber DA, et al. The 2008 revision of the WHO classification of myeloid neoplasms and acute leukemia: rationale and important changes. *Blood*. 2009;114:937-951.
6. Vardiman JW, Harris NL, Brunning RD. The World Health Organization (WHO) classification of the myeloid neoplasms. *Blood*. 2002;100:2292-2302.
7. Greenberg P, Cox C, LeBeau MM, et al. International scoring system for evaluating prognosis in myelodysplastic syndromes. *Blood*. 1997;89:2079-2088.
8. 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;25:3503-3510.
9. Bowen DT, Fenaux P, Hellstrom-Lindberg E, de Witte T. Time-dependent prognostic scoring system for myelodysplastic syndromes has significant limitations that may influence its reproducibility and practical application. *J Clin Oncol*. 2008;26:1180-1182.
10. Kohlmann A, Haschke-Becher E, Wimmer B, et al. Intraplatform reproducibility and technical precision of gene expression profiling in 4 laboratories investigating 160 leukemia samples: the DACH study. *Clin Chem*. 2008;54:1705-1715.
11. Shi L, Reid LH, Jones WD, et al. The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat Biotechnol*. 2006;24:1151-1161.
12. Golub TR, Slonim DK, Tamayo P, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*. 1999;286:531-537.
13. Bullinger L, Dohner K, Bair E, et al. Use of gene-expression profiling to identify prognostic subclasses in adult acute myeloid leukemia. *N Engl J Med*. 2004;350:1605-1616.
14. Haferlach T, Kohlmann A, Schnittger S, et al. A global approach to the diagnosis of leukemia using gene expression profiling. *Blood*. 2005;106:1189-1198.
15. Valk PJ, Verhaak RG, Beijen MA, et al. Prognostically useful gene-expression profiles in acute myeloid leukemia. *N Engl J Med*. 2004;350:1617-1628.
16. Mano H. DNA micro-array analysis of myelodysplastic syndrome. *Leuk Lymphoma*. 2006;47:9-14.
17. Pellagatti A, Cazzola M, Giagounidis AA, et al. Gene expression profiles of CD34+ cells in myelodysplastic syndromes: involvement of interferon-stimulated genes and correlation to FAB subtype and karyotype. *Blood*. 2006;108:337-345.
18. Pellagatti A, Fidler C, Wainscoat JS, Boultwood J. Gene expression profiling in the myelodysplastic syndromes. *Hematology*. 2005;10:281-287.
19. Qian J, Chen Z, Wang W, Cen J, Xue Y. Gene expression profiling of the bone marrow mononuclear cells from patients with myelodysplastic syndrome. *Oncol Rep*. 2005;14:1189-1197.
20. Haferlach T, Kohlmann A, Basso G, et al. The clinical utility of microarray-based gene expression profiling in the diagnosis and subclassification of leukemia: final report on 3252 cases from the International MILE Study Group. *ASH Annual Meeting Abstracts*. 2008;112:753.
21. Mills KI, Gilkes AF, Hernandez JM, et al. A molecular classification of leukaemia reveals MDS as a disease continuum with non-leukaemia and AML sub groups. *Haematologica*. 2007;92:165.
22. Kohlmann A, Kipps TJ, Rassenti LZ, et al. An international standardization programme towards the application of gene expression profiling in routine leukaemia diagnostics: the Microarray Innovations in LEukemia study prephase. *Br J Haematol*. 2008;142:802-807.
23. Liu WM, Li R, Sun JZ, et al. PQN and DQN: algorithms for expression microarrays. *J Theor Biol*. 2006;243:273-278.
24. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A*. 1998;95:14863-14868.
25. Barrett T, Suzek TO, Troup DB, et al. NCBI GEO: mining millions of expression profiles—database and tools. *Nucleic Acids Res*. 2005;33:D562-D566.
26. Chang CC, Lin CJ. LIBSVM: a library for support vector machines, 2001. http://www.csie.ntu.edu.tw/~cjlin/libsvm.
27. Vapnik VN. *The Nature of Statistical Learning Theory*. New York, NY: Springer-Verlag; 1995.
28. Pellagatti A, Esoof N, Watkins F, et al. Gene expression profiling in the myelodysplastic syndromes using cDNA microarray technology. *Br J Haematol*. 2004;125:576-583.
29. Camos M, Esteve J, Jares P, et al. Gene expression profiling of acute myeloid leukemia with translocation t(8;16)(p11;p13) and MYST3-CREBBP rearrangement reveals a distinctive signature with a specific pattern of HOX gene expression. *Cancer Res*. 2006;66:6947-6954.
30. Grubach L, Juhl-Christensen C, Rethmeier A, et al. Gene expression profiling of Polycomb, Hox and Meis genes in patients with acute myeloid leukaemia. *Eur J Haematol*. 2008;81:112-122.
31. Mullighan CG, Kennedy A, Zhou X, et al. Pediatric acute myeloid leukemia with NPM1 mutations is characterized by a gene expression profile with dysregulated HOX gene expression distinct from MLL-rearranged leukemias. *Leukemia*. 2007;21:2000-2009.
32. Roche J, Zeng C, Baron A, et al. Hox expression in AML identifies a distinct subset of patients with intermediate cytogenetics. *Leukemia*. 2004;18:1059-1063.
33. Abu-Duhier FM, Goodeve AC, Wilson GA, et al. FLT3 internal tandem duplication mutations in adult acute myeloid leukaemia define a high-risk group. *Br J Haematol*. 2000;111:190-195.
34. Bacher U, Haferlach C, Kern W, Haferlach T, Schnittger S. Prognostic relevance of FLT3-TKD mutations in AML: the combination matters—an analysis of 3082 patients. *Blood*. 2008;111:2527-2537.
35. Frohling S, Schlenk RF, Breitruck J, et al. Prognostic significance of activating FLT3 mutations in younger adults (16 to 60 years) with acute myeloid leukemia and normal cytogenetics: a study of the AML Study Group Ulm. *Blood*. 2002;100:4372-4380.
36. Sheikhha MH, Awan A, Tobal K, Liu Yin JA. Prognostic significance of FLT3 ITD and D835 mutations in AML patients. *Hematol J*. 2003;4:41-46.
37. Virappane P, Gale R, Hills R, et al. Mutation of the Wilms' tumor 1 gene is a poor prognostic factor associated with chemotherapy resistance in normal karyotype acute myeloid leukemia: the United Kingdom Medical Research Council Adult Leukaemia Working Party. *J Clin Oncol*. 2008;26:5429-5435.
38. Pedersen-Bjergaard J, Andersen MK, Andersen MT, Christiansen DH. Genetics of therapy-related myelodysplasia and acute myeloid leukemia. *Leukemia*. 2008;22:240-248.