MYELOID NEOPLASIA

Comparing cancer vs normal gene expression profiles identifies new disease entities and common transcriptional programs in AML patients

Nicolas Rapin,[1-4] Frederik Otzen Bagger,[1-4] Johan Jendholm,[1,2,4] Helena Mora-Jensen,[5] Anders Krogh,[2,3] Alexander Kohlmann,[6] Christian Thiede,[7] Niels Borregaard,[5] Lars Bullinger,[8] Ole Winther,[2,3,9] Kim Theilgaard-Mönch,[1,2,10] and Bo T. Porse[1,2,4]

[1]The Finsen Laboratory, Rigshospitalet, Faculty of Health Sciences, [2]Biotech Research and Innovation Centre, [3]The Bioinformatics Centre, Department of Biology, Faculty of Natural Sciences, [4]Danish Stem Cell Centre, Faculty of Health Sciences, and [5]The Granulocyte Research Laboratory, Rigshospitalet, Faculty of Health Sciences, University of Copenhagen, Copenhagen, Denmark; [6]Munich Leukemia Laboratory, Munich, Germany; [7]University Hospital Carl Gustav Carus, Medical Department 1, University of Technics Dresden, Germany; [8]Department of Internal Medicine III, University Hospital of Ulm, Ulm, Germany; [9]DTU Compute, Technical University of Denmark, Lyngby, Denmark; and [10]Department of Hematology, Skanes University Hospital, University of Lund, Sweden

Key Points

- This study describes a method for the comparison of gene expression data of any type of cancer cells with their corresponding normal cells.
- Our analyses reveal novel disease entities, identify common deregulated transcriptional networks, and predict survival.

Gene expression profiling has been used extensively to characterize cancer, identify novel subtypes, and improve patient stratification. However, it has largely failed to identify transcriptional programs that differ between cancer and corresponding normal cells and has not been efficient in identifying expression changes fundamental to disease etiology. Here we present a method that facilitates the comparison of any cancer sample to its nearest normal cellular counterpart, using acute myeloid leukemia (AML) as a model. We first generated a gene expression-based landscape of the normal hematopoietic hierarchy, using expression profiles from normal stem/progenitor cells, and next mapped the AML patient samples to this landscape. This allowed us to identify the closest normal counterpart of individual AML samples and determine gene expression changes between cancer and normal. We find the cancer vs normal method (CvN method) to be superior to conventional methods in stratifying AML patients with aberrant karyotype and in identifying common aberrant transcriptional programs with potential importance for AML etiology. Moreover, the CvN method uncovered a novel poor-outcome subtype of normal-karyotype AML, which allowed for the generation of a highly prognostic survival signature. Collectively, our CvN method holds great potential as a tool for the analysis of gene expression profiles of cancer patients. (*Blood*. 2014;123(6):894-904)

Introduction

Global gene expression profiling (GEP) has been used for more than a decade to uncover the underlying transcriptional programs of normal and malignant cells. In cancer, GEP has been used successfully to identify cancer subtypes, to stratify patients into responders vs nonresponders, and to predict survival but has, to a large extent, failed to uncover genes that are causally involved in cancer initiation and maintenance.[1-8] These genes are obviously of great interest because they constitute potential targets for therapeutic intervention.

The reasons for the failure of GEP to uncover targets of therapeutic relevance are potentially many and may differ between different cancer types. However, as most studies compare cancer with cancer, a lot of the detected transcriptional changes between different cancer samples may arise from differences in cell type and developmental stage and, consequently, will not identify those gene expression programs that underlie the malignant phenotype. Moreover, in the few instances in which comparisons of cancer cells with normal cells have been attempted, they have generally used heterogeneous cell mixtures of the organ in question (such as whole bone marrow [BM] or $CD34^+$ cells in acute myeloid leukemia [AML]), which is not ideal when the aim is to precisely map changes in gene expression between cancer cells and their nearest normal counterpart. Given the caveats with the standard procedures of analyzing large-scale GEP data sets from cancer patients, we hypothesized that a method that facilitates a comparison of gene expression profiles between cancer samples and their nearest normal counterpart holds the potential to significantly improve disease stratification, the identification of novel disease subtypes, prognostication, and the identification of gene expression changes underlying the malignant phenotype.

Here we present such a method and apply it retrospectively to 5 publically available data sets from AML patients.[9-15] We show that comparison of cancer vs normal (CvN) is equivalent to cancer vs cancer (CvC) in terms of GEP-based prediction of AML subtypes

Submitted February 20, 2013; accepted December 15, 2013. Prepublished online as *Blood* First Edition paper, December 20, 2013; DOI 10.1182/blood-2013-02-485771.

K.T.-M. and B.T.P. contributed equally to this study.

The data reported in this article have been deposited in the Gene Expression Omnibus database (accession number GSE42519).

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.

harboring different aberrant karyotypic lesions (AK-AML). Moreover, applying this method to data sets of AML patients with normal karyotype AML (NK-AML) lead to identification of a novel AML subtype associated with poor outcome. Finally, the method provides a list of deregulated genes and common transcriptional programs that differ from those of normal progenitors. These may potentially underlie the malignant phenotype in different AML subtypes.

Materials and methods

GEP of normal hematopoietic stem and progenitor cells

BM samples were aspirated from the posterior iliac crest of healthy subjects, according to the standard protocol of the Department of Hematology, Rigshospitalet University of Copenhagen, and in accordance with the Declaration of Helsinki. Normal hematopoietic stem and progenitor cells (HSCs/HPCs), as well as mature myeloid cells representing successive developmental stages of the hematopoietic differentiation, were purified from BM samples of healthy subjects by flow cytometry-based cell sorting and subjected to GEP (see the supplemental Methods, available on the *Blood* Web site). All microarray files of our newly generated data set are available at the Gene Expression Omnibus database (accession number GSE42519).

Bioinformatics analyses

Gene expression-based map of the normal hematopoietic hierarchy. The normalized[16] and batch-corrected[17] data set of normal blood and BM populations was filtered to exclude probe sets with low variance and low expression (see Table 1 for sample list and origin[18-22] and the supplemental Methods for data normalization and batch correction). Probe sets with a standard deviation 3 times lower than the average standard deviation of the entire data set and with an average log2 expression below 6 were excluded from the analysis, yielding a total of 2119 probe sets (1367 unique gene symbols). These were used in a principal component analysis (PCA)[23] to generate a map of the normal hematopoietic hierarchy by projection of the data onto the first 2 principal components (ie, the 2 directions, which explain the most variance in the 2119 dimensional space of standardized gene expression values).

Mapping of AML samples onto the gene expression-based map of the normal hematopoietic hierarchy. Each AML sample was normalized together with the data set of normal cells, using the same procedure described earlier. The normal populations closest to each individual AML sample were identified by a 2-step approach. First, we computed the Euclidian distances between the AML sample and all normal populations, using the expression profiles projected to the first 6 principal components, which explain more than 90% of the variation for more than 95% of the samples. Second, for each individual AML sample, we selected the 50% most varying probe sets within the 15 closest normal samples and subsequently mapped the AML sample in this reduced local gene expression space. Using PCA projections in these steps reduces noise and computation time while preserving relevant information. For each AML sample mapped in the gene expression landscape of normal hematopoiesis, we next calculated a weighted average gene expression profile of the 5 nearest normal samples. The weights were based on the Euclidian distance between normal and AML samples in the reduced second PCA and were set to decrease exponentially with distance and, subsequently, renormalize to sum to 1. This approach gives more weight to populations closest to the AML sample. Finally, gene expression changes between individual AML samples and their corresponding individual average-weighted normal counterpart were computed.

For the standard CvC method, we generated an average GEP profile based on all the samples in a given AML data set. We used this to compute the log2 fold-changes for individual AML samples.

Table 1. Data sources

Data set	Sample number	Cell types	Reference
Normal blood and BM populations			
GSE42519	34	HSC, MPP, CMP, MEP, GMP, early PM, late PM, MY, MM, BC, PMN	This study
GSE17054	2	HSC	Majeti et al[19]
GSE19599	4	GMP, MEP	Andersson et al[20]
GSE11864	2	Monocytes	Hu et al[21]
E-MEXP-1242	2	Monocytes	Wildenberg et al[22]
Total	44		
AML patient samples			
GSE13159	191	t(8;21), inv(16), t(15;17), t(11q23), complex	Haferlach et al[9]
GSE14468	526	MDS:-5/7(q), −9q, 11q23, 8, complex; AML: NK, complex, inv(3q), inv(16), t(15;17), t(6;9), t(8;21), t(9;22)	Wouters et al[11]
GSE15434	251	NK-AML	Kohlmann et al,[2] Klein et al[12]
TCGA	183	Various genetic aberrations, including t(8;21), inv(16), t(15;17), t(11q23), complex	Cancer Genome Atlas Research Network[13]
GSE6891	91	Various genetic aberrations, including t(8;21), inv(16), t(15;17), t(11q23)	de Jonge et al[18]
GSE12417	79	NK-AML	Metzeler et al[15]
Total	1321		

Description and origin of all microarray data sets used in this study, including normal blood and BM populations and AML patient samples. All samples were run on the HG-U133 Plus 2.0 Array Affymetrix platform.

Gene set overlap analysis. We created gene sets using the 1% most up- and downregulated genes for each AML sample compared with its computed normal counterpart. Using a hypergeometric test, we calculated the significance of the overlap between gene sets of up- and downregulated genes against various AML-subtype signatures, as well as curated gene signatures (C2), gene ontology signatures (C5), and oncogenic signatures (C6) from the MSigDB molecular signature database (www.broadinstitute.org/gsea/msigdb/).

Further information on experimental procedures and bioinformatics analyses are provided in the supplemental Methods. Here we describe methods for supervised and unsupervised classification of AML, the identification of deregulated transcriptional programs in AML compared with normal, and the generation and performance assessment of a novel survival signature for NK-AML.

Results

The gene expression-based landscape of the normal hematopoietic hierarchy

Using our own and publically available microarray data sets (Table 1) of highly purified HSCs/myeloid HPCs and their mature progeny (see supplemental Figure 1A-B for cell sorting strategy), we generated a gene expression-based landscape representing the developmental hierarchy of the hematopoietic system using the first 6 components of a PCA (Figure 1A-B; gene lists provided in supplemental Table 1A). Significantly, this landscape not only faithfully reconstructed the

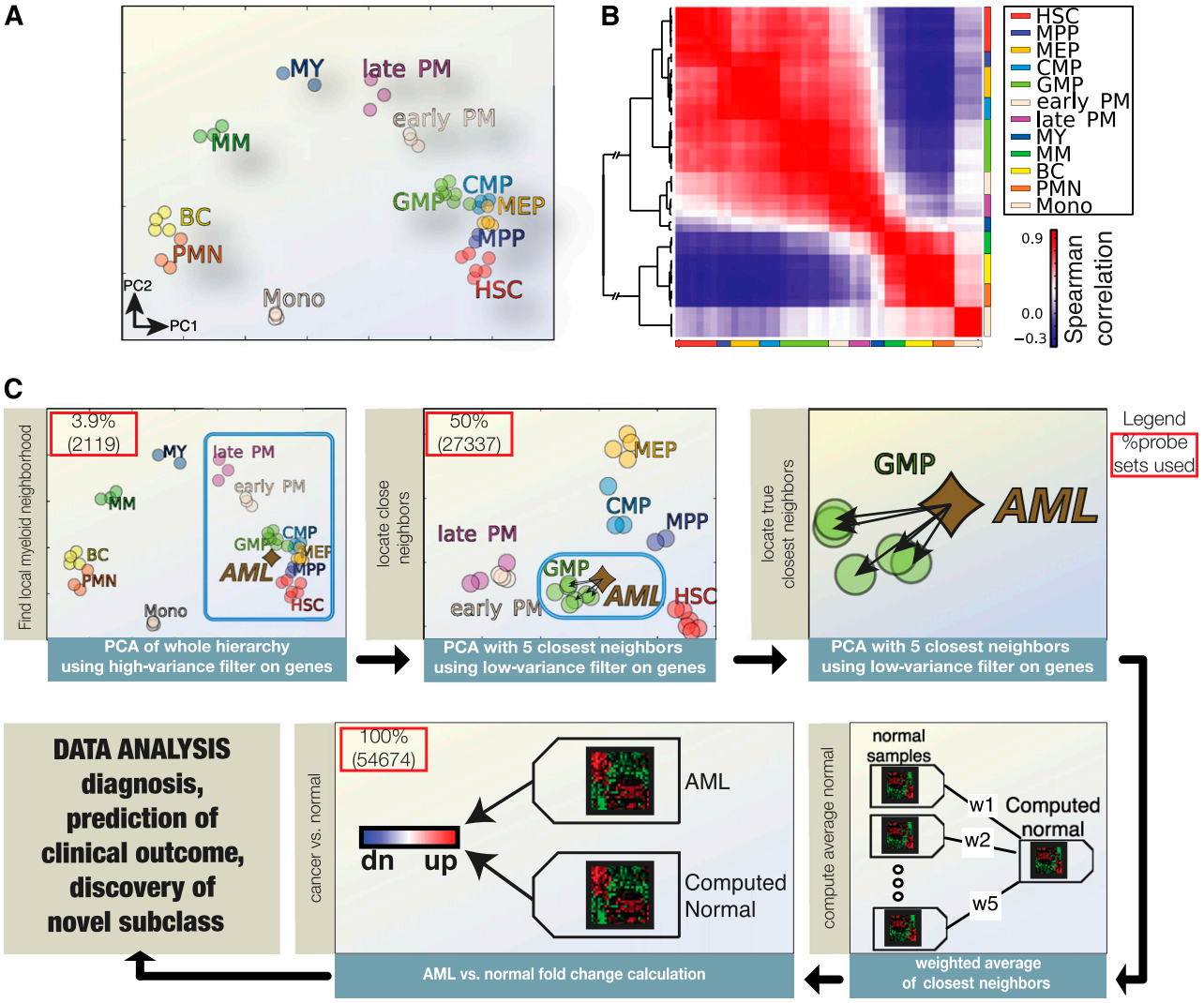

Figure 1. CvN method: identification of the nearest normal population for individual AK-AML samples, using a gene expression-based landscape of the normal hematopoietic hierarchy. (A) PCA of gene expression profiles from the following normal purified BM populations: HSCs, multipotent progenitors (MPPs), common myeloid progenitors (CMPs), granulocyte-monocyte progenitors (GMPs), megakaryocyte-erythrocyte progenitors (MEPs), early PM, late PM, myelocytes (MY), metamyelocytes (MM), band cells (BC), polymorphonuclear neutrophilic granulocytes (PMN_BM), and monocytes (Mono). The PCA was performed on 2119 probe sets (1367 genes) selected by variance filtering. (B) Spearman correlation matrix of gene expression of the samples from A. (C) Workflow of the CvN method for the identification of the nearest normal counterpart for individual AK-AML samples (CvN method): AML samples are normalized individually together with the data set of the normal hematopoietic hierarchy shown in A and B, and the normal populations closest to AML samples are identified by a 2-step approach. First, the Euclidian distances between each individual AML sample and all the normal blood and BM populations are calculated using gene expression profiles projected onto the first 6 principal components. Next, the 50% most varying probe sets within the 15 closest normal populations are selected for each individual AML samples and used in a second PCA to map the AML sample to its 5 nearest normal populations. Subsequently, a weighted average gene expression profile based on the Euclidian distance between the 5 normal populations and AML sample is calculated. Finally, gene expression profiles of AML samples and their corresponding individual average-weighted normal population are compared with defined differentially expressed genes in individual AML samples for enrichment analysis, prognostification, and further analyses.

normal hierarchical order of myeloid differentiation but also demonstrated a tight clustering of replicates from the same normal populations, thereby highlighting both the high quality of data processing and the data itself.

A method for assessment of gene expression changes between AML and its nearest normal counterpart (CvN method)

We next applied a 2-step approach to identify the closest normal population for individual AML patient samples using publically available AK-AML gene expression data sets (Table 1; supplemental Table 2). In the first step, we mapped individual AML samples onto the PCA space of normal hematopoietic differentiation, using genes selected by a high-stringency variance filter (Figure 1C). Next we reduced the filter stringency to increase mapping precision and identify the 5 closest normal BM populations. Finally, the GEPs of these 5 BM populations were merged into a "virtual" distance-weighted GEP representing the closest normal counterpart of the tested AML sample and subsequently used to calculate gene expression changes between normal and the individual AML sample (Figure 1C).

Mapping of individual AML patient samples to the gene expression landscape of the normal hematopoietic hierarchy demonstrated varying normal counterparts for different AML subtypes. Whereas samples of AML patients with a complex karyotype mapped to different normal populations, ranging from HSCs to monocytes, the more defined t(15;17) AMLs predominantly mapped closely to the related granulocyte-monocyte progenitors and early promyelocytes (early PMs), representing the developmental stage of this particular

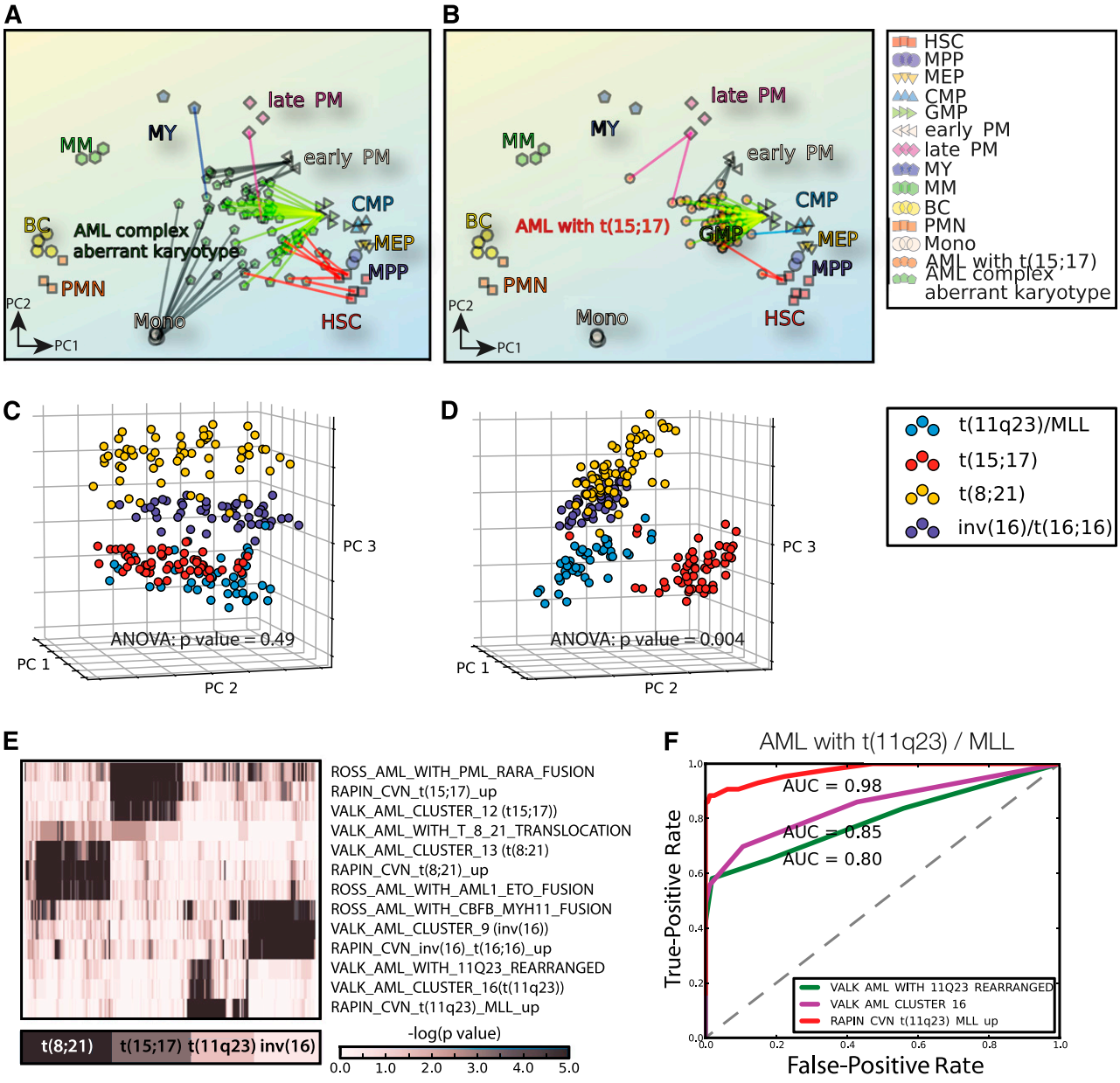

Figure 2. Cluster analysis of AK-AML samples using the CvN method and the CvC method. (A-B) Example of PCA plots of individual gene expression profiles of complex karyotype AML (A) and t(15;17) AML patient samples (B) projected to the gene expression-based map of the normal hematopoietic hierarchy, using the CvN method (see "Bioinformatics analyses" and Figure 1). Only the 2 first (ie, PC1 and PC2) PCs are given in the PCA plot. A line indicates the nearest normal counterpart for each of the AK-AML samples. (C-D) Unsupervised clustering of AK-AML. PCA of AK-AML based on genes identified by the CvC- (C) and CvN method (D). Genes were selected by variance (1545 and 1449 probe sets in (C) and (D); respectively). ANOVA analysis of the segregation of the clusters using the first 5 PCs reports *P* values of .49 for the CvC method (inter-group variance: 6.08, intra-group variance: 495.85) and .004 for the CvN method (inter-group variance: 115.58, intra-group variance: 1685.95) for the CvC and CvN methods, respectively. (E) ROC curves (classification performances) for 2 published and 1 novel AML t(11q23) gene signatures. Areas under ROC curve (AUCs) are reported in the graph. (F). Heat map representing the degree of enrichment (-log10[*P* value]) in the 1% upregulated genes in AML patients[9] of known and novel AK-AML gene signatures.

AML subtype (Figure 2A-B). Other AK-AML subtypes demonstrated mapping patterns between these extremes (supplemental Figure 2).

The CvN method is comparable to the classical CvC method in stratifying AK-AML patients

We next performed a side-by-side comparison of the CvN and CvC methods, using unsupervised standard analysis of a GEP data set derived from 4 distinct AML subtypes [inv16/t(16;16), t(11q23), t(15;17), t(8;21)]. Both methods generated distinct clusters, each representing a genetically defined AK-AML subclass ($P < 1e-5$; Kruskal–Wallis test), as visualized by either unsupervised PCA (Figure 2C-D) or hierarchical clustering (supplemental Figure 3A-B). However, testing the intercluster and intracluster variance of the first 5 principal components in an analysis of variance (ANOVA) test (Figure 2C-D) demonstrated that genes identified by the CvN method, but not the CvC method, form significant clusters in an unsupervised analysis ($P = .004$ and $P = .49$, respectively).

We next used a standard supervised classification analysis and found that genes selected by the CvN method performed as well as those selected by the CvC method for the training set[9] (error rates, CvC: 2.05%; CvN: 1.55%; supplemental Figure 4), as well as in an

independent AK-AML test set[18] (error rates, CvC: 8.62%; CvN: 6.46%). To test whether the CvN method correctly identifies transcriptional changes specific for distinct AK-AML subtypes, we generated signatures of the most discriminatory genes for each AML subtype (supplemental Table 3; supplemental Methods).[24] We next generated patient-specific signatures for each individual AML patient based on the 1% most up- and downregulated genes between AML and normal. Finally, using a hypergeometric test, we calculated the significance of the overlap between the subtype-specific CvN-defined, as well as previously reported CvC-defined, AK-AML signatures and the patient-specific signatures for upregulated genes (Figure 2E). Both types of subtype-specific AK-AML signatures displayed a strong overlap with the patient-specific signatures, but receiver-operator characteristic curves showed that the CvN-defined signatures outperformed their CvC-defined counterparts (Figure 2F; supplemental Figure 5A-D; supplemental Table 4). Significantly, this was also demonstrated in an independent data set[11] (supplemental Figure 5E-I).

In conclusion, our CvN method allows for the efficient classification of AK-AML subtypes with defined cytogenetic aberrations and outperforms the CvC method in unsupervised and supervised classification analysis.

The CvN method identifies genes and common transcriptional programs potentially linked to malignant transformation and maintenance of AK-AML

Although the CvN method allows for classification of AML subtypes, its main strength lies in its potential to identify changes in gene expression between cancer and normal. We therefore identified genes that exhibit deregulated expression ($|\log2\ FC| > 2$; $P < 10^{-5}$; see supplemental Table 1D-G for gene lists) between AML and normal for patients belonging to different AK-AML subtypes. This yielded complex patterns of gene expression changes between different AK-AML subtypes and their respective normal counterparts, which is also evident from the corresponding hierarchical clustering (supplemental Figure 6A-C). Of particular interest is the 1018 probe sets that are commonly deregulated ($|\log2\ FC| > 1$; $P < 10^{-5}$; supplemental Table 1C) in all the 4 AK-AML subtypes. Among the genes upregulated compared with their nearest normal counterpart, we find the RAS homology gene *RHOB* as well as the epigenetic regulators *JMJD3* and *BRD4* (supplemental Figure 7). Notably, the latter was recently identified as a therapeutic target in AML.[25]

The CvN method also identified high expression of EVI1 to be specifically associated with t(11q23) AML (supplemental Table 3). This concurs with the previously reported cooccurrence of t(11q23) lesions with high EVI1 expression, and it has been suggested that the latter correlates with the maturation stage of the leukemic blasts.[26] To test this, we separated the t(11q23) cohort into $EVI1^{high}$ and $EVI1^{low}$ patients (supplemental Figure 8A). We next assessed the average expression values of genes belonging to a novel stem cell signature (supplemental Methods) and find the $EVI1^{high}$ group to exhibit a higher score, suggesting that this subgroup of t(11q23) is more immature (supplemental Figure 8B). As high EVI1 expression correlates with adverse outcome,[26] we predict the $EVI1^{high}$ subgroup to have a poor overall survival (OS).

To further explore the transcriptional programs underlying the leukemic phenotypes, we used a hypergeometric test to compare the significance of the overlap between the patient-specific signatures, described earlier, and known gene expression signatures representing curated gene sets (C2), gene ontology gene sets (C5), and oncogenic signatures (C6) from the MSigDB database. Using this gene set overlap analysis, we identified the 200 best correlated MSigDB signatures for each AK-AML subclass ($P < 1e{-}5$; supplemental Table 5) and selected for signatures that, based on literature review of experimental design, represented bona fide correlates of normal cellular activities and responses (ie, cell cycle, signaling, and inflammatory and hypoxia response). We report this selection of signatures and their median log2 fold-change when compared with normal for patients derived from each AK-AML subtype (Figure 3A).

Strikingly, our analysis identified a predominance of transcriptional programs in all AK-AML subtypes, reflecting a low cell cycle activity combined with elevated activities of inflammatory response, hypoxia, and signaling. High cell cycle activity was most abundant among AK-AML patients with inv16/t(16;16) and t(11q23) and was low among t(15;17) and t(8;21) patients. Significantly, these findings demonstrate that our CvN method-based gene set overlap analysis allows for the identification of common sets of transcriptional programs shared by AK-AML patients of different genetic subclasses. Because of the lack of publicly available survival data for the AK-AML cohort, we were unfortunately not able to assess the relevance of differences in common transcriptional programs with respect to clinical outcome. It is, however, likely that individual patients of genetically defined AK-AML subclasses whose common transcriptional programs differ substantially (high vs low cell cycle activity, etc) also may exhibit differential survival.

Finally, to validate the functional relevance of one of the transcriptional programs identified by the CvN method, we performed cell cycle analysis on $CD34^+$ cells from t(8;21) AK-AML patients predicted to exhibit low cell cycle activity and $CD34^+$ populations of healthy subjects (Figure 3B-C). Indeed, this analysis demonstrated a low proliferation rate of leukemic compared with normal $CD34^+$ cells, which is consistent with the predicted low "transcriptional" cell cycle activity of t(8;21) AK-AML. Importantly, the GEP-based prediction of proliferation could also be extended to normal myeloid progenitors (Figure 3D).

Overall, our CvN-based analysis suggests that genetically and clinically diverse AK-AML subclasses share a common set of transcriptional programs that potentially represent abnormal activity of core cellular functions associated with transformation and maintenance of the leukemic phenotype.

Comparison of NK-AML patient samples to their nearest normal counterpart identifies novel subtypes

Having demonstrated the ability of the CvN method to correctly classify subtypes of AK-AML patients and identify common transcriptional programs, we next tested its potential on a data set of NK-AML patients, including survival rates.[2] NK-AML is associated with mutations in key hematopoietic and epigenetic regulators (*NPM1*, *CEBPA*, *FLT3*, *RUNX1*, *TET2*, *DNMT3A*, and others[27,28]), but only *NPM1* and *CEBPA* mutant AML constitutes distinct subtypes approved by the World Health Organization.[29] We reasoned that, similar to AK-AML, NK-AML would harbor distinct subtypes that could be identified through the CvN method.

As a first approach to estimate the number of potential subtypes in NK-AML patients, we used a similar strategy as that outlined earlier to perform hierarchical clustering on an data set of 218 NK-AML patients,[2] including information on survival and mutational status of *CEBPA*, *FLT3*, and *NPM1*. Visual inspection of this initial clustering analysis suggested the presence of 6 subtypes in the NK-AML data set (supplemental Figure 9). To further refine the analysis, we next performed K-means clustering to assign the patients to 6 clusters using variance-selected genes (Figure 4A-B,D-E; supplemental

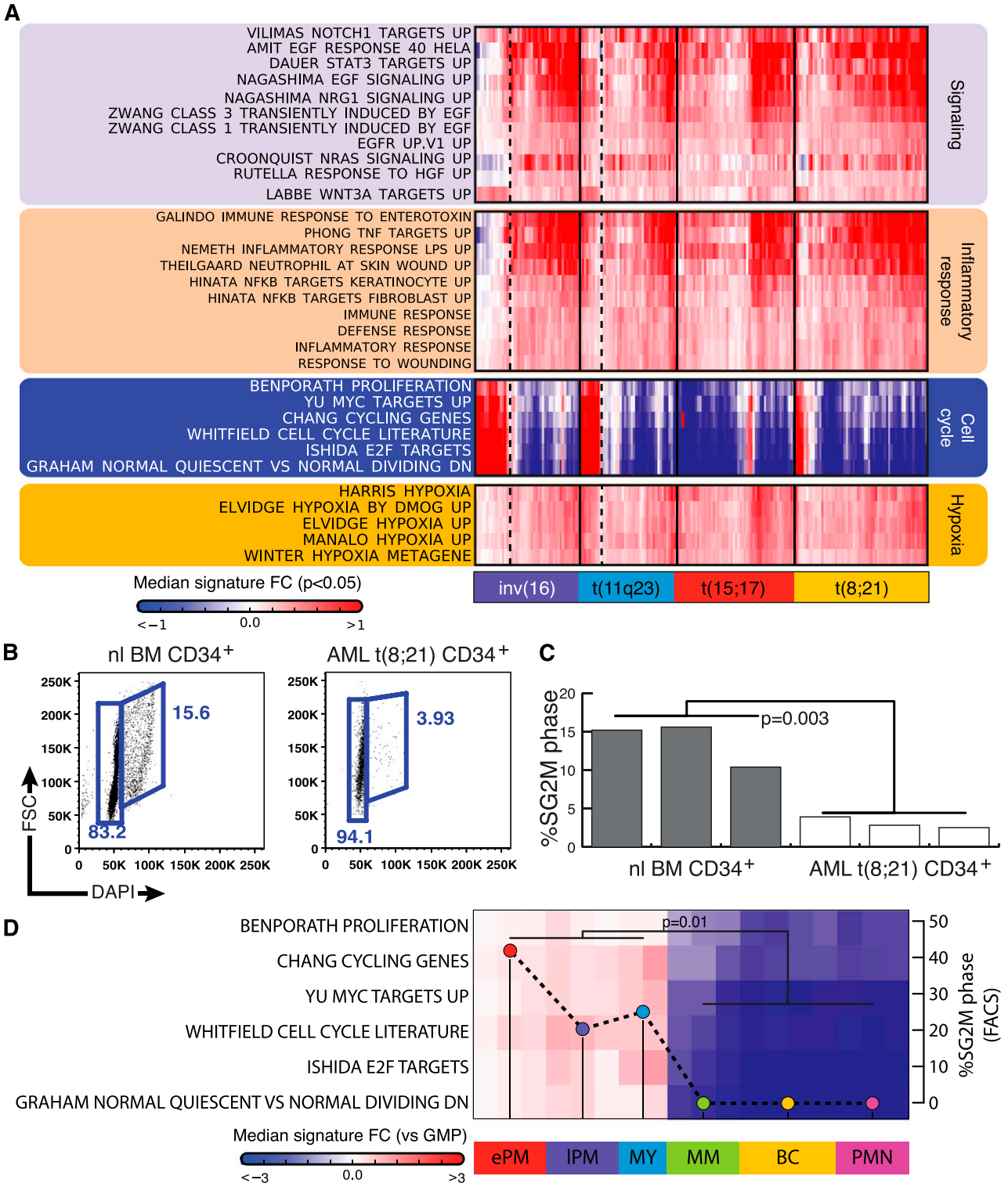

Figure 3. Identification of deregulated gene expression programs in AK-AML. (A) Median gene expression fold change of selected MsigDB gene signatures that overlap significantly ($P < 1e\text{-}5$, median, subclass-wise) with patient-specific AK-AML signatures. (B-C) Cell cycle analysis of $CD34^+$ cells from healthy subjects (n = 3) and t(8;21) AML patients (n = 3). (D) Median gene expression fold-change (*vs* normal GMPs) in cell cycle-related gene signatures for purified normal BM populations together with the experimentally determined cell cycle status (cell cycle profiles were presented in Mora-Jensen et al[48]). The correlation coefficient between "percentage of cells in SG2M" and the average median fold change for the 6 cell cycle signatures was $r^2=0.8$. The following populations are depicted: early promyelocytes (ePM), late promyelocytes (lPM), MY, MM, band cells (BC) and polymorphonuclear neutrophilic granulocytes (PMN).

Figure 10A-G). We note that *NPM1* and *FLT3* mutations did not segregate to any distinct cluster with either the CvN nor the CvC method, which likely reflects the high frequency of patients with combined *NPM1* and *FLT3* mutations in our NK-AML cohort. In contrast, patients with *CEBPA* mutations formed a distinct cluster with both methods (CvC cluster_5, Figure 4A; CvN cluster_3,

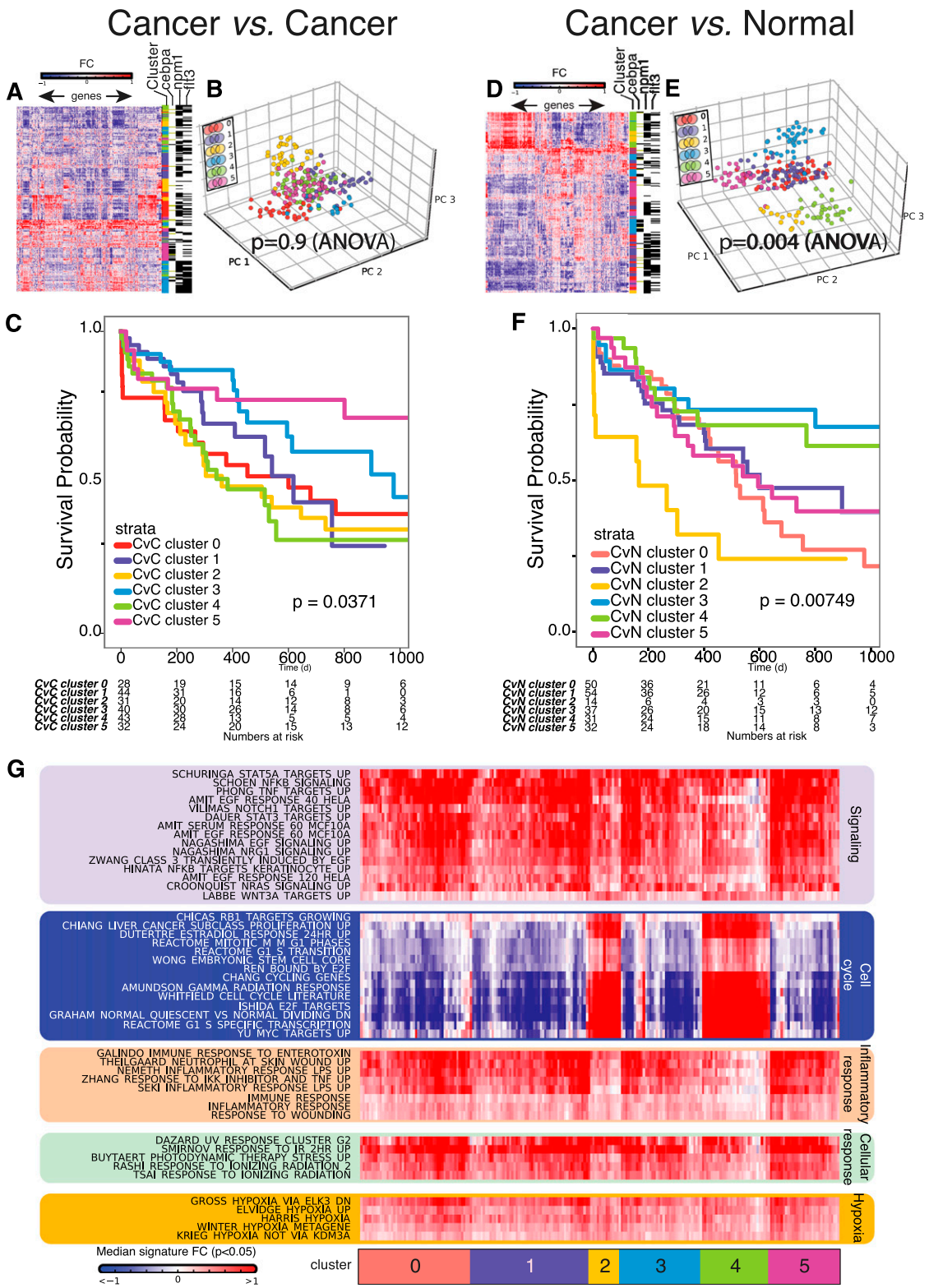

Figure 4. The CvN method improves classification of NK-AML patients. Side-by-side comparison of clustering performance of the CvC (A-C) and CvN (D-F) methods on a NK-AML data set (GSE15434). Heat maps (hierarchical clustering) of genes identified by the CvC method (A) and CvN methods (D), using a NK-AML patient data set. Differentially expressed genes identified by each method were selected by variance (1614 and 1383 probe sets in A and D, respectively) and rescaled gene wise. An initial hierarchical clustering was used to identify the optimal number of patient clusters (n = 6; supplemental Figure 9). This was followed by K-means clustering (K = 6), which distributed the samples into 6 patient clusters (color labeled). (B,E) 3-dimensional-PCA plots of the 6 K-means-derived patient clusters identified by the CvC (B) and CvN (E). (C,F) Kaplan-Meier plots depicting the OS curves for of the 6 NK-AML clusters assessed by (C) the CvC method and (F) the CvN method ($P = .04$ and $P = .007$, respectively, χ-square). (G) Median gene expression fold change of selected MsigDB gene signatures that overlap significantly ($P < 1e{-}5$, median, subclass-wise) with patient-specific NK-AML signatures.

Figure 4D). Although the data set does not contain information on the presence of bi- vs monoallelic *CEBPA* mutations, the published frequency of biallelic *CEBPA* AML (≈70%) suggest that these cluster contains the biallelic *CEBPA* AMLs (76% in CvC cluster_5; 71% in CvN cluster_3).[11]

To assess the relative performance of the 2 methods, we performed a silhouette analysis,[30] which demonstrated that both the CvN and CvC clusters were robust (with the former being slightly better [$P < 1e-5$, *t* test]; supplemental Figure 10H-I). However, when the intercluster and intracluster variance of the first 3 principal components were tested by ANOVA, only the CvN method yielded significant clusters ($P = .9$ vs $P = .004$).

To determine to what extent the CvC and CvN methods used different or overlapping genes to separate their respective clusters, we next merged the lists of genes selected by the 2 methods. Of the approximately 450 genes that were used for clustering by each method, 54% were shared. Importantly, when we analyzed the contribution of CvC-specific, shared, and CvN-specific genes to cluster formation, we found that the CvN-specific genes were better than the CvC-specific genes in separating the clusters (supplemental Figure 10B-G; CvN: $P = .001$; CvC: $P = .02$; ANOVA test on first 3 principal components, using the method-specific probe sets). Hence, the residual predictive power of the CvC genes in our analysis of NK-AML is primarily driven by a subgroup of genes, which is also selected by the CvN method, thus highlighting its excellent performance.

Overall, our cluster analysis demonstrates that the CvN method is capable of identifying potential subtypes of NK-AML patients with distinct patterns of aberrantly expressed genes.

The CvN-predicted NK-AML clusters display differential OS

We next assessed the potential of the CvN and CvC clusters to predict OS in NK-AML (Figure 4C,F). Interestingly, the clusters generated by the CvN method displayed distinct distributions of OS rates among NK-AML patients, suggesting that this method is capable of extracting prognostic relevant disease entities that are not defined by specific genetic lesions but, rather, by distinct gene expression programs representing surrogates of their leukemic phenotype. Of the 6 CvN clusters, we found cluster_2 to be associated with significantly worse outcome compared with the remaining 5 clusters (Figure 4F; supplemental Table 6). Moreover, multivariate Cox regression analysis identified cluster_2 as the strongest independent prognostic factor for OS in NK-AML patients, performing better than known risk factors such as *FLT3* mutations and age (Table 2). These findings were corroborated by a random forest analysis[31] that highlighted cluster_2 as the most important variable for the prediction of OS (supplemental Figure 11). Collectively, this demonstrates the ability of the CvN method to identify novel prognostic relevant subtypes of NK-AML patients.

To explore potentially disrupted core cellular functions associated with the leukemic phenotype and OS of the 6 NK-AML clusters, we again generated patient-specific signatures (as described earlier) and scored the significance of the overlap for each individual NK-AML signature against the gene expression signatures from the MSigDB database. We identified the 200 most significant and positively correlated MSigDB signatures for each NK-AML cluster ($P < 1e-5$; supplemental Table 5) and selected for signatures based on literature review, as described earlier. This analysis allowed us to identify NK-AML patients with high (cluster_2 cluster_4) vs low (cluster_0, cluster_1, cluster_3, cluster_5) cell cycle activity compared with their normal counterparts. Similar to AK-AML, the majority of NK-AML patients shared a common transcriptional program reflecting elevated activity of inflammatory response, hypoxia, and signaling activities independent of their mutation and cluster status (Figure 4G).

Table 2. Multivariate Cox regression analyses of the NK-AML data set

	OS	
Analysis	**Hazard ratio**	***P* value**
NK-AML patients		
Age	1.741	.018
Blast cell count	1.177	.264
CEBPA status	0.786	.190
Cluster_2	2.722	.004
FLT3	2.299	.002
Sex	0.962	.439
NPM1	0.432	.001
CvN method		
Cluster_1	0.80	.276
Custer_2	2.51	.017
Cluster_3	0.63	.144
Cluster_0	0.67	.154
Cluster_4	0.64	.161
Cluster_5	did not converge	

Multivariate Cox regression analyses illustrating the prognostic power of the NK-AML cluster_2. Cluster_2 was analyzed along with other clinical risk factors for the NK-AML patient data set. Cluster_2 was analyzed together with the other clusters identified by the CvN method. Cluster_5 fitting did not converge because of the few events in this cluster. None of the NK-AML clusters identified by the CvC method were significant in the multivariate Cox regression analysis.

Surprisingly, cluster_2 and cluster_4, which differed widely with respect to clinical outcome, shared high cell cycle activity and did not differ markedly with respect to activity of other common transcriptional programs. Consistently, we noted that cluster_2, with a poor outcome, expressed a very similar set of deregulated genes compared with cluster_4 with a favorable outcome ($|\log2\ FC| > 2$; $P < 10^{-5}$ vs normal cells; Figure 4D; supplemental Tables 1H and 4). We therefore hypothesized that the few differentially expressed genes between these clusters would be highly enriched in genes that account for chemotherapy resistance, and thus could predict OS for the entire NK-AML patient data set. To test this, we identified differentially expressed genes between cluster_2 and cluster_4 ($|\log2\ FC| > 2$; $P < .05$) and determined the significance by which they could predict differences in OS in the upper and lower fold-change quartiles, using the entire NK-AML patient data set (supplemental Table 1I). We next used these genes to build a poor-outcome signature, a good-outcome signature, and a combined survival signature, which were all able to efficiently allocate the entire NK-AML data set, as well as 2 independent data sets,[13,15] into patients with good and poor outcome (Figure 5A-E). Importantly, the genes in these signatures are predicted to be enriched for genes directly involved in disease etiology, including resistance to chemotherapy (supplemental Table 7). On a final note, we found that a previously reported hematopoietic stem cell signature[32] was unable to predict survival in the NK-AML patient data set[15] (Figure 5F).

Collectively, our analyses demonstrate that the CvN method is able to stratify NK-AML patients into known subtypes (cluster_3 with *CEBPA* mutations), stratify patients into new subtypes exhibiting differential OS, and extract a set common transcriptional programs that likely represent disrupted core cellular functions underlying the leukemic phenotype (see supplemental Table 5 for additional data). In addition, our analysis might imply that increased chemotherapy resistance in NK-AML clusters with poor vs favorable

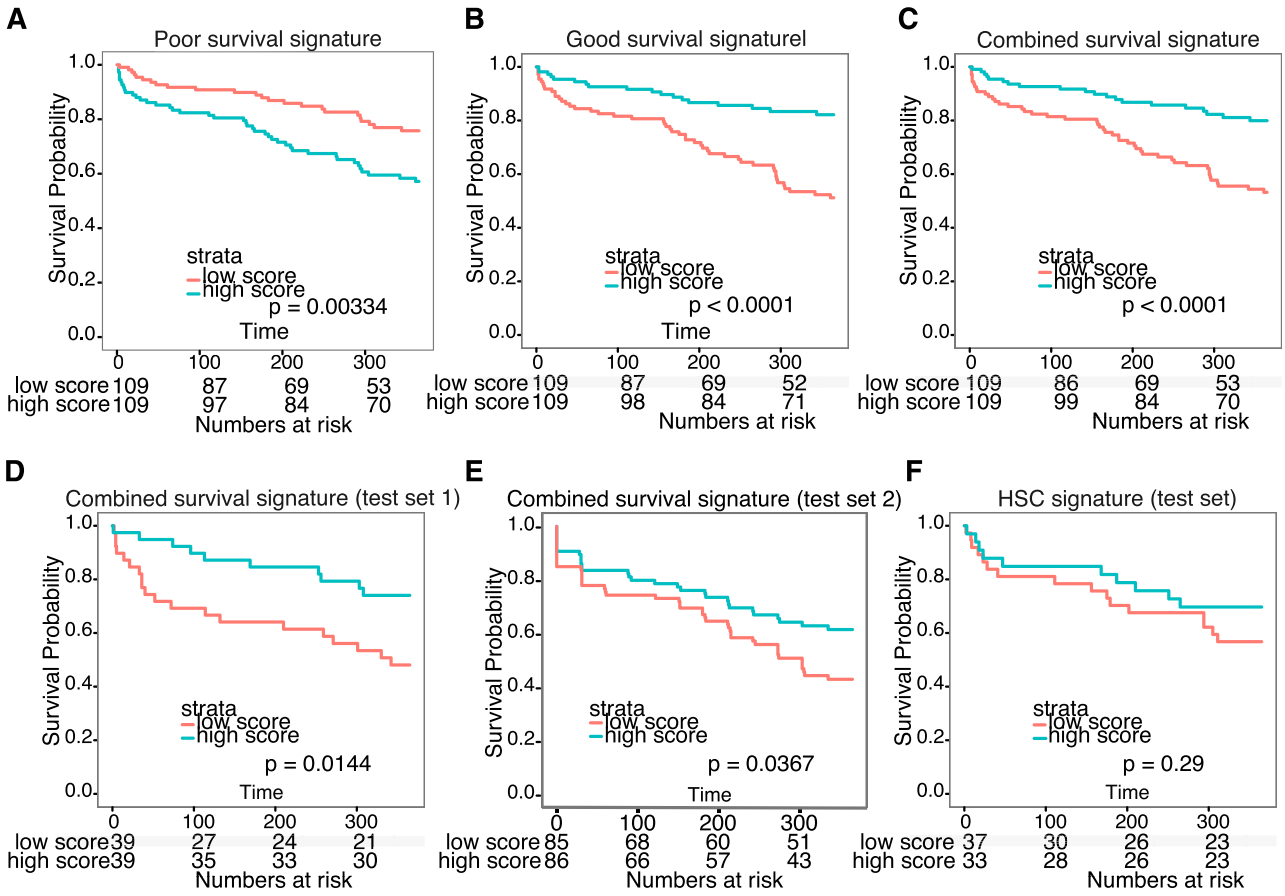

Figure 5. Survival signature predicts survival of patients with NK-AML. (A-C) Survival analysis based on 3 survival signatures derived from genes differentially expressed in patient cluster_2 and cluster_4. The effect of the expression of individual probe sets on survival was tested by dividing the entire data set into low- and high-scoring samples (median). Probe sets associated with poor and good OS ($P < .05$, moderated t test) and the ability to separate the data set ($P < .05$, log-rank test) were used to generate a poor outcome signature (A), good outcome signature (B), and combined survival signature (C). (D-E) Testing of the combined survival signature on 2 independent NK-AML patient data sets.[13,15] (F) Testing of a previously published HSC signature[32] revealed its inability to predict survival in the NK-AML patient data set in D.

outcome is primarily driven by a limited number of highly prognostically relevant genes rather than higher or lower activity of some of the common transcriptional programs.

Discussion

GEP has the potential to yield fundamental insights into the transcriptional programs of cancer cells and has thus been used for more than a decade to probe tumor phenotypes. However, with few exceptions, these analyses have all compared cancer to cancer, with the obvious risk that differences in cell type and developmental stage may render the identification of truly malignant gene expression programs impossible. Here we present a simple method, referred to as the CvN method, that allows us to identify the nearest normal counterpart for individual AML patient samples and calculate gene expression differences between AML and normal. Our method performed extremely well in classification of AK-AML and identified gene expression programs associated with distinct AK-AML subtypes. Moreover, we were able to clearly separate an NK-AML patient data set into several clusters according to transcriptional differences between individual AML patient samples and their closest normal counterpart. These clusters were associated with distinct OS and were predictive in multivariate analysis, highlighting their biological relevance.

Recent epidemiologic and clinical studies have demonstrated a higher incidence and aggressiveness of cancer in patients with diabetes and in patients with inflammatory and autoimmune diseases.[33-35] Consistently, treatment with metformin reduces cancer in patients with diabetes, and inflammatory ligands are elevated and promote maintenance and proliferation of malignant cells in different cancer entities.[33,36-39] In addition, NF-κB, the key transcriptional regulator of inflammatory response, was demonstrated to be constitutively activated in various types of cancers, including AML, and to play an important role in malignant transformation in mouse models.[39,40-43] Finally, the ability of various solid cancer cells to adapt to hypoxia and switch metabolism from oxidative phosphorylation toward glycolysis has emerged as a novel hallmark of cancer that defines more aggressive cancer phenotypes.[44]

Consistent with these reports, a previous study[45] identified a cancer signature of malignant transformation that is not only shared by various types of cancers but also overlaps significantly with gene expression signatures of chronic inflammatory conditions (colitis ulcerosa, rheumatoid arthritis, systemic lupus erythematosus, Crohn's disease) and metabolic diseases (diabetes, obesity, hypercholesterolemia, atherosclerosis, cardiomyopathy). On the basis of their findings, the authors argued that physiological and/genetic disruption of core biological pathways maintaining normal cell functions generates a gene expression program that is common to a diverse set of human diseases.[45]

In line with these findings, our analysis demonstrated a significant overlap of the "common cancer signature" and AK-AML and NK-AML signatures generated by the CvN method (supplemental Table 5; HIRSCH_CELLULAR_TRANSFORMATION_SIGNATURE_UP). Strikingly, our analysis also unraveled common transcriptional programs among all AML patients that are associated with elevated signaling activity, inflammatory response, and hypoxia. Indeed, these programs might reflect a disruption of normal core cellular functions that are shared by most AML patients despite their otherwise profound clinical and genetic heterogeneity.

Significantly, our gene set overlap analysis allowed us to discriminate AML patients with a high vs low cell cycle activity compared with their normal counterpart. Whereas the majority of AK- and NK-AML patients demonstrated a low cell cycle activity combined with elevated activities of inflammatory response, hypoxia, and signaling, a minor number of patients demonstrated a program of high cell cycle activity. The latter included a significant number of AK-AML patients with inv16/t(16;16) and t(11q23) as well as all cluster_2 and cluster_4 NK-AML patients. Surprisingly, cluster_2 and cluster_4 demonstrated poor and favorable outcomes, respectively, despite comparable high cell cycle activity combined with a similar activity of inflammatory response, hypoxia, and signaling. Consistently, they shared a high number of aberrantly expressed genes compared with normal but also demonstrated a limited number of differentially expressed genes that formed the basis for a powerful NK-AML survival signature. These findings suggest that resistance to chemotherapy in NK-AML patients with poor vs favorable outcome is primarily driven by a minor number of prognostic relevant genes and not by the differential activity of common transcriptional programs. Importantly, as these common transcriptional programs likely represent disrupted core cellular functions, some of them may be relevant for future targeting.

The CvN method may, in principle, be improved by several means. As an example, the data used in the present work originate from AML data sets derived from bulk tumor material. Hence, an obvious improvement of the precision of the CvN method would be to perform the analysis on purified AML subpopulations and compare those with their respective normal counterparts. Furthermore, our method is dependent on the availability of GEPs from normal cells to construct a gene expression landscape of the hematopoietic hierarchy onto which we can map AML samples. Obviously, the precision of the mapping and the subsequent deduction of gene expression changes between cancer cells, and their corresponding normal counterpart is dependent on the number and quality of normal reference populations. Given the high density of functionally defined intermediate HPCs on the path from HSCs to mature blood cells within the hematopoietic hierarchy, its associated malignancies are ideally suited for the CvN method. However, by combining multiparameter cell sorting with highly innovative methods for the analysis of flow cytometry data, such as the recently published Cyto-Spanning Tree Progression of Density Normalized Events (CytoSPADE) method, it should be possible to isolate novel intermediate HPCs for subsequent GEP, thereby further refining the resolution of the gene expression landscape of the normal hematopoietic hierarchy.[46,47] This will in turn improve the extent to which transcriptional changes between normal and cancer cells can be detected. Importantly, as no functionally characterization is required, implementation of SPADE or similar protocols may be used to isolate a suitable number of stem/progenitor cells from other organs, thereby making the CvN method amenable for analysis of solid tumors. Thus, our approach has the potential to be widely applicable to a substantial number of cancer types and promises to expand the clinical use of GEP.

Acknowledgments

This work was supported by grants from the Danish Council for Strategic Research (09-065157, 10-092798), the Danish Cancer Society (R2-Rp1425), the NovoNordisk Foundation (11628, R168-A14079), and the Lundbeck Foundation (R34-A3620) and through a center grant from the Novo Nordisk Foundation Section for Stem Cell Biology in Human Disease. This work is based on the joint research activities under the framework of the European Program for Cooperation in Science and Technology (Action BM0801, WG1). L.B. was supported in part by the Deutsche Forschungsgemeinschaft (Heisenberg-Stipendium BU 1339/3-1). K.T.-M. is supported by a clinical research fellowship from the Novo Nordisk Foundation (R191-A15986). O.W. acknowledges funding from the Novo Nordisk Foundation (05-04-2005).

Authorship

Contribution: N.R., N.B., A. Krogh, O.W., K.T.-M., and B.T.P. conceived and designed the study; N.B. and H.M.J. provided BM samples; A. Kohlmann, C.T., L.B., J.J., and H.M.-J. collected and assembled the data; N.R. produced the figures; N.R., F.O.B., K.T.-M., and B.T.P. analyzed and interpreted the data; and N.R., B.T.P., K.T.-M., O.W., and F.O.B. wrote the manuscript.

Conflict-of-interest disclosure: A. Kohlman is employed by the Munich Leukemia Laboratory. The remaining authors declare no competing financial interests.

Correspondence: Bo Porse, Finsen Laboratory/Rigshospitalet, University of Copenhagen, Ole Maaløesvej 5, 2200 Copenhagen N, Denmark; e-mail: bo.porse@finsenlab.dk.

References

1. Golub TR, Slonim DK, Tamayo P, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science.* 1999;286(5439):531-537.
2. Kohlmann A, Bullinger L, Thiede C, et al. Gene expression profiling in AML with normal karyotype can predict mutations for molecular markers and allows novel insights into perturbed biological pathways. *Leukemia.* 2010;24(6): 1216-1220.
3. Valk PJM, Verhaak RGW, Beijen MA, et al. Prognostically useful gene-expression profiles in acute myeloid leukemia. *N Engl J Med.* 2004; 350(16):1617-1628.
4. Bullinger L, Döhner 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(16): 1605-1616.
5. Raponi M, Lancet JE, Fan H, et al. A 2-gene classifier for predicting response to the farnesyltransferase inhibitor tipifarnib in acute myeloid leukemia. *Blood.* 2008;111(5): 2589-2596.
6. Ebert BL, Galili N, Tamayo P, et al. An erythroid differentiation signature predicts response to lenalidomide in myelodysplastic syndrome. *PLoS Med.* 2008;5(2):e35.
7. Theilgaard-Mönch K, Boultwood J, Ferrari S, et al. Gene expression profiling in MDS and AML: potential and future avenues. *Leukemia.* 2011; 25(6):909-920.
8. 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.

9. Haferlach T, Kohlmann A, Wieczorek L, et al. Clinical utility of microarray-based gene expression profiling in the diagnosis and subclassification of leukemia: report from the International Microarray Innovations in Leukemia Study Group. *J Clin Oncol.* 2010;28(15): 2529-2537.
10. 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(5):802-807.
11. Wouters BJ, Löwenberg B, Erpelinck-Verschueren CAJ, van Putten WLJ, Valk PJM, Delwel R. Double CEBPA mutations, but not single CEBPA mutations, define a subgroup of acute myeloid leukemia with a distinctive gene expression profile that is uniquely associated with a favorable outcome. *Blood.* 2009;113(13): 3088-3091.
12. Klein H-U, Ruckert C, Kohlmann A, et al. Quantitative comparison of microarray experiments with published leukemia related gene expression signatures. *BMC Bioinformatics.* 2009; 10(1):422.
13. 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.
14. Tomasson MH, Xiang Z, Walgren R, et al. Somatic mutations and germline sequence variants in the expressed tyrosine kinase genes of patients with de novo acute myeloid leukemia. *Blood.* 2008;111(9):4797-4808.
15. Metzeler KH, Hummel M, Bloomfield CD, et al; Cancer and Leukemia Group B; German AML Cooperative Group. An 86-probe-set gene-expression signature predicts survival in cytogenetically normal acute myeloid leukemia. *Blood.* 2008;112(10):4193-4201.
16. Gautier L, Cope L, Bolstad BM, Irizarry RA. affy—analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics.* 2004;20(3):307-315.
17. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics.* 2007;8(1): 118-127.
18. de Jonge HJM, Valk PJM, Veeger NJGM, et al. High VEGFC expression is associated with unique gene expression profiles and predicts adverse prognosis in pediatric and adult acute myeloid leukemia. *Blood.* 2010;116(10): 1747-1754.
19. Majeti R, Becker MW, Tian Q, et al. Dysregulated gene expression networks in human acute myelogenous leukemia stem cells. *Proc Natl Acad Sci U S A.* 2009;106(9):3396-3401.
20. Andersson A, Edén P, Olofsson T, Fioretos T. Gene expression signatures in childhood acute leukemias are largely unique and distinct from those of normal tissues and other malignancies. *BMC Med Genomics.* 2010;3(1):6.
21. Hu X, Chung AY, Wu I, et al. Integrated regulation of Toll-like receptor responses by Notch and interferon-γ pathways. *Immunity.* 2008;29(5): 691-703.
22. Wildenberg ME, van Helden-Meeuwsen CG, van de Merwe JP, Drexhage HA, Versnel MA. Systemic increase in type I interferon activity in Sjögren's syndrome: a putative role for plasmacytoid dendritic cells. *Eur J Immunol.* 2008; 38(7):2024-2033.
23. de Hoon MJL, Imoto S, Nolan J, Miyano S. Open source clustering software. *Bioinformatics.* 2004; 20(9):1453-1454.
24. Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments [published ahead of print February 12, 2004]. *Stat Appl Genet Mol Biol.*
25. Zuber J, Shi J, Wang E, et al. RNAi screen identifies Brd4 as a therapeutic target in acute myeloid leukaemia. *Nature.* 2011;478(7370): 524-528.
26. Lugthart S, van Drunen E, van Norden Y, et al. High EVI1 levels predict adverse outcome in acute myeloid leukemia: prevalence of EVI1 overexpression and chromosome 3q26 abnormalities underestimated. *Blood.* 2008; 111(8):4329-4337.
27. Bacher U, Schnittger S, Haferlach T. Molecular genetics in acute myeloid leukemia. *Curr Opin Oncol.* 2010;22(6):646-655.
28. Shih AH, Abdel-Wahab O, Patel JP, Levine RL. The role of mutations in epigenetic regulators in myeloid malignancies. *Nat Rev Cancer.* 2012; 12(9):599-612.
29. Döhner H, Estey EH, Amadori S, et al; European LeukemiaNet. Diagnosis and management of acute myeloid leukemia in adults: recommendations from an international expert panel, on behalf of the European LeukemiaNet. *Blood.* 2010;115(3):453-474.
30. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J Comput Appl Math.* 1987;20(1):53-65.
31. Ishwaran H, Kogalur UB, Chen X, Minn AJ. Random survival forests for high-dimensional data. *Stat Anal Data Mining.* 2011;4(1):115-132.
32. 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.
33. Pierce BL, Ballard-Barbash R, Bernstein L, et al. Elevated biomarkers of inflammation are associated with reduced survival among breast cancer patients. *J Clin Oncol.* 2009;27(21): 3437-3444.
34. Mantovani A, Allavena P, Sica A, Balkwill F. Cancer-related inflammation. *Nature.* 2008; 454(7203):436-444.
35. Calle EE, Kaaks R. Overweight, obesity and cancer: epidemiological evidence and proposed mechanisms. *Nat Rev Cancer.* 2004;4(8): 579-591.
36. Balkwill F, Mantovani A. Inflammation and cancer: back to Virchow? *Lancet.* 2001;357(9255): 539-545.
37. Karin M. Nuclear factor-kappaB in cancer development and progression. *Nature.* 2006; 441(7092):431-436.
38. De Marzo AM, Platz EA, Sutcliffe S, et al. Inflammation in prostate carcinogenesis. *Nat Rev Cancer.* 2007;7(4):256-269.
39. Naugler WE, Karin MNF. NF-kappaB and cancer-identifying targets and mechanisms. *Curr Opin Genet Dev.* 2008;18(1):19-26.
40. Luedde T, Beraza N, Kotsikoris V, et al. Deletion of NEMO/IKKgamma in liver parenchymal cells causes steatohepatitis and hepatocellular carcinoma. *Cancer Cell.* 2007;11(2):119-132.
41. Sakurai T, He G, Matsuzawa A, et al. Hepatocyte necrosis induced by oxidative stress and IL-1 α release mediate carcinogen-induced compensatory proliferation and liver tumorigenesis. *Cancer Cell.* 2008;14(2): 156-165.
42. Hassane DC, Guzman ML, Corbett C, et al. Discovery of agents that eradicate leukemia stem cells using an in silico screen of public gene expression data. *Blood.* 2008;111(12):5654-5662.
43. Marstrand TT, Borup R, Willer A, et al. A conceptual framework for the identification of candidate drugs and drug targets in acute promyelocytic leukemia. *Leukemia.* 2010;24(7): 1265-1275.
44. Harris AL. Hypoxia—a key regulatory factor in tumour growth. *Nat Rev Cancer.* 2002;2(1):38-47.
45. Hirsch HA, Iliopoulos D, Joshi A, et al. A transcriptional signature and common gene networks link cancer with lipid metabolism and diverse human diseases. *Cancer Cell.* 2010; 17(4):348-361.
46. Qiu P, Simonds EF, Bendall SC, et al. Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. *Nat Biotechnol.* 2011;29(10):886-891.
47. Bendall SC, Simonds EF, Qiu P, et al. Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. *Science.* 2011;332(6030):687-696.
48. Mora-Jensen H, Jendholm J, Fossum A, Porse B, Borregaard N, Theilgaard-Mönch K. Technical advance: immunophenotypical characterization of human neutrophil differentiation. *J Leukoc Biol.* 2011;90(3):629-634.