Supplementary Data for:

Specific molecular signatures underlie response to Decitabine in CMML

Kristen Meldi[1*], Tingting Qin[1*], Francesca Buchi[2], Nathalie Droin[3], Jason Sotzen[1], Jean-Baptiste Micol3,[4], Dorothée Selimoglu-Buet[3], Erico Masala[2], Bernardino Allione[5,6], Daniela Gioia[6,7], Antonella Poloni[6,8], Monia Lunghi[6,9], Eric Solary[3], Omar Abdel-Wahab[4], Valeria Santini[2,6*], Maria E. Figueroa[1*]

1. University of Michigan Medical School, Dept. Of Pathology, Ann Arbor, MI 48109-2200, USA
2. University of Florence AOU Careggi, Hematology, Florence, Italy.
3. Inserm U1009, Gustave Roussy Cancer Center, Villejuif, France.
4. Memorial Sloan Kettering Cancer Center, Human Oncology and Pathogenesis Program and Leukemia Service, New York, 10021, USA
5. Hematology, Ospedale S. Giovanni Battista Molinette,Torino, Italy
6. Fondazione Italiana per le Sindromi Mielodisplastiche, Italy
7. AO SS Antonio e Biagio e C. Arrigo, Alessandria, Italy
8. Hematology, Università Politecnica delle Marche, Ancona, Italy
9. Department of Hematology, Amedeo Avogadro University, Novara, Italy

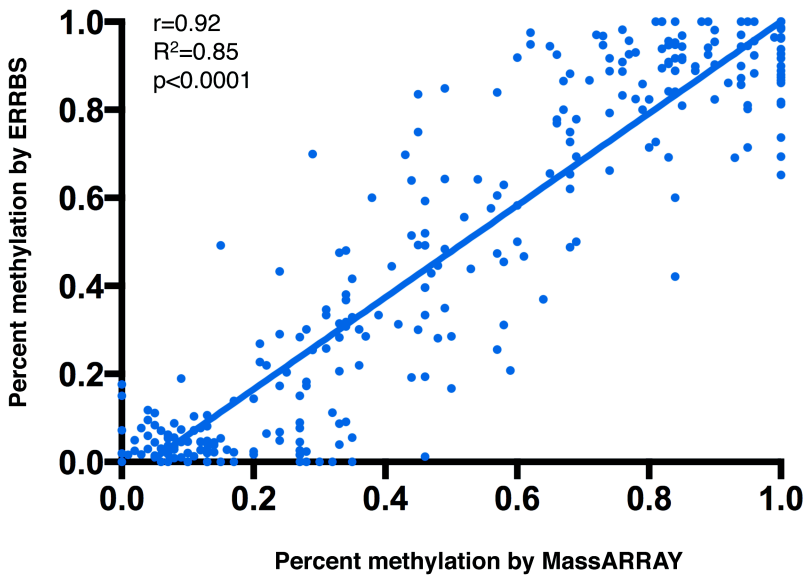
* These authors contributed equally

Key words: DNA methylation, Bisulfite sequencing, ERRBS, MDS, CMML, Decitabine, Methyltransferase inhibitors.

Corresponding authors:

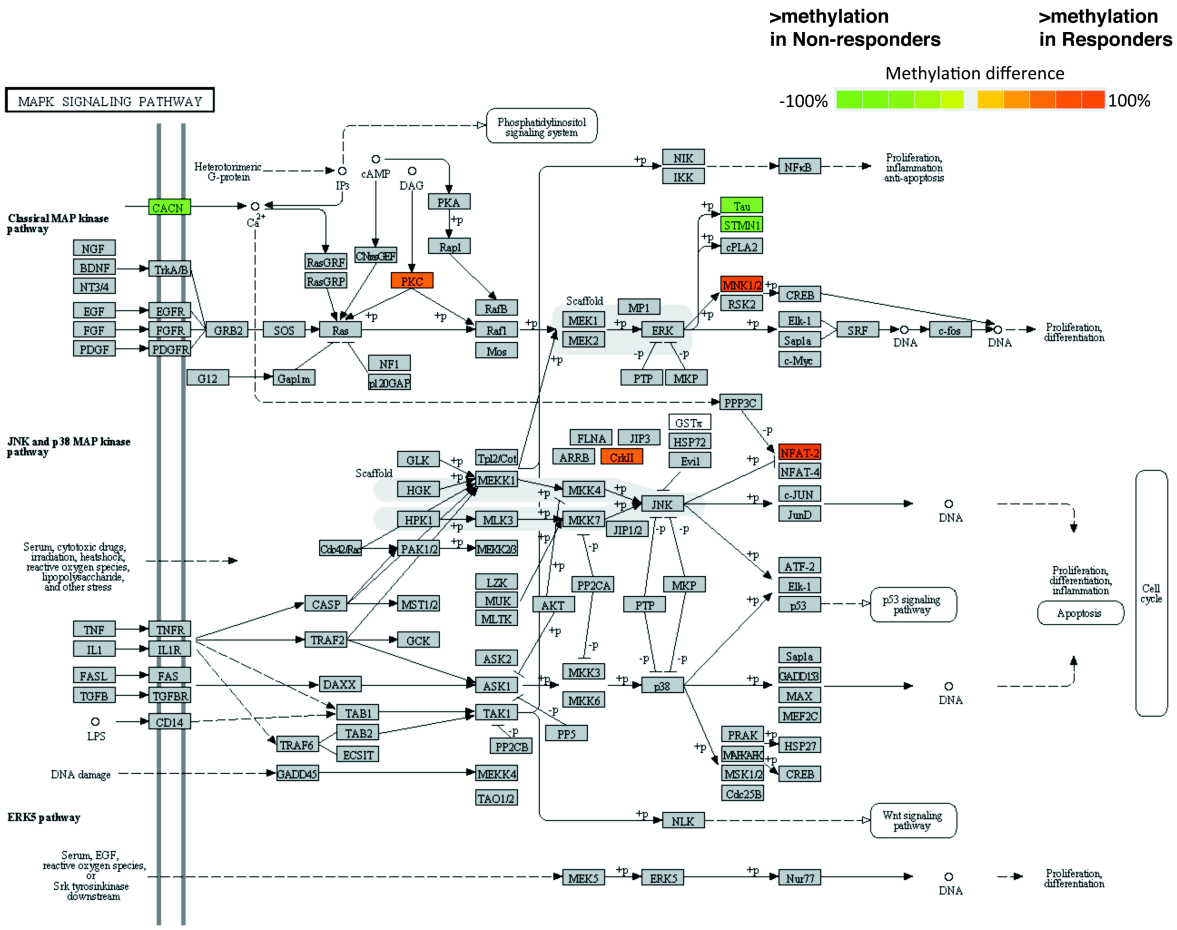
Maria E. Figueroa, M.D.
Assistant Professor,
Department of Pathology,
University of Michigan Medical School,
109 Zina Pitcher Pl, BSRB 2019,
Ann Arbor, MI 48176
Phone: +1 (734) 763-1865
Fax: +1 (734) 763-2162
Email: marfigue@med.umich.edu

Valeria Santini, M.D.
University of Florence AOU Careggi,
Hematology,
Florence, Italy
Phone: +39 (055) 7947296
Fax: +39 (055) 7947343
Email: santini@unifi.it

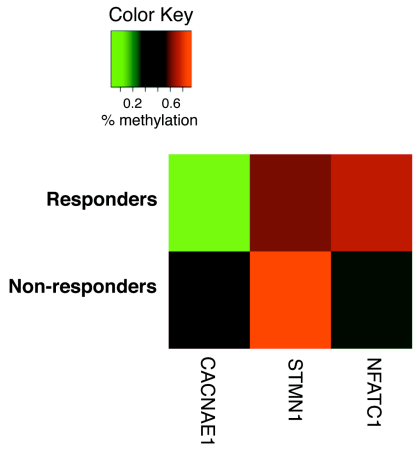


Supplementary Figure 1. Technical validation of ERRBS performance accuracy by MassARRAY EpiTYPER. DNA from 8 randomly-selected patients was used for an independent round of bisulfite treatment followed by site-specific PCR amplification and processed for MALDI-TOF analysis as previously described (1). The primers and corresponding amplicon sequences are listed in **Supplementary Table 6**. Forty-seven CpG sites were covered by both MassARRAY and ERRBS, and the methylation status at most, but not all, CpGs was available in all 8 patient samples by both methods. This resulted in a total of 292 CpGs analyzed by both methods with r=0.92, R^2=0.85, p<0.0001.

A

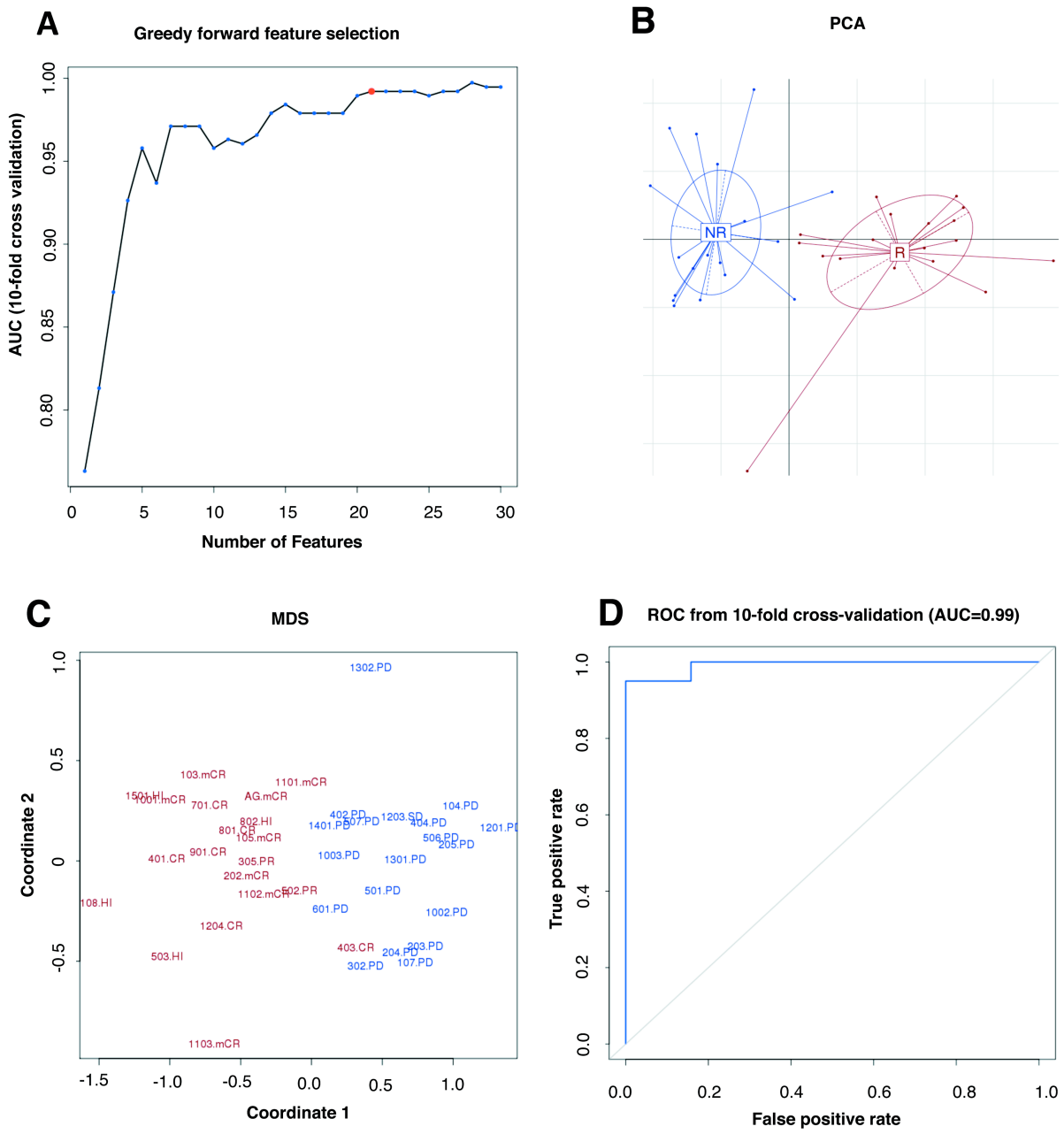


B



Supplementary Figure 2. Pathway analysis for DMR-annoted genes. A. Kegg pathway analysis revealed the enrichment of DMR at genes involved in MAPK

signaling**. B.** Heatmap of 3 MAPK-annotated DMRs in responders and non-responders as determined by MassARRAY EpiTYPER.



Supplementary Figure 3. Methylation status at 21 genomic regions can be used to predict Decitabine (DAC) response at diagnosis. A. Forward feature selection identified 21 CpG regions with the highest performance in classifying the samples. **B.** Principle components analysis (PCA) using the 21 features **C.** Multi-dimensional scaling

(MDS) analysis of patients using the 21 genomic features. Patients are labeled with their specific type of response; CR: complete response; HI: hematological improvement; mCR: marrow complete response; PD: progressive disease; PR: partial response; SD: stable disease. **D.** Receiver Operating Characteristic (ROC) after 10-fold cross-validation reflects the performance of the SVM classifier with the 21 selected predictors.

A

30 training, 9 testing
Accuracy = 100%

Patient ID	Original label	Prediction
1002	NR	NR
0107	NR	NR
0402	NR	NR
1401	NR	NR
1203	NR	NR
0401	R	R
0302	NR	NR
0701	R	R
1204	R	R

25 training, 14 testing
Accuracy = 92.86%

Patient ID	Original labels	Prediction
1002	NR	NR
0104	NR	NR
0402	NR	NR
0403	R	NR
0502	R	R
0103	R	R
0105	R	R
1203	NR	NR
0507	NR	NR
0401	R	R
0802	R	R
1102	R	R
0108	R	R
1103	R	R

B

6 features, 28 GFM patients
Accuracy = 71%

Patient ID	Original label	Prediction
2A	R	R
3A	R	R
4A	R	R
5A	NR	NR
6A	R	R
7A	NR	R
9A	NR	NR
10A	R	R
11A	NR	R
12A	R	NR
13A	R	R
15A	NR	NR
16A	NR	R
18A	NR	NR
22A	NR	R
23A	R	R
25A	R	R
26A	R	R
27A	NR	NR
30A	NR	NR
32A	NR	R
33A	R	R
34A	NR	R
35A	NR	R
36A	NR	NR
37A	NR	NR
39A	NR	NR
41A	R	R

14 features, 19 GFM patients
Accuracy = 79%

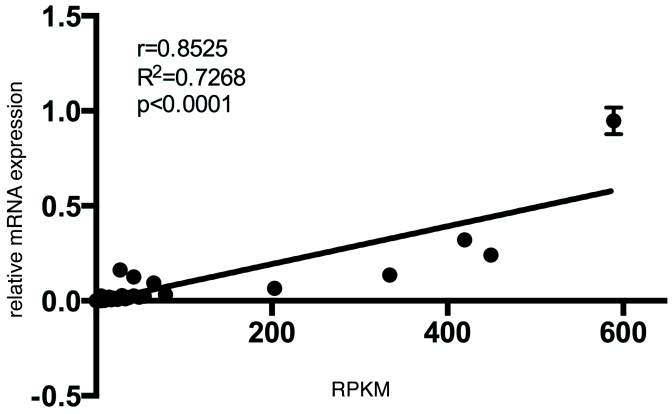
Patient ID	Original label	Prediction
2A	R	NR
4A	R	R
5A	NR	NR
6A	R	R
9A	NR	NR
11A	NR	NR
12A	R	NR
16A	NR	NR
18A	NR	NR
26A	R	R
27A	NR	NR
30A	NR	NR
32A	NR	R
33A	R	NR
35A	NR	NR
36A	NR	NR
37A	NR	NR
39A	NR	NR
41A	R	R

16 features, 15 GFM patients
Accuracy = 87%

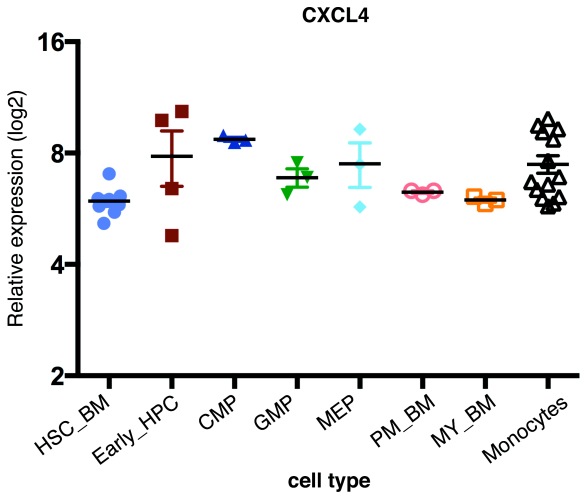
Patient ID	Original label	Prediction
2A	R	R
4A	R	R
5A	NR	NR
6A	R	R
9A	NR	NR
11A	NR	NR
12A	R	NR
18A	NR	NR
26A	R	R
27A	NR	NR
30A	NR	NR
32A	NR	R
36A	NR	NR
37A	NR	NR
39A	NR	NR

Supplementary Figure 4: Epigenetic classifier accurately predicts patient response to Decitabine. A. Predictions of the 21-feature SVM classifier on two randomly selected training sets of the FISM cohort, which were trained by the corresponding exclusive testing sets. **B.** Validations of the SVM classifier on the independent GFM cohort using 6, 14, and 16 out of the 21 features selected from FISM cohort (*left, middle*, and *right panels*, respectively).

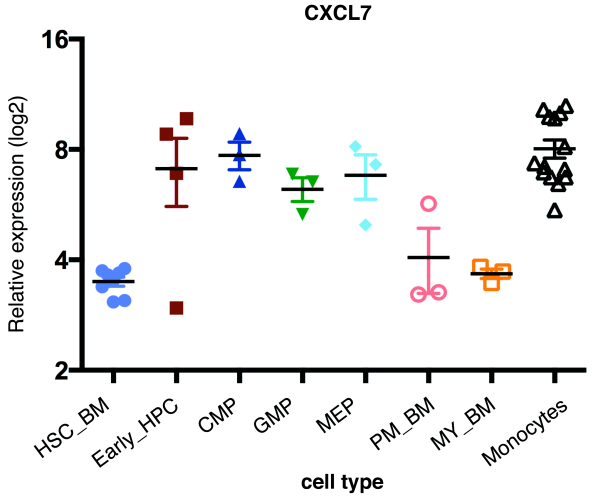
A


r=0.8525
R2=0.7268
p<0.0001
relative mRNA expression
1.5
1.0
0.5
0.0
-0.5
200
400
600
RPKM

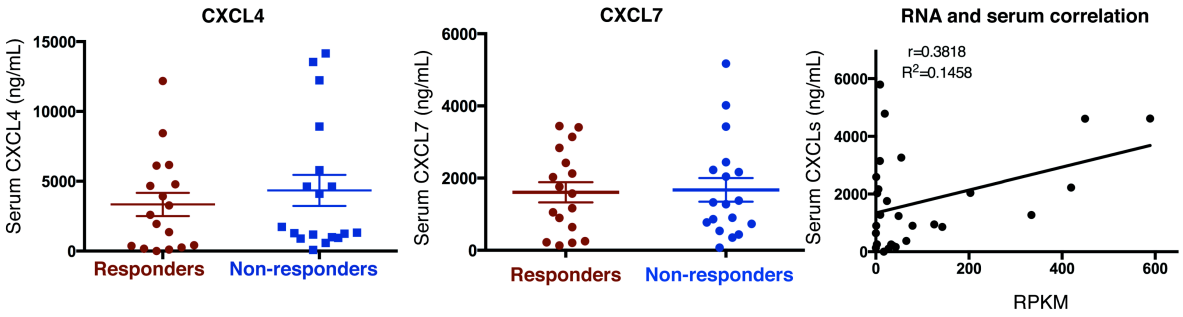
B


CXCL4
Relative expression (log2)
16
8
4
2
HSC_BM
Early_HPC
CMP
GMP
MEP
PM_BM
MY_BM
Monocytes
cell type

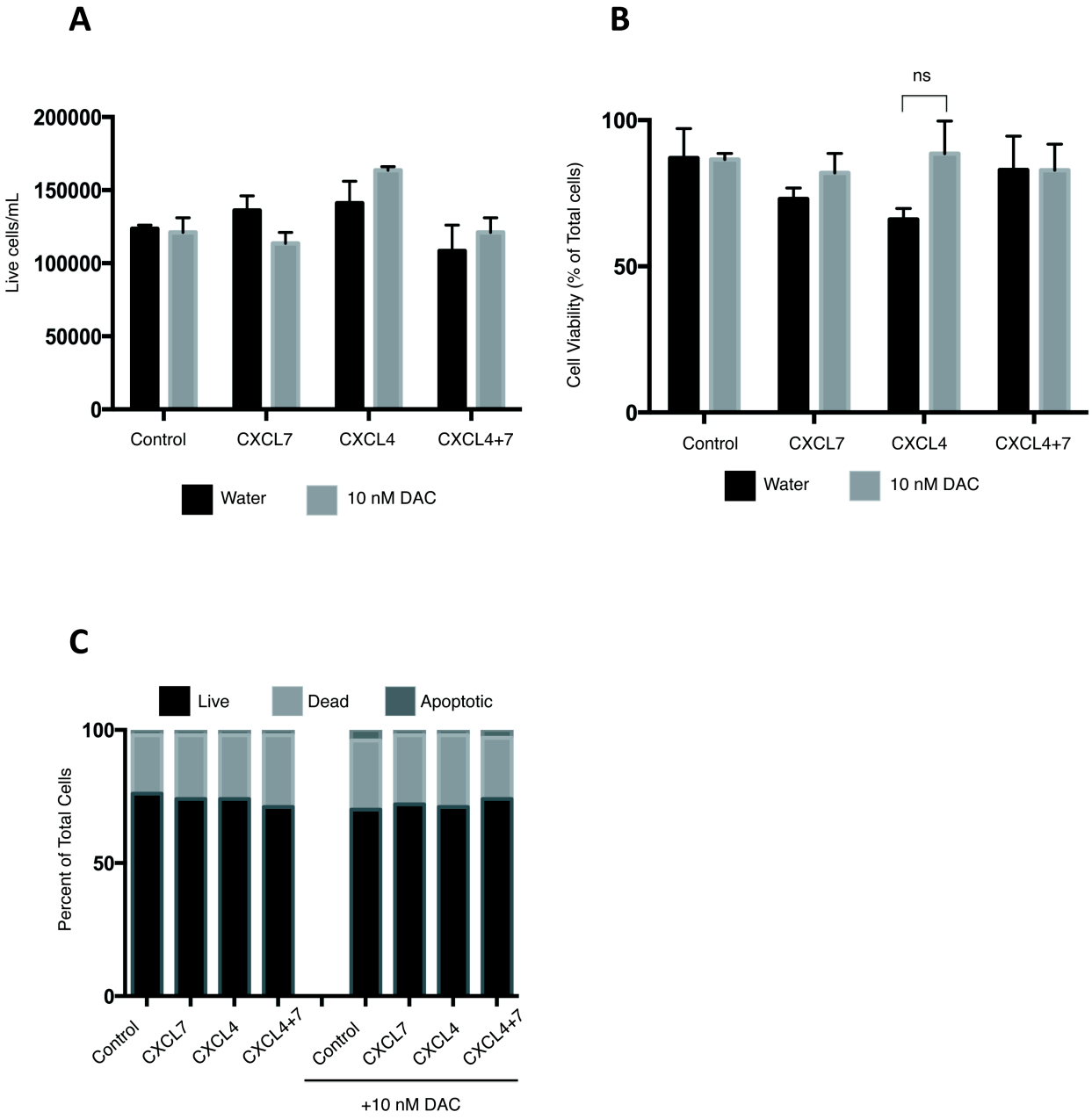
C


CXCL7
Relative expression (log2)
16
8
4
2
HSC_BM
Early_HPC
CMP
GMP
MEP
PM_BM
MY_BM
Monocytes
cell type

Supplementary Figure 5. CXCL4 and 7 are overexpressed in patients resistant to Decitabine (DAC). A. RNA-seq validation by qRT-PCR**.** Quantitative reverse transcription PCR (qRT-PCR) was used to technically validate the RNA-seq analysis on 13 out of 14 patients and indicates strong correlation between the RNA-seq results (reads per kilobase per million mapped reads, RPKM) and qRT-PCR analysis (relative mRNA to RPL19) for CXCL4, CXCL7, and ITGβ3. **B-C.** CXCL4 **(B)** and CXCL7 **(C)** are expressed in multiple cell types in the hematopoietic system. Expression data from Hemaexplorer (2, 3) in the normal human hematopoetic system indicate that expression of CXCL4 and CXCL7 is not limited to megakaryocytes. Early HPC_BM; early hematopoietic progenitor cell, bone marrow; HSC_BM, hematopoetic stem cell, bone marrow; CMP, committed myeloid progenitor; GMP, granulocyte-macrophage progenitor; MEP, megakaryocyte ethryroid progenitor; PM_BM, promyeloblast, bone marrow; MY_BM; myelocyte, bone marrow.



Supplementary Figure 6: Serum levels of CXCL4 and CXCL7 are not different in Decitabine responders and non-responders. *Left and middle panels*: Enzyme-linked immunosorbent assays (ELISAs) for CXCL4 and CXCL7. *Right panel:* Correlation between levels CXCL4/7 mRNA by RNA-seq (RPKM) and serum levels by ELISA (ng/mL protein)



Supplementary Figure 7. Low-dose Decitabine (DAC) does not impact CD34^{+} cell proliferation, viability, and apoptosis. A-B. Low-doses of DAC do not affect cell proliferation, viability, or apoptosis with or without CXCL4 and 7. CD34+ cells were treated with one dose of CXCL4, CXCL7, or both (50 ng/mL each) or vehicle (PBS/0.1% BSA) and daily doses of 10 nM DAC for 3 days. After 3 days, live cell number **(A)** and viability (**B)** were assessed by trypan blue exclusion, and apoptosis **(C) was** analyzed

by Annexin V and propodium iodide staining**.** The results shown are from one representative experiment out of two independent experiments.

Supplementary Table 1: Complete list of 167 differentially methylated regions between Decitabine (DAC) responders and non-responders. Please see spreadsheet 1 within the separate Excel file. Negative methylation differences indicate hypermethylation in non-responders while positive methylation differences indicate hypermethylation in responders.

Supplementary Table 2: Complete list of differentially methylated regions (DMRs) enriched in MAPK pathway-associated genes. Please see spreadsheet 2 within the separate Excel file.

Supplementary Table 3: List of 21 CpG regions used as predictors in the SVM classifier to predict Decitabine (DAC) response. Please see spreadsheet 3 within the separate Excel file.

Supplementary Table 4: Somatic mutations in GFM CMML cohort

Gene	Responder n (%)	Non-responder n (%)	p-value[A]
SRSF2*	7/8 (87.5)	3/9 (30)	p=0.05
ASXL1	7/12 (58.3)	7/16 (43.7)	p=0.70
TET2	10/12 (83.3)	6/16 (37.5)	p=0.02
RAS	6/12 (50)	3/16 (18.8)	n=0.40
U2AF1	1/12 (8.3)	1/16 (6.25)	n=1.0
JAK2	0/12 (0)	0/16 (0)	n=1.0
SF3B1	0/12 (0)	1/16 (6.25)	n=1.0
KIT	0/12 (0)	1/16 (6.25)	n=1.0
RUNX1	2/12 (16.7)	5/16 (31.3)	n=0.66

**SRSF2* mutational status was not available for all patients; the percentage was calculated based on the number of patients for whom mutational status was known.

[A]Fisher's exact test

Supplementary Table 5: Complete list of genes with at least a two-fold change in expression level between Decitabine (DAC) responders and non-responders identified by RNA-seq after applying the cutoff of p value <0.05. Please see spreadsheet 4 within the separate Excel file. Negative log2 fold-change (-log2FC) values indicate overexpression in non-responders while positive values indicate overexpression in responders.

Supplementary Table 6: Primers and amplicon sequences for the EpiTYPER MassARRAY validation of ERRBS and the MAPK signaling pathway. Please see spreadsheet 5 within the separate Excel file.

Supplementary Table 7: Primers used in the qRT-PCR validation of the RNA-seq results

Primer Name	Sequence
CXCL4 Fwd	5’ GCGCTGAAGCTGAAGAAGAT 3’
CXCL4 Rvs	5’ TTCAGCGTGGCTATCAGTTG 3’
CXCL7/PPBP Fwd	5’ GCGAAAGGCAAAGAGGAAAGTC 3’
CXCL7/PPBP Rvs	5’ CCTTTCCCGATCACTTCCAAAC 3’
ITGb3 Fwd	5’ GCAGGCATTGTCCAGCCTAA 3’
ITGb3 Rvs	5’ AGTCATCAGCCCCAAAGAGG 3’
RPL19 Fwd	5’ TGGCAAGAAGAAGGTCTGGT 3’
RPL19 Rvs	5’ AAGGTGTTTTTCCGGCATC 3’

Supplementary Methods:

Complete code for building the SVM classifier:

```
#########################################################################
################################
### R function to assign group index to samples
### Input: num.sample (the number of samples)
### Input: num.fold (the fold of CV)
assignGroupIndex <- function(num.sample,num.fold){
	groupIndex <- rep(1:num.fold,each=floor(num.sample/num.fold))
	if(length(groupIndex) < num.sample){
		groupIndex <- c(groupIndex, sample(groupIndex, (num.sample-
length(groupIndex)), replace=F))
	}
	return(groupIndex)
}

#########################################################################
################################
### R function to make classification and evaluate the result using cross validation
### Input: dataFrame (a perc.meth data frame with rows being samples and columns
being CpG features)
### Input: sampelLabels (a matrix stroring sample name vs. classes)
### Input: sampleIndex (a vector storing the randomized sample index)
### Input: groupIndex (a vector storing the group ids for the randomized sample index)
### Input: FUN.model (a classifier function)
### Input: FUN.pred (a prediction function)
getAccuROCforClassifierCV <- function(dataFrame, sampleLabels, sampleIndex,
groupIndex, FUN.model,FUN.pred){
	require("e1071")
	require("ROCR")
	sampleID <- rownames(dataFrame)[sampleIndex]
	groups <- unique(groupIndex)
	result <- c()
	for(i in groups){
		print(i)
		sample.test <- sampleID[groupIndex==i]
		sample.train <- sampleID[groupIndex!=i]

		data.train <- dataFrame[sample.train,,drop=F]
		data.test <- dataFrame[sample.test,,drop=F]

		label.train <-
```

```
as.factor(sampleLabels[match(sample.train,sampleLabels[,1]),2])
		label.test <-
as.factor(sampleLabels[match(sample.test,sampleLabels[,1]),2])
		#print(label.test)

		model <- FUN.model(data.train,label.train)
		pred.label <- FUN.pred(model,data.test)

		result <- rbind(result,data.frame("pred"=pred.label,
"obs"=(label.test=="1")))
	}
	#ROC
	pred <- prediction(result[,1],result[,2])
	perf.auc <- performance(pred,"auc")
	perf.mcc <- performance(pred, "mat")
	perf.roc <- performance(pred,"tpr","fpr")
	perf.precrecall <- performance(pred, "prec","rec")

	return(list("labels.pred"=result,"performance"=list("auc"=perf.auc,"mcc"=perf.mcc,
"roc"=perf.roc,"precrecall"=perf.precrecall)))
}
##########################################################################
################################

### feature  filtering by raw p value by mehtylSig and meth.diff
library(methylSig)

myDiffSigBothTile25_coverage10vs500_p0.05d20 <-
myDiffSigBothTile25_coverage10vs500[myDiffSigBothTile25_coverage10vs500[,"pvalue"]
< 0.05 & abs(myDiffSigBothTile25_coverage10vs500[,"meth.diff"]) > 20,]

ID_myDiffSigBothTile25_coverage10vs500_p0.05d20 <-
paste(myDiffSigBothTile25_coverage10vs500_p0.05d20@data.chr,myDiffSigBothTile25_co
verage10vs500_p0.05d20@data.start,myDiffSigBothTile25_coverage10vs500_p0.05d20@d
ata.end,sep=":")

### feature selection
set.seed(10)
sampleIndex <- sample(1:39,39,replace=F)
groupIndex <- assignGroupIndex(39,10)

result_10Fold_tiles25_p0.05d20  <- c()
for (i in 1:length(ID_filtered)){
	print(i)
	feature <- ID_filtered[i]
```

```
	result_10Fold_tiles25_p0.05d20  <- c(result_10Fold_tiles25_p0.05d20 ,
getAccuROCforClassifierCV(data.frame(t(perc.meth_Tile25_coverage10vs500_rmNArows[f
eature,, drop=F])), sampleInfo,  sampleIndex, groupIndex,
FUN.model=function(x,y){svm(x,y, kernel="sigmoid", probability=T)},
FUN.pred=function(model,newdata){ pred=predict(model,newdata,probability=T,
decision.value=T); prob.class1=attr(pred, "probabilities")[,
"1"];return(prob.class1)})$performance$auc@y.values[[1]])
	}
names(result_10Fold_tiles25_p0.05d20) <- ID_filtered


### search for the best set of tiles as classifier features

ID_filtered_sort <- ID_filtered[order(result_10Fold_tiles25_p0.05d20,decreasing=T)]

ForwardSelectionResult_10Fold_1337Tiles25_p0.05d20  <- list()

sampleIndex <- sample(1:39,39,replace=F)
groupIndex <- assignGroupIndex(39,10)

feature <- c()
for (i in 1:length(ID_filtered_sort)){
	print(i)
	feature <- c(feature, ID_filtered_sort[i])
	auc <-
getAccuROCforClassifierCV(data.frame(t(perc.meth_Tile25_coverage10vs500_rmNArows[f
eature,, drop=F])), sampleInfo,  sampleIndex, groupIndex,
FUN.model=function(x,y){svm(x,y, kernel="sigmoid", probability=T)},
FUN.pred=function(model,newdata){ pred=predict(model,newdata,probability=T,
decision.value=T); prob.class1=attr(pred, "probabilities")[,
"1"];return(prob.class1)})$performance$auc@y.values[[1]]
ForwardSelectionResult_10Fold_1337Tiles25_p0.05d20[[i]] <-
list("AUC"=auc,"feature"=feature)
	}

### plot feature forward selection auc values
plot(unlist(lapply(ForwardSelectionResult_10Fold_1337Tiles25_p0.05d20,function(x){x$A
UC}))[1:30],type="l",lwd=2,ylab="AUC (10-fold cross-validation)", xlab="Number of
features",main="Greedy forward feature selection", cex.lab=1,cex.main=1.5)

points(unlist(lapply(ForwardSelectionResult_10Fold_1337Tiles25_p0.05d20,function(x){x
$AUC}))[1:30],pch=20,col="blue")

points(21,unlist(lapply(ForwardSelectionResult_10Fold_1337Tiles25_p0.05d20,function(x
){x$AUC}))[21],pch=20,cex=2,col="red")
```

```
#############################################################
### independent validation on GFM cohort using 6 out of the 21 pre-selected features ###

Features6_25Tile_inBothSantiniAndSolary <-
ForwardSelectionResult_10Fold_1337Tiles25_p0.05d20[[21]]$feature[!is.na(match(Forwa
rdSelectionResult_10Fold_1337Tiles25_p0.05d20[[21]]$feature,rownames(perc.meth_Tile
25_coverage10vs500_Solary_rmNArow)))]

## train SVM model using the 6 common features

model_SVM_6FeatureIinBothSantiniAndSolary_Tile25_39SantiniSamples <-
svm(data.frame(t(perc.meth_Tile25_coverage10vs500[Features6_25Tile_inBothSantiniAnd
Solary, sampleInfo[,1],drop=F])),as.factor(sampleInfo[,2]), kernel="sigmoid",
probability=T, decision.value=T)

## make prediction

pred_Solary28Samples_6FeatureIinBothSantiniAndSolary <-
predict(model_SVM_6FeatureIinBothSantiniAndSolary_Tile25_39SantiniSamples,
data.frame(t(perc.meth_Tile25_coverage10vs500_Solary_rmNArow[Features6_25Tile_inBo
thSantiniAndSolary, ,drop=F])), probability=T, decision.value=T)

## calculate accuracy
accuracy <-
sum(SampleInfo_28SolaryAnd39SantiniSamples[match(names(pred_Solary28Samples_6Fe
atureIinBothSantiniAndSolary[1:28]),SampleInfo_28SolaryAnd39SantiniSamples[,1]),2]==
pred_Solary28Samples_6FeatureIinBothSantiniAndSolary[1:28])/28
```