Phenotype in combination with genotype improves outcome prediction in acute myeloid leukemia: a report from Children's Oncology Group protocol AAML0531

Andrew P. Voigt,[1*] Lisa Eidenschink Brodersen,[1*] Todd A. Alonzo,[2,3] Robert B. Gerbing,[2] Andrew J. Menssen,[1] Elisabeth R. Wilson,[1] Samir Kahwash,[4] Susana C. Raimondi,[5] Betsy A. Hirsch,[6] Alan S. Gamis,[7] Soheil Meshinchi,[2,8] Denise A. Wells[1] and Michael R. Loken[1]

[1]Hematologics, Inc, Seattle, WA; [2]Children's Oncology Group, Monrovia, CA; [3]University of Southern California, Los Angeles, CA; [4]Nationwide Children's Hospital, Columbus, OH; [5]St. Jude's Children's Research Hospital, Memphis, TN; [6]University of Minnesota Medical Center, Minneapolis, MN; [7]Children's Mercy Hospitals & Clinics, Kansas City, MO and [8]Fred Hutchinson Cancer Research Center, Seattle, WA, USA

**APV and LEB contributed equally to this study*

 doi:10.3324/haematol.2017.169029

Received: March 21, 2017.

Accepted: September 6, 2017.

Pre-published: September 7, 2017.

Correspondence: lisa@hematologics.com

Supplementary material:

Methods:

Immunophenotypic expression profile

The diagnostic AML population was identified by CD45 versus log-SSC gating with WinList (Verity Software House, Topsham, ME). The log mean fluorescence intensities of twelve cell surface antigens were calculated and incorporated into each patient's IEP along with average forward scatter (FSC) and side scatter (SSC) of the leukemic cells. Further, the coefficient of variation (CV) of CD34 was included in the IEP to assess the extent of maturation of leukemic cells.[5,9,10] Together, these independently quantified characteristics defined the IEP for each patient as a location in a 15-dimensional data space (Figure 1f–g).

Statistical analyses

Data from AAML0531 were current as of March 31, 2015. The significance of the observed difference in proportions was tested by Pearson's χ^2 test or Fisher's exact test when data were sparse. The Kruskal–Wallis test was used to test for differences in medians of continuous variables. The Kaplan–Meier method[18] was used to estimate 5-year overall survival (OS) and event-free survival (EFS). Differences between groups were tested by the log-rank test. OS was defined as the time from study entry until death. EFS was defined as the time from study entry until induction failure, relapse, or death. Survival probabilities were reported with 95% confidence intervals (CIs) calculated by the complementary log–log transformation. Cox proportional hazard models estimated hazard ratios (HRs) for univariable and multivariable analyses of OS and EFS.

Tables

Supplementary Table 1. Classification of AML according to the combined FAB and 2001 WHO classification that was used to assign subtypes for patients in the central review.

FAB Classification of AML	**2001 WHO** Classification of AML					
	AML with recurrent cytogenetics abnormality	AML- NOS	**AML with Multi-lineage Dysplasia**	***AML-MDS therapy related***	**Acute Panmyelosis/ Myelofibrosis**	**Myeloid Sarcoma**
FAB: M0		AML, NOS, minimal differentiation				
FAB: M1		AML, NOS, without maturation				
FAB: M2	t (8;21)	AML, NOS with maturation				
	Other FAB: M2					
FAB: M3	APL or AML with t (15;17)(PML/RARA)	Note: APL cases were excluded from AAML 0531 trial				
	Other APL variants					
FAB: M4	inv 16; or t (16;16)	AML, NOS Myelomonocytic				
	Other FAB: M4					
FAB: M5	AML with 11q23/ MLL abnormalities	AML, NOS Monocytic /Monoblastic				
	Other FAB: M5					
FAB: M6		Acute Erythroid Leukemia				
FAB: M7		Acute Megakaryoblastic Leukemia				
AML, Other						

Blue : FAB (Morphologic Diagnosis)

Tan : WHO (Genetics/Immuno/Morphologic)

Gold : Both FAB & WHO designations

Abbreviations: FAB classification, French–American–British classification; WHO, World Health Organization; AML. Acute myeloid leukemia; NOS, not otherwise specified

Supplementary Table 2. Summary of age and morphology and cytogenetic/molecular abnormalities associated with risk stratification at diagnosis for the phenotypic clusters determined by HCA.

	Cluster A (N=266)	Cluster B (N=77)	Cluster C (N=106)	Cluster D (N=68)	Cluster E (N=52)	Cluster F (N=28)	Cluster G (N=11)	Cluster H (N=81)	Cluster I (N=23)	Cluster J (N=41)	Cluster K (N=16)
	N (%)	N (%)	N (%)	N (%)	N (%)	N (%)	N (%)	N (%)	N (%)	N (%)	N (%)
FAB Class											
FAB M0	3 (1%)	0 (0%)	1 (1%)	2 (3%)	1 (2%)	0 (0%)	0 (0%)	7 (9%)	4 (21%)	3 (8%)	1 (8%)
FAB M1	28 (13%)	14 (19%)	15 (16%)	2 (3%)	0 (0%)	0 (0%)	0 (0%)	13 (17%)	8 (42%)	7 (19%)	3 (23%)
FAB M2	63 (29%)	52 (71%)	23 (25%)	2 (3%)	3 (7%)	3 (13%)	0 (0%)	12 (16%)	2 (11%)	1 (3%)	0 (0%)
FAB M4	91 (42%)	4 (5%)	24 (26%)	9 (16%)	5 (11%)	6 (25%)	0 (0%)	19 (25%)	2 (11%)	5 (14%)	0 (0%)
FAB M5	13 (6%)	1 (1%)	16 (17%)	40 (69%)	35 (78%)	15 (63%)	9 (100%)	7 (9%)	1 (5%)	7 (19%)	0 (0%)
FAB M6	4 (2%)	0 (0%	0 (0%)	1 (2%)	0 (0%)	0 (0%)	0 (0%)	1 (1%)	1 (5%)	2 (6%)	0 (0%)
FAB M7	1 (0%)	1 (1%)	1 (1%)	1 (2%)	0 (0%)	0 (0%)	0 (0%)	12 (16%)	1 (5%)	6 (17%)	7 (54%)
Other	15 (7%)	1 (1%)	12 (13%)	1 (2%)	1 (2%)	0 (0%)	0 (0%)	5 (7%)	0 (0%)	5 (14%)	2 (15%)
Unknown	48	4	14	10	7	4	2	5	4	5	3
Cytogenetic Abnormality											
Normal	61 (23%)	4 (5%)	48 (47%)	11 (16%)	0 (0%)	7 (26%)	0 (0%)	11 (14%)	12 (52%)	12 (30%)	6 (38%)
t(8;21)	45 (17%)	55 (72%)	1 (1%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	8 (10%)	0 (0%)	0 (0%)	0 (0%)
Inv(16)	75 (29%)	2 (3%)	7 (7%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	9 (12%)	2 (9%)	0 (0%)	0 (0%)
11q23	11 (4%)	0 (0%)	10 (10%)	46 (69%)	35 (69%)	16 (59%)	9 (82%)	12 (16%)	1 (4%)	12 (30%)	0 (0%)
t(6;9)	4 (2%)	0 (0%)	4 (4%)	1 (1%)	0 (0%)	0 (0%)	0 (0%)	3 (4%)	0 (0%)	0 (0%)	0 (0%)
Monosomy 7	9 (3%)	2 (3%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	3 (4%)	0 (0%)	1 (3%)	0 (0%)
Del(7q)	8 (3%)	1 (1%)	0 (0%)	0 (0%)	1 (2%)	0 (0%)	1 (9%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Monosomy 5/Del(5q)	3 (1%)	0 (0%)	0 (0%)	0 (0%)	1 (2%)	0 (0%)	0 (0%)	3 (4%)	1 (4%)	0 (0%)	0 (0%)
Trisomy 8	11(4%)	2 (3%)	21 (21%)	5 (7%)	2 (4%)	1 (4%)	0 (0%)	3 (4%)	2 (9%)	2 (5%)	0 (0%)
Other	33 (13%)	10 (13%)	11 (11%)	4 (6%)	12 (24%)	3 (11%)	1 (9%)	25 (32%)	5 (22%)	13 (33%)	10 (63%)
Unknown	6	1	4	1	1	1	0	4	0	1	0
Molecular Abnormality											

FLT3/ITD+	49 (19%)	4 (5%)	47 (54%)	5 (8%)	1 (2%)	4 (15%)	0 (0%)	6 (8%)	6 (26%)	9 (22%)	0 (0%)
WT1 mutation	30 (11%)	2 (3%)	7 (7%)	1 (1%)	0 (0%)	1 (4%)	0 (0%)	9 (11%)	2 (9%)	0 (0%)	0 (0%)
NPM1 mutation	4 (2%)	1 (1%)	23 (22%)	8 (12%)	1 (2%)	7 (26%)	0 (0%)	5 (6%)	1 (4%)	8 (20%)	0 (0%)
CEBPA mutation	22 (8%)	5 (6%)	11 (10%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	3 (4%)	4 (17%)	1 (2%)	0 (0%)
Risk group (cyto/mutation)											
Standard	80 (31%)	11 (14%)	38 (36%)	56 (84%)	48 (94%)	21 (75%)	11 (100%)	43 (54%)	11 (48%)	25 (63%)	16 (100%)
Low	140 (54%)	63 (82%)	34 (32%)	6 (9%)	1 (2%)	6 (21%)	0 (0%)	25 (32%)	6 (26%)	9 (23%)	0 (0%)
High	41 (16%)	3(4%)	33 (31%)	5 (7%)	2 (4%)	1 (4%)	0 (0%)	11 (14%)	6 (26%)	6 (15%)	0 (0%)
Unknown	5	0	1	1	1	0	0	2	0	1	0
Cytogenetic risk group											
Standard	128 (49%)	17 (22%)	94 (92%)	67 (100%)	50 (98%)	27 (100%)	11 (100%)	54 (70%)	20 (87%)	39 (98%)	16 (100%)
Low	120 (46%)	57 (75%)	8 (8%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	17 (22%)	2 (9%)	0 (0%)	0 (0%)
High	12 (5%)	2 (3%)	0 (0%)	0 (0%)	1 (2%)	0 (0%)	0 (0%)	6 (8%)	1 (4%)	1 (3%)	0 (0%)
Unknown	6	1	4	1	1	1	0	4	0	1	0
Age											
Median (range)	11.7 (0.17 - 29.8)	11.2 (1.4 -19.0)	13.1 (0.41 -20.9)	4.2 (0.02 - 18.6)	2.4 (0.15 -17.8)	1.7 (0.01 - 18.0)	0.69 (0.21 - 18.9)	9.0 (0.06 - 23.9)	13.6 (1.2 - 23.5)	4.3 (0.14 - 18.8)	2.1 (0.75 - 13.8)

Supplementary Table 3. Presence of cytogenetic or molecular abnormalities within the genotypic subclusters.

	inv(16) (N=95)	t(8;21) (N=105)	11q23 (N=153)	FLT3/ITD (N=127)	NPM (N=58)	CEBPA (N=46)	WT1 (N=55)	GLIS (N=17)
	N (%)*	N (%)*	N (%)*	N (%)*	N (%)*	N (%)*	N (%)*	N (%)*
Subcluster								
A-i (N=29)	2 (2.1%)	3 (2.8%)	1 (0.7%)	4 (3.1%)	0 (0%)	14 (30.4%)	5 (9.1%)	2 (11.8%)
A-ii (N=58)	50 (52.6%)	1 (0.9%)	0 (0%)	4 (3.1%)	0 (0%)	0 (0%)	3 (5.5%)	0 (0%)
A-iii (N=26)	0 (0%)	22 (20.2%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (1.8%)	0 (0%)
A-iv (N=32)	0 (0%)	1 (0.9%)	0 (0%)	17 (13.4%)	2 (3.4%)	5 (10.9%)	13 (23.6%)	0 (0%)
A-v (N=55)	19 (20%)	17 (15.6%)	1 (0.7%)	3 (2.4%)	1 (1.7%)	0 (0%)	3 (5.5%)	0 (0%)
A-vi (N=36)	0 (0%)	0 (0%)	6 (3.9%)	17 (13.4%)	1 (1.7%)	1 (2.2%)	6 (10.9%)	0 (0%)
B-i (N=65)	2 (2.1%)	54 (49.5%)	0 (0%)	2 (1.6%)	0 (0%)	2 (4.3%)	1 (1.8%)	1 (5.9%)
C-i (N=53)	5 (5.3%)	0 (0%)	6 (3.9%)	25 (19.7%)	5 (8.6%)	11 (23.9%)	6 (10.9%)	0 (0%)
C-ii (N=42	2 (2.1%)	0 (0%)	2 (1.3%)	18 (14.2%)	18 (31%)	0 (0%)	0 (0%)	1 (5.9%)
D-i (N=68)	0 (0%)	0 (0%)	45 (29.4%)	5 (3.9%)	8 (13.8%)	0 (0%)	1 (1.8%)	0 (0%)
E-i (N=52)	0 (0%)	0 (0%)	35 (22.9%)	1 (0.8%)	1 (1.7%)	0 (0%)	0 (0%)	0 (0%)
F-i (N=28)	0 (0%)	0 (0%)	16 (10.5%)	3 (2.4%)	7 (12.1%)	0 (0%)	1 (1.8%)	0 (0%)
G-i (N=11)	0 (0%)	0 (0%)	9 (5.9%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
H-i (N=19)	0 (0%)	2 (1.8%)	10 (6.5%)	0 (0%)	0 (0%)	1 (2.2%)	1 (1.8%)	0 (0%)
K-i (N=16)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	10 (58.8%)

*** Indicates that the (%) shown is calculated with the numerator as the total of number of patients that have the genetic abnormality within a sub-cluster and the denominator as the total number of patients**

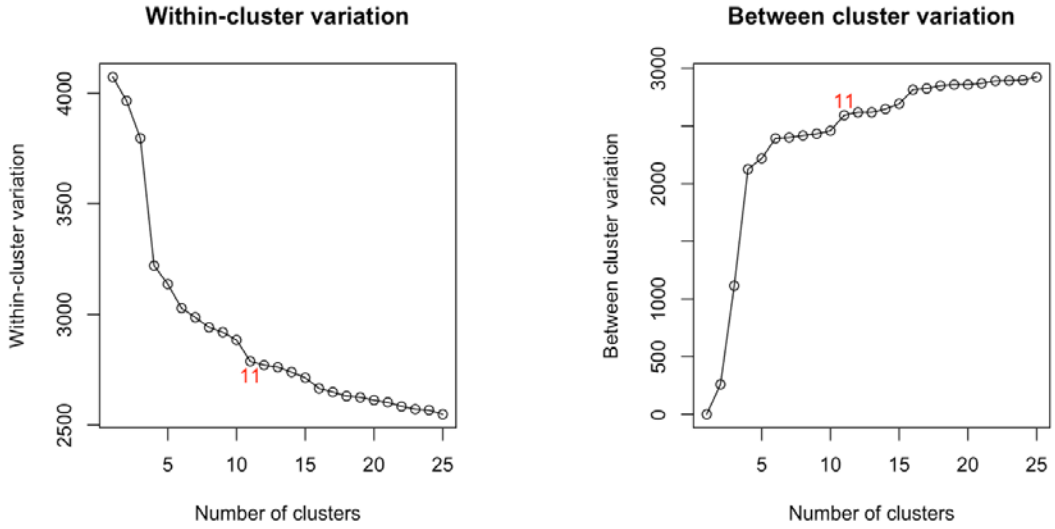
Supplementary Table 4: Performance of Supervised Bagged Decision Trees in Replicating Cluster Classifications from Unsupervised Hierarchical Clustering. A boosted decision tree model was trained to replicate the original unsupervised hierarchical clustering eleven-cluster classification using only the IEP. The decision tree model was trained with the gbm package in R. Selection of the number of trees, shrinkage parameter, and interaction depth were tuned in cross-validation in the training cohort. In the test cohort, 84.0% of patients were accurately classified. The sensitivity, specificity, and F1-score (the harmonic mean of precision and recall) are displayed for the test cohort below.

	Sensitivity	Specificity	F1-score
Test Cohort Patient Cluster			
A (N=87)	0.954	0.929	0.912
B (N=28)	0.786	0.987	0.830
C (N=37)	0.811	0.950	0.770
D (N=27)	0.741	0.991	0.816
E (N=17)	1	0.987	0.919
F (N=7)	0.714	0.996	0.769
G (N=3)	0.667	1	0.8
H (N=26)	0.577	0.983	0.667
I (N=4)	1	1	1
J (N=16)	0.813	0.979	0.765
K (N=4)	1	1	1
Mean	**0.824**	**0.982**	**0.841**

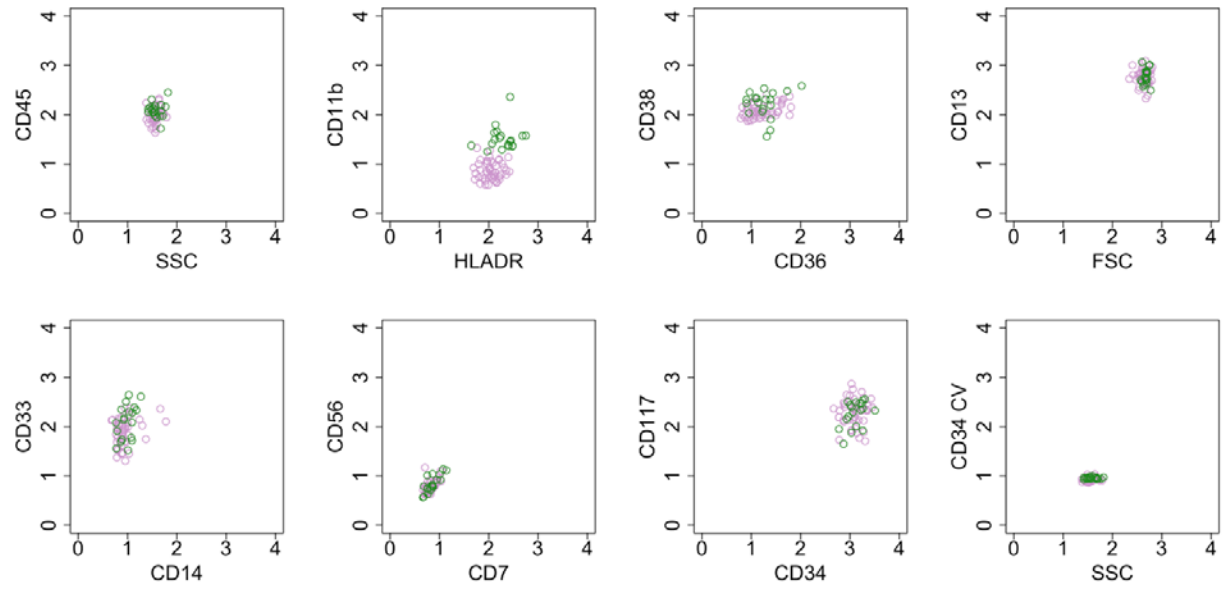
Supplementary Table 5: Performance of Supervised Bagged Decision Trees in Replicating Subcluster Classifications. In the training cohort, a boosted decision tree model was trained to identify patients in Subcluster A-ii and Subcluster A-v. The decision tree model was trained with the gbm package in R. Selection of the number of trees, shrinkage parameter, and interaction depth were tuned in cross-validation in the training cohort. Performance was evaluated only on inv(16) patients in the testing cohort (N=26). The inv(16) boosted decision tree model accurately classified the inv(16) patients into Subcluster A-ii, Subcluster A-v, or non-A-ii-non-A-v subclusters with an accuracy of 92.3%. For patients with 11q23, Subclusters D-i, E-i, F-i, and G-i completely overlap with Clusters D, E, F, and G. Hence no additional boosted decision tree models were trained to identify these subclusters. A boosted decision tree model was trained to identify patients in Subcluster H-i. Performance was evaluated only on 11q23 patients in the testing cohort (N=52). The combination of boosted decision trees identifying Clusters D, E, F, and G with the Subcluster H-i boosted decision tree correctly classified 11q23 patients into Subclusters D-i, E-i, F-i, G-i, H-i, or other groupings with an accuracy of 95.3%.

	Sensitivity	Specificity	F1-score
inv(16) subcluster			
A-ii (N=14)	1	0.833	0.933
A-v (N=3)	0.667	0.957	0.667
Mean	0.833	0.895	0.800
11q23 subcluster			
D-i (N=15)	0.800	0.946	0.828
E-I (N=12)	1	0.950	0.923
F-i (N=6)	0.667	1	0.800
G-i (N=2)	1	1	1
H-i (N=4)	0.25	1	0.40
Mean	**0743**	**0.979**	**0.790**

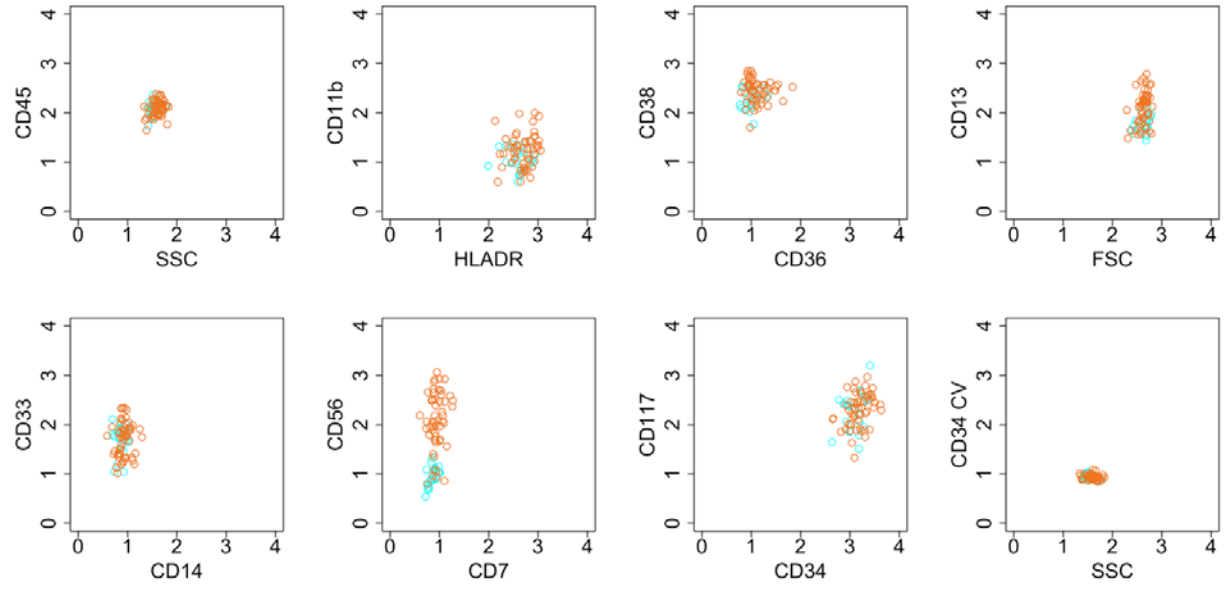
Figures:



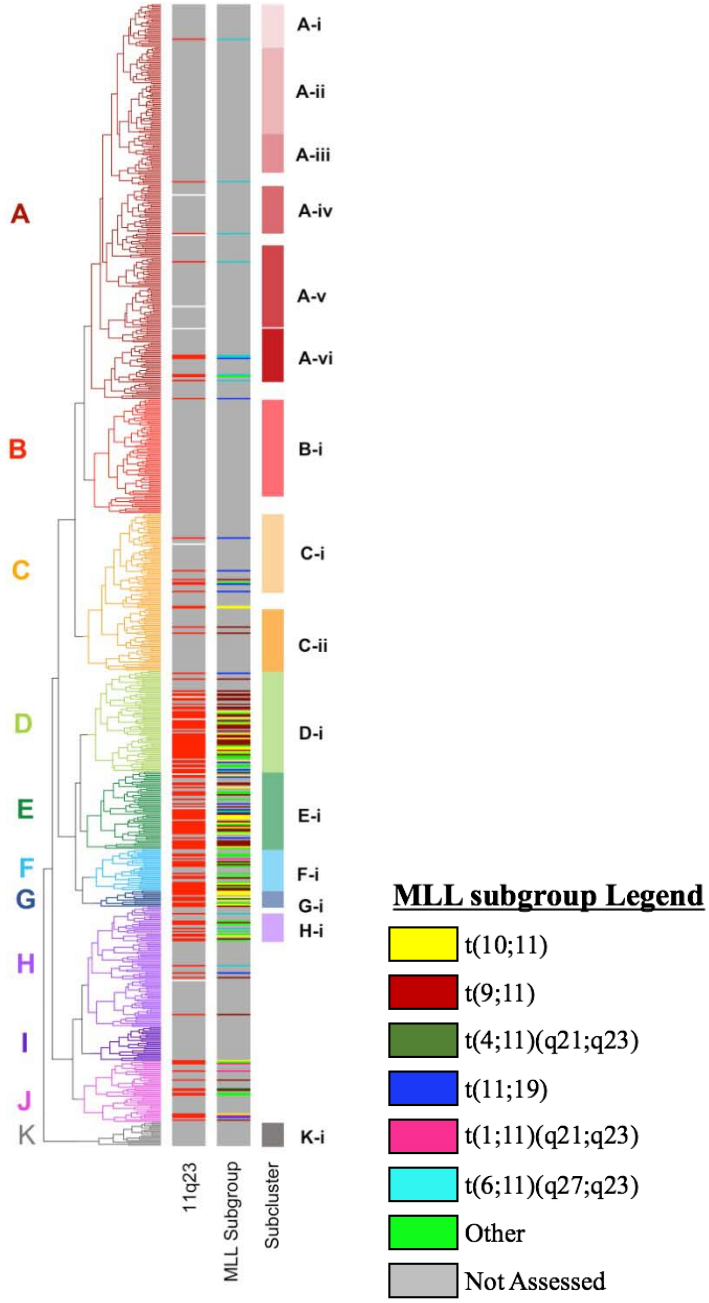
Supplementary Figure 1. Cluster selection. (A) Compared to 10 clusters, the division of the dendrogram into 11 clusters resulted in an abrupt decrease in within-cluster variation (W). W = $\sum_{k=1}^{K}\sum_{C(i)=k}\|X_i - \bar{X}_k\|_2^2$ over clustering assignments *C*, where X_i is the IEP for patient *i* in cluster *k*, $\bar{X}_k$is the average IEP in cluster *k*, and *K* is the total number of clusters. **(B)** Likewise, the division of the dendrogram into 11 clusters resulted in an abrupt increase in between-cluster variation (B). B = $\sum_{k=1}^{K} n_K \|\bar{X}_k - \bar{X}\|_2^2$, where $\bar{X}_k$is the average IEP in cluster *k* and $\bar{X}_k$is the average IEP of all 769 patients.



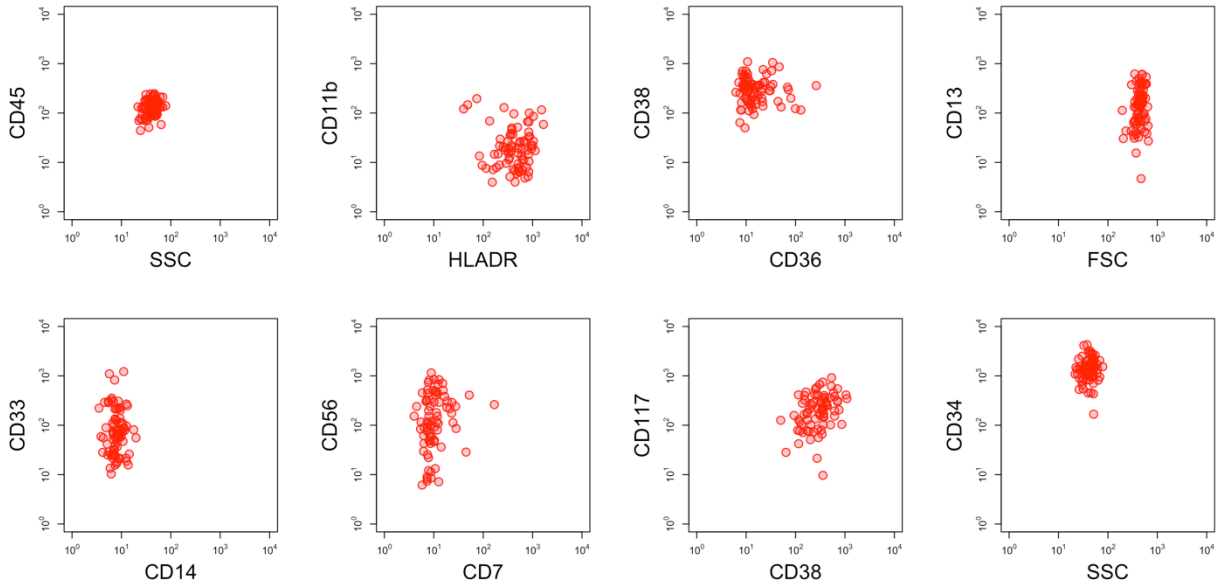
Supplementary Figure 2. IEPs of inv(16)-enriched regions. IEPs for patients with inv(16) abnormalities in Subclusters A-ii (purple) and A-v (green) of the clustering heatmap are displayed in 2 dimensions. Each dot represents the IEP of 1 patient. Patients with inv(16) abnormalities in Subcluster A-v had a slightly higher expression of CD11b than those in Subcluster A-ii.



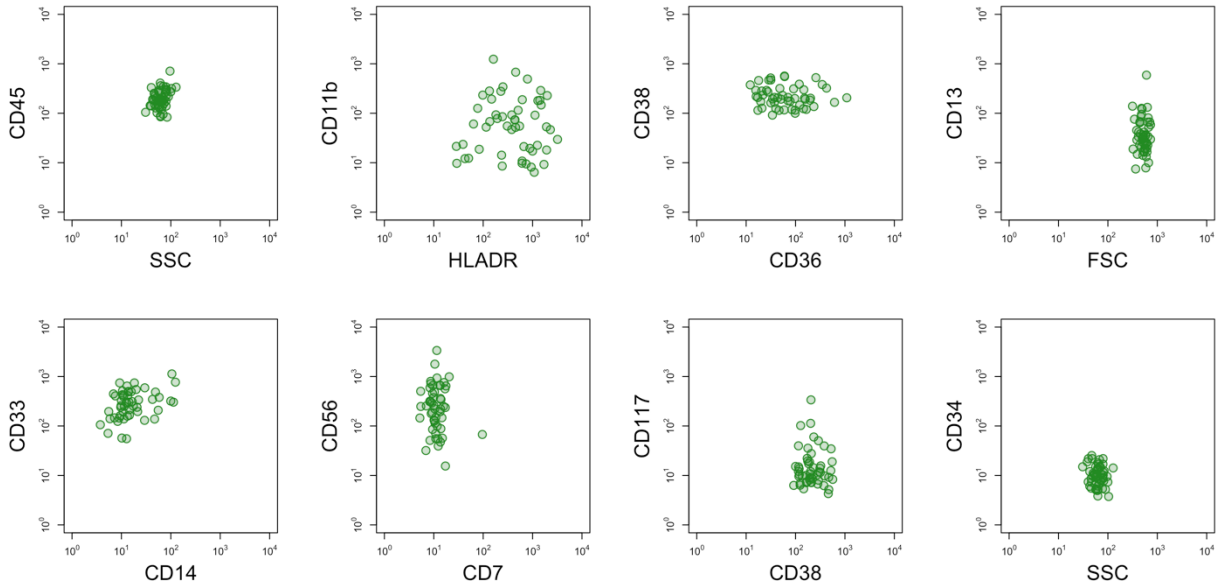
Supplementary Figure 3. IEPs of t(8;21)-enriched regions. IEPs for patients with t(8;21) abnormalities in Subclusters A-iii (light blue) and B-i (orange) of the clustering heatmap are displayed in 2 dimensions. Each dot represents the IEP of 1 patient. Patients with t(8;21) abnormalities in Subcluster B-i had higher expression of CD56 than those in Subcluster A-iii.



Supplementary Figure 4. Subanalysis of 11q23 translocation partners. Data on translocation partners for each patient with 11q23 abnormalities were appended to the dendrogram. Specific translocation partners are given in the MLL subgroup legend.

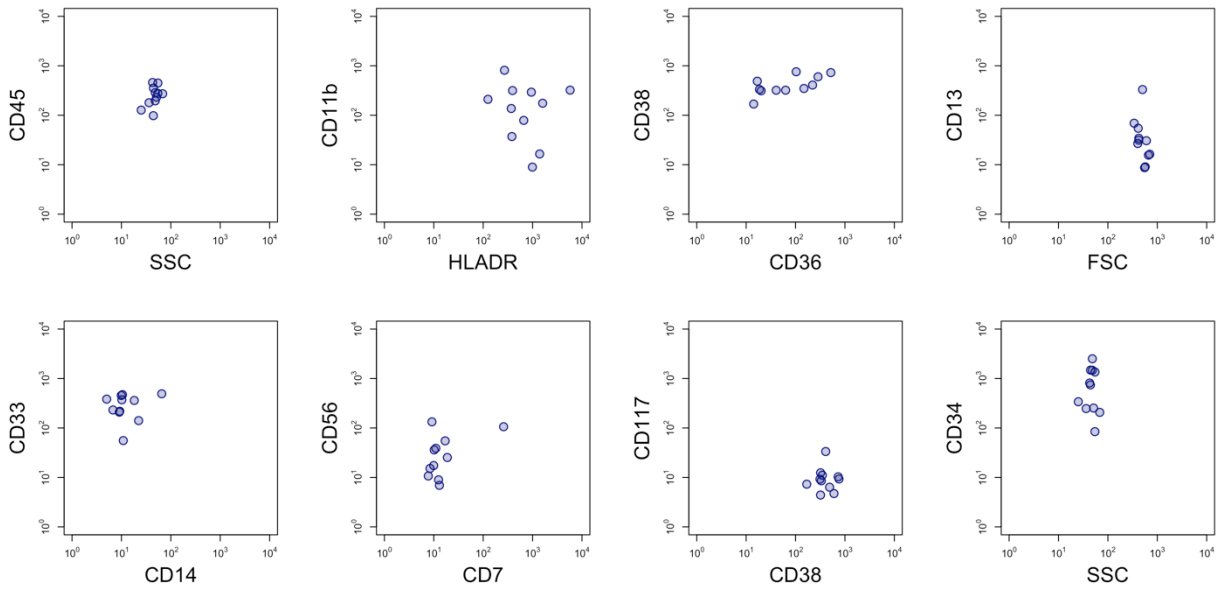


Supplementary Figure 5. Summary of patient IEPs in Cluster B. A 2-dimenional representation of each patients IEP within cluster B is shown. While this representation distorts the high dimensional nature of the IEPs it does allow for an understanding of the immunophenotype for the independent clusters. Cluster B is notable in that all patients have bright CD34 expression and frequently co-express CD56.

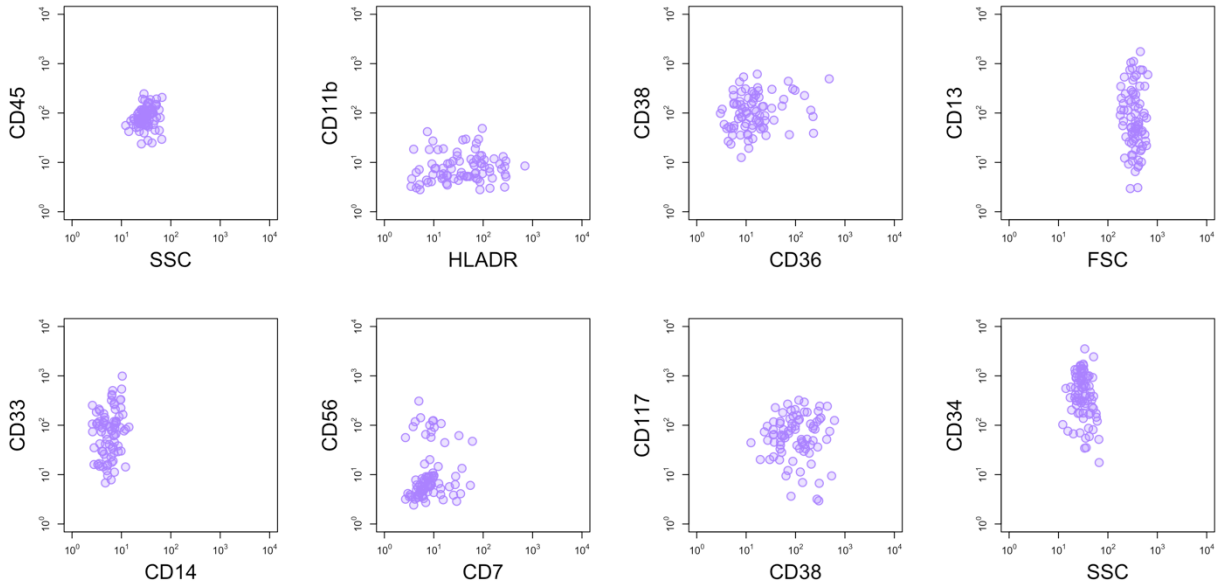


Supplementary Figure 6. Summary of patient IEPs in Cluster E. A 2-dimenional representation of each patients IEP within cluster E is shown. While this representation distorts the high dimensional nature of the IEPs it does allow for an understanding of the immunophenotype for the independent clusters. Cluster E is notable in that these patients are uniformly CD34 negative, frequently express CD56 and

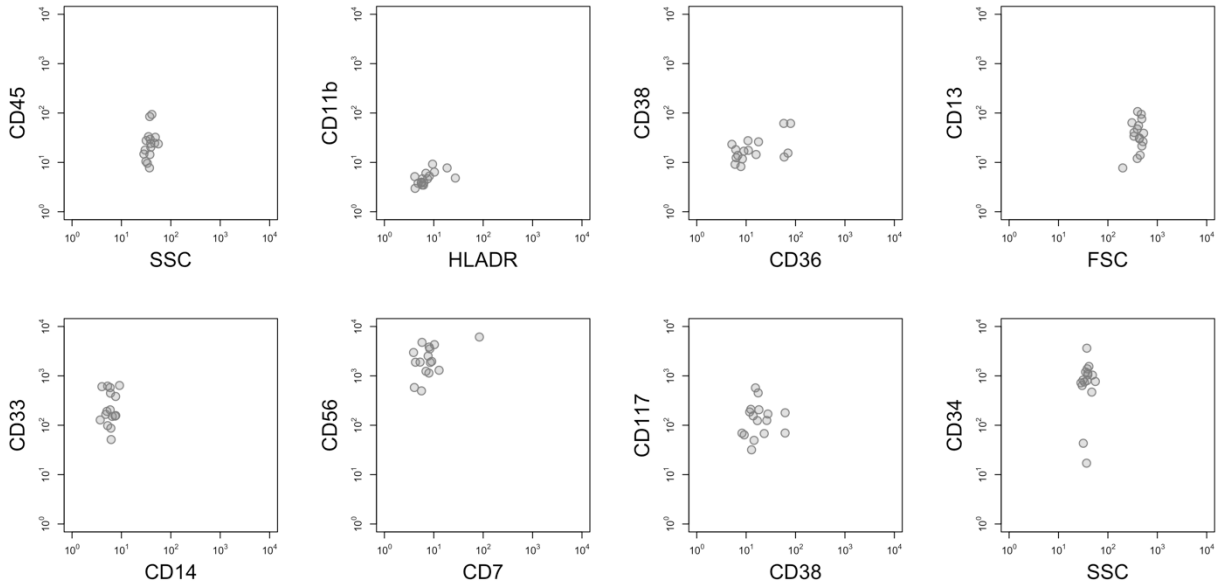
often have phenotypes consistent with mature monocytes (co-expression of CD11b and HLA-DR, expression of CD36 and some expression of CD14).



Supplementary Figure 7. Summary of patient IEPs in Cluster G. A 2-dimenional representation of each patients IEP within cluster G is shown. While this representation distorts the high dimensional nature of the IEPs it does allow for an understanding of the immunophenotype for the independent clusters. Cluster G is notable in that all patients express CD34 but lack expression of CD117. In addition CD33 and CD13 are dimly expressed.

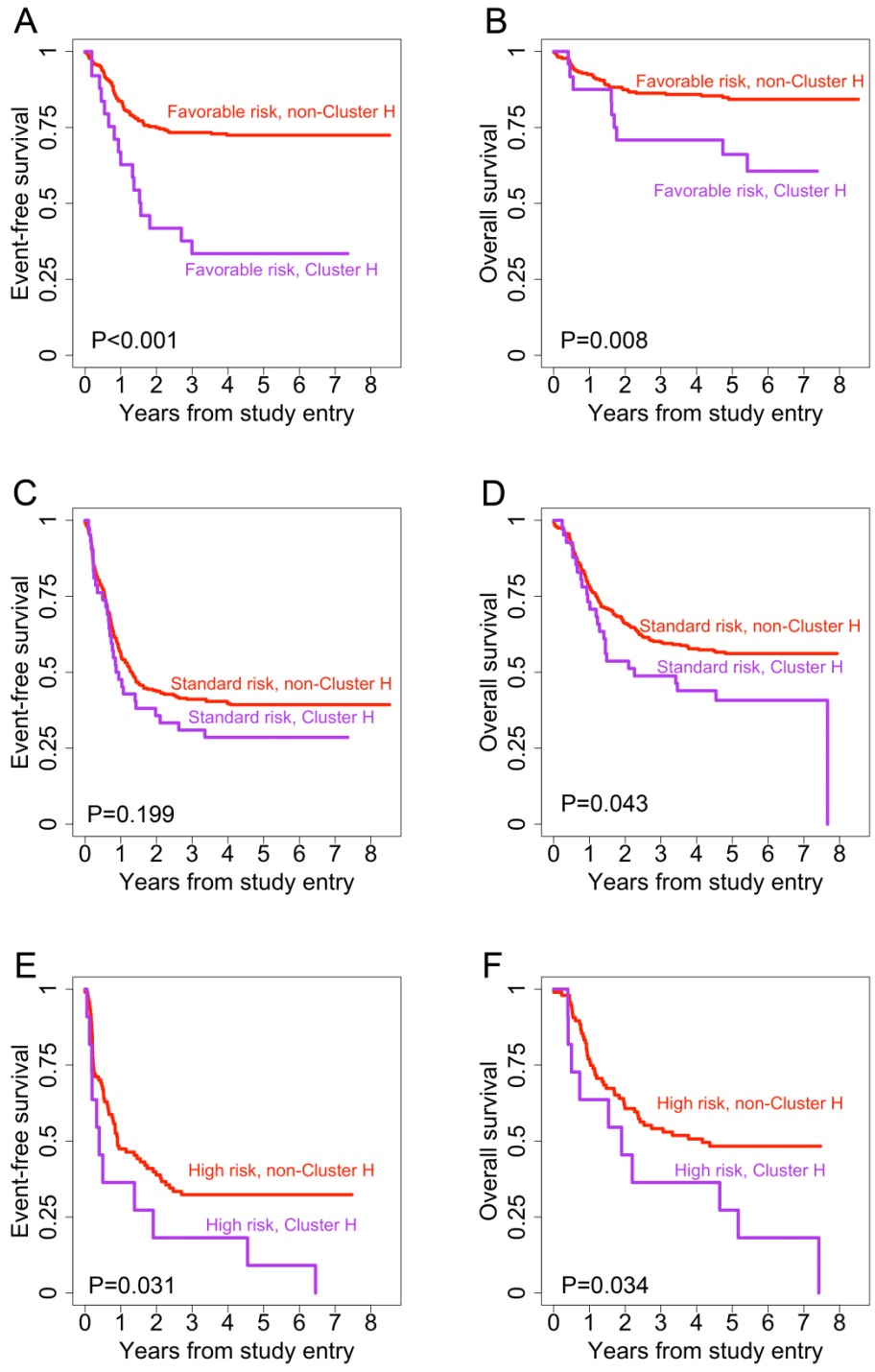


Supplementary Figure 8. Summary of patient IEPs in Cluster H. A 2-dimenional representation of each patients IEP within cluster H is shown. While this representation distorts the high dimensional nature of the IEPs it does allow for an understanding of the immunophenotype for the independent clusters. Cluster H is notable in that all patients have decreased CD38 expression and frequently express CD34. In addition significant heterogeneity is observed for CD13 and CD33.



Supplementary Figure 9. Summary of patient IEPs in Cluster K. A 2-dimenional representation of each patients IEP within cluster K is shown. While this representation distorts the high dimensional nature of the IEPs it does allow for an understanding of the immunophenotype for the independent

clusters. Cluster K is notable in that all patients have very high intensity CD56 expression while lacking HLA-DR and dim to negative expression of CD45 and CD38. This phenotype is consistent with the previously reported RAM Phenotype,[20] which is associated with poor prognosis.



Supplementary Figure 10. Cluster H subanalysis of EFS and OS for favorable, standard and high risk based on cytogenetic/molecular stratification. (A-B) Patients with low-risk cytogenetic/molecular markers in Cluster H had significantly poorer 5-year EFS and 5-year OS than low-risk patients in all other clusters (EFS: 33% vs 72%, $P<0.001$; OS: 66% vs 84%, P=0.008). **(C-D)** Patients with standard-risk cytogenetic/molecular markers in Cluster H had significantly poorer 5-year EFS and 5-year OS than low-risk patients in all other clusters (EFS: 29% vs 39%, $P=0.199$; OS: 41% vs 56%, P=0.043). **(E-F)** Patients with high-risk cytogenetic/molecular markers in Cluster H had significantly poorer 5-year EFS and 5-year OS than high-risk patients in all other clusters (EFS: 9% vs 32%, $P=0.031$; OS: 18% vs 48%, P=0.034).