Association of Circulating Transcriptomic Profiles with Mortality in Sickle Cell Disease

Ankit A. Desai MD[1], Zhengdeng Lei PhD[2], Neil Bahroos[2], Mark Maienschein-Cline[2], Santosh L Saraf MD[3], Xu Zhang[3], Binal N Shah PhD[3], Seyed M Nouraie PhD[4], Taimur Abbasi MD[3], Amit R. Patel MD[5], Roberto M. Lang MD[5], Yves Lussier MD[1], Joe GN Garcia MD[1], Victor R Gordeuk MD[3], Roberto F. Machado MD[3]

[1]Department of Medicine and Arizona Health Sciences Center, University of Arizona, Tucson, AZ USA
[2]Center for Research Informatics, Research Resources Center and Center for Clinical and Translational Science, University of Illinois at Chicago, Chicago, IL, USA
[3]Department of Medicine, University of Illinois Hospitals and Health Sciences System, Chicago, IL, USA
[4]Department of Medicine, Howard University, Washington DC, USA
[5]Department of Medicine, University of Chicago Medical Center, Chicago, Illinois, USA

Supplemental Methods and Data

Supplemental Methods

Clinical Trial Registration. Clinical trial registration information is provided at the following website: http://clinicaltrials.gov/ct2/show/NCT01044901 for the University of Chicago cohort. For Howard University, trial registration details are provided at the following website: https://clinicaltrials.gov/ct2/show/NCT00005541?term=gordeuk&rank=1.

Study design and Cohorts. Hemoglobin subtype was demonstrated by high-performance liquid chromatographic separation or gel electrophoresis for all patients. Subjects from all centers were excluded if they had vaso-occlusive crisis, acute chest syndrome, or unscheduled blood transfusions within 3 weeks of the study ("defined as steady-state").

Microarray preparation and analysis. Blood was drawn via peripheral venipuncture and care was taken to standardize blood sample collection and preparation. Peripheral blood mononuclear cells (PBMCs) were isolated from blood and stored at -80^0C as described previously [1]. Total RNA was isolated from these lysates using Qiagen's RNeasy plus kit as per the manufacturer's protocol. Affymetrix Human Exon 1.0 ST array (for the University of Chicago and Howard University) and Human Gene 2.0 ST (for UIC) was employed for mRNA profiling. RNA quality control was performed with the use of automated electrophoresis system Experion (BioRad). ARTS (automated randomization of

multiple traits for study design) tool was employed for automated randomization of batch processing of microarray processing including labeling and hybridization as previously described [2]. Labeling reactions and hybridizations were carried out according to the standard GeneChip® WTsense target labeling protocol [3]. 100 ng of total cellular RNA per sample was used for each labeling reaction for gene arrays. Hybridizations were followed by binding to streptavidin-conjugated fluorescent marker. Detection of bound probe was achieved following laser excitation of the fluorescent marker and scanning of the resultant emission spectra using a scanning confocal laser microscope [3]. Data acquisition was performed using Affymetrix AGCC suite. Hybridization images were subjected to quality control with the use of Expression Console analysis tool [3]. CEL files were normalized with the RMA algorithm using Affymetrix Power Tools [4]. Adjustment for possible batch effect was conducted by COMBAT (http://jlab.byu.edu//ComBat/) [5]. DAVID tools were used to identify the enriched KEGG (Kyoto Encyclopedia of Genes and Genomes, http://www.genome/jp/kegg) physiological pathways among the potentially differentially expressed genes [6].

Microarray analysis and Consensus Cluster analysis. For pre-processing of expression data used in consensus clustering, all probe sets with median expression levels less than the 20th percentile or with variances less than the 20th percentile were not included for analysis. The expression data was then normalized to mean zero and standard deviation one, first by probe sets and then

by arrays. Consensus clustering was then performed on the remaining expression data. In the Training cohort, unbiased consensus clustering on 172 available gene expression profiles was performed. The consensus clustering approach used a resampling-based procedure that repeatedly samples 80% of SCD arrays and then uses hierarchical clustering to find intrinsic clusters [7]. To determine this optimal number of clusters, the change in area under the CDF was evaluated in response to increasing the number of clusters, *K*. As *K* increases, the area under the CDF markedly increases as long as *K is* less than or equal to (≤) the optimal number of clusters. However, when the optimal number of clusters is reached, further increases of *K* do not lead to any marked increases in the area under the CDF.

Pathway signature and analysis. The application of gene signatures derived from gene expression profiles have been hampered by the lack of reproducibility and a shortage of functional interpretation. Recent studies have shown that pathway-based genomic classifiers could provide greater stability and more robust validation in contrast to traditional gene signature-based classifiers [8,9]. FAIME signatures were derived using the Wilcoxon rank sum test using the Training cohort to compare FAIME scores between the observed clusters identified, support vector machine (SVM) model was then trained on the FAIME signature using the R package “e1071” and utilized to predict clusters in in the Testing and West African cohorts. The optimal SVM parameters were obtained by 10-fold cross validation.

Gene signature. Based on the clusters identified in the Training cohort, a gene signature (fold change ≥ 1.5 and false discovery rate ≤ $1x10^{-5}$) was derived using the R package “limma” differentiating Clusters 1 and 2 in the Training cohort. A separate gene signature was also developed in the Testing cohort, in this case using ComBat to correct for batch effects in the combined the expression data (fold change ≥ 1.5 and false discovery rate ≤ 0.05). The overlap between the Training and Test gene signatures defined a 31-gene signature associated with cluster-specific profiles in SCD. An SVM classifier based on this derived gene signature was constructed, and tested to determine its ability to predict clustering and severity of SCD in the West African cohort.

RT-qPCR. RNA extracted from PBMCs of 8 sickle cell disease (SCD) subjects and 6 African-American control subjects without SCD were evaluated with RT-qPCR. Selection of the top 10 ten differentially expressed transcripts [4 down-regulated: dedicator of cytokinesis 9 (*DOCK9*), lymphoid enhancer binding factor 1 (*LEF1*), neural EGFL like 2 (*NELL2*), golgin A8 family member B (*GOLGA8B*) and 6 upregulated: selenium binding protein 1 (*SELENBP1*), solute carrier family 4, member 1 (*SLC4A1*), erythrocyte membrane protein band 4.2 (*EPB42*), 5'-aminolevulinate synthase 2 (*ALAS2*), glycophorin A (*GYPA*), ferrochelatase (*FECH*)] for confirmation and validation by RT-qPCR were based upon their ability to discriminate patients’ vital status. Results for target gene expression were standardized to the expression of the control gene *RPS11*, selected as a

reference gene based on exhibiting one of the lowest coefficient of variation across all microarray samples (**Supplement Table 6**). First strand cDNA synthesis was carried out from 1.5 µg of total RNA extracted from isolated PBMC of 8 SCD subjects and 6 African-American control subjects without SCD in a 20 µl reaction using Superscript III reverse transcriptase and 0.5 µg of oligodT (12-18mer) following manufacturer's protocol. After 4-fold dilution of the cDNA, an aliquot of each sample was mixed to create the highest standard for verifying primer efficiency. The remaining cDNA was diluted five-fold from which 4 µl was used in a 10 µl RT-qPCR reaction using SSO Advanced Universal Sybr green supermix and a final primer concentration of 300 nM. Primers used for quantitative-PCR are described in **Supplement Table 7**. These primers had an amplification efficiency of 97-105%, over a 1000-fold range. RT-qPCR was carried out using the BioRad CFX384 Real-Time PCR Detection System (Bio-Rad) and a thermal cycling protocol used was denaturation at 95°C for 30 seconds followed by 40 cycles of denaturation at 95°C for 3 seconds and annealing/extension at 60°C for 30 seconds. All experimental samples and standards were run in duplicate and averaged. The qPCR product of one of the sample was run on a 3% agarose gel to determine specificity (**Supplement Figure 4**). The C_t values were obtained using the regression method in Biorad CFX manager 3.1. The relative quantification [$2(-DDC_t)$] method was used to analyze the relative changes in gene expression [10].

Supplement Figure Legends.

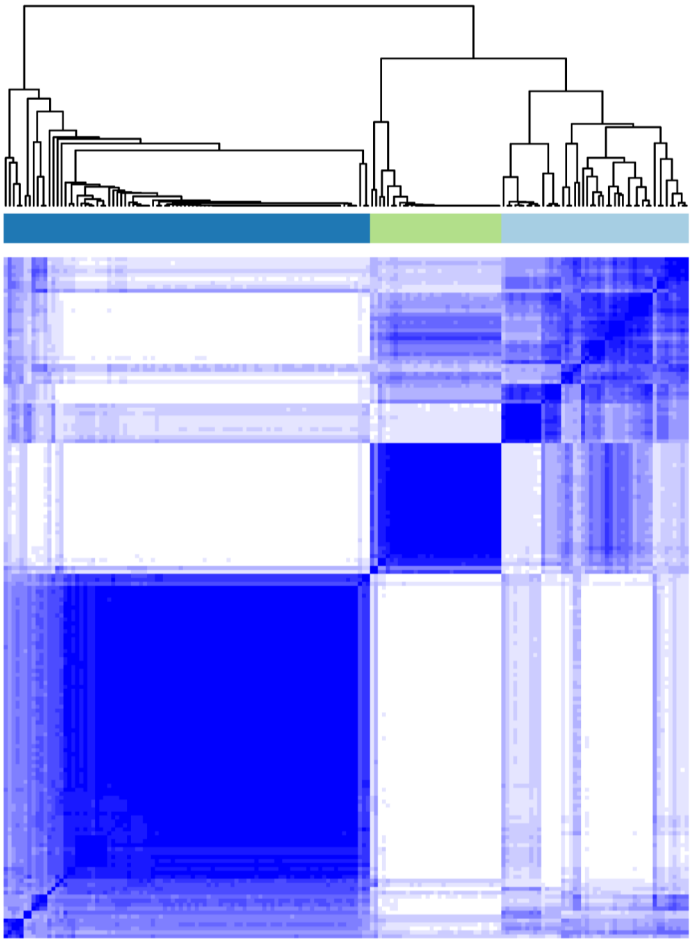
Supplemental Figure 1. The consensus matrix heatmap for three clusters (*K* = 3). The consensus matrix heatmap for *K* = 3 demonstrated a high proportion of samples with ambiguous clustering. Thus we determined the optimal number of clusters to be 2.

Supplemental Figure 2. Transcriptomic profiling and disease severity in West African children with SCD. A). Children with SCD who demonstrate Cluster 1 transcriptomes have significantly higher severity scores than patients in cluster 2 (p= 9.994e-05). **B**) Similarly, children with SCD who demonstrate the 31-gene signature also predict Cluster 1 profiling with significantly higher severity scores those that exhibit a signature which predicts Cluster 2 (p= 2.284e-05).

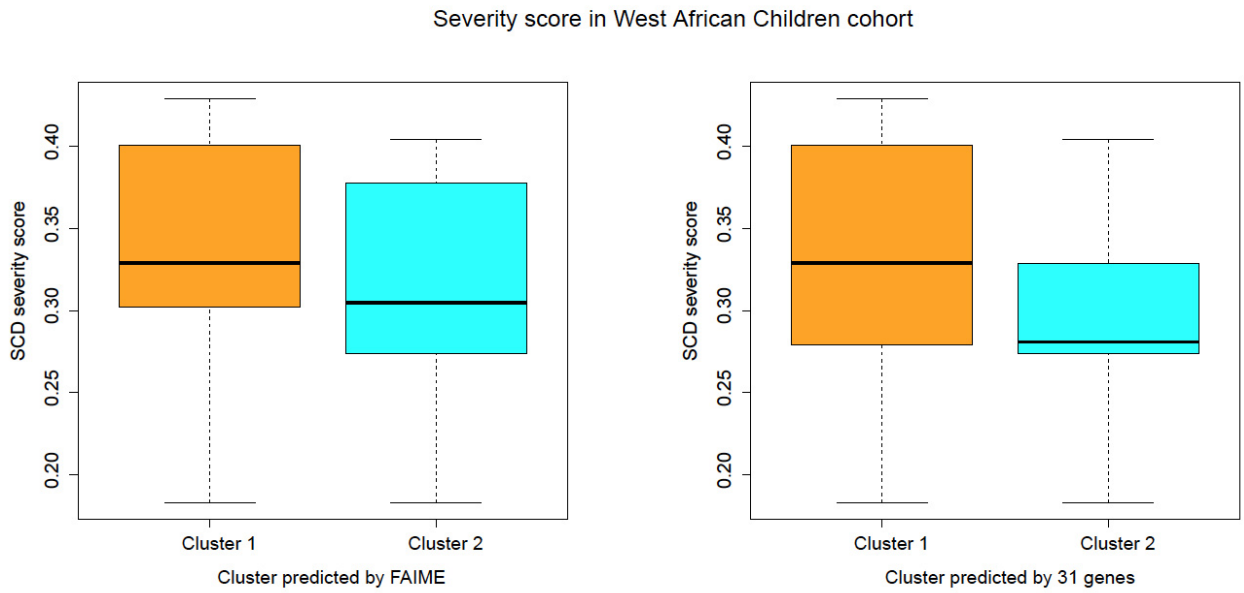
Supplemental Figure 3. Correlation of gene expression between RT-qPCR and microarray-derived profiling. A subset of 10 genes were further evaluated by RT-qPCR for levels of gene expression in SCD patients compared to non-SCD African American control subjects. Log_2-fold expression by qPCR was correlated against values derived by microarray of these 10 genes. The fold changes of gene expression determined by RT-qPCR highly correlated with those estimated by microarray profiling (Pearson r=0.99, *P* =7.628e-08), with the direction of gene expression alteration consistent between the 2 platforms for 9 out of 10 genes.

Supplemental Figure 4. Validation of primer specificity. RT-qPCR product of one SCD sample for each primer-pair from the two plates were run on 3% agarose gel with SYBR green I detection dye. Lanes left to right (1) 100 bp DNA ladder [11] 25 bp Bioline hyperladder (3) *RPS11* -78 bp (4)*FECH* – 95 bp, (5) *GYPA* – 110 bp (6) *EPB42* 134 bp (7) *GOLA8B*-153 bp (8) *DOCK9* – 79 bp (9) *RPS11* (Plate 2) – 38 bp (10) *SELENBP1* -109 bp (11) *ALAS2* – 80 bp (12) *SLC4A1* – 80 bp (13) *NELL2* – 73 bp (14) *LEF1* – 196 bp (15) 25 bp Bioline hyperladder.

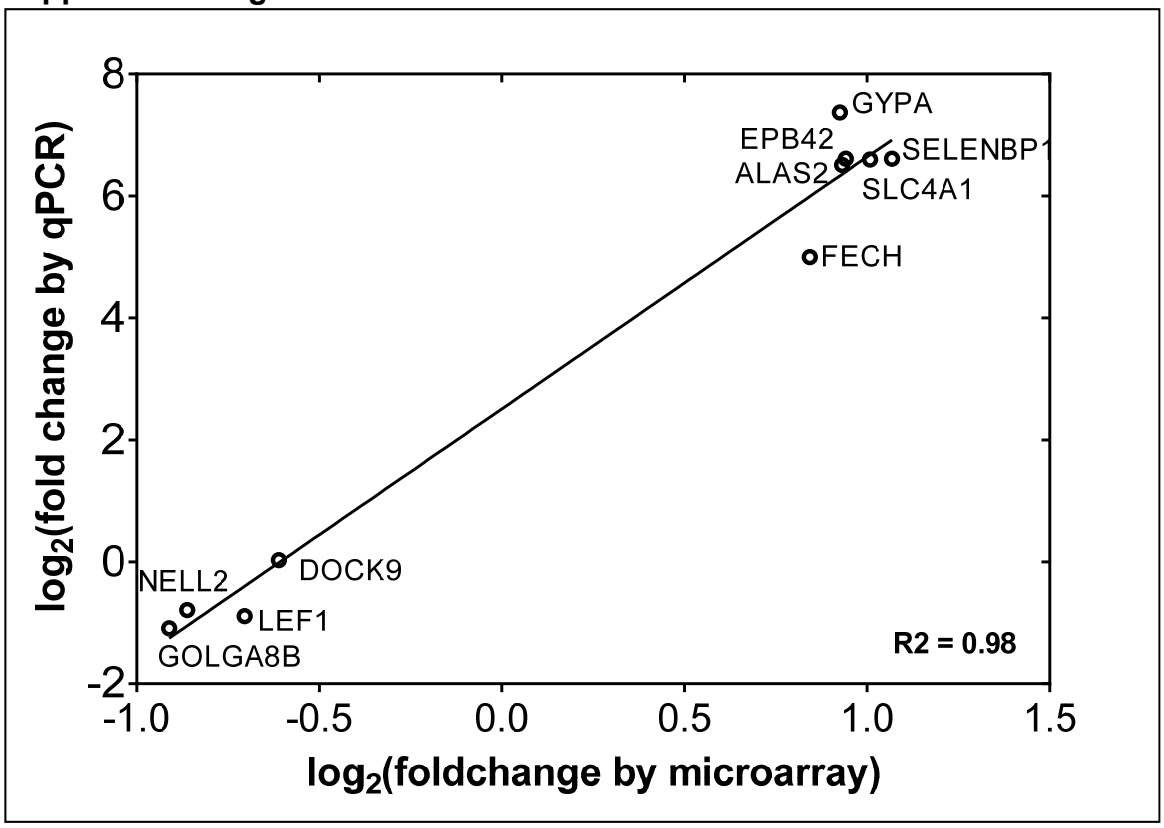
Supplemental Figure 1



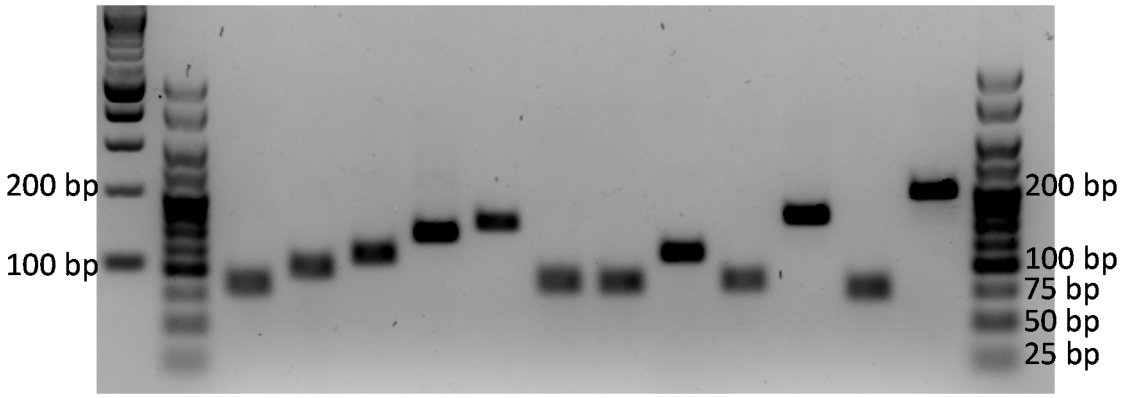
Supplemental Figure 2



Supplemental Figure 3



Supplemental Figure 4



Supplemental Table 1. Clinical and demographic characteristics of the Training and Testing cohorts

	Training Cohort (n=172)	Testing Cohort (n=78)	P Value
Age (yrs)	35.6 ± 11.9	39.3 ± 10.4	0.014
Female / Male (n=)	94/78	45/44	0.6
Hemoglobin SS	147	60	0.106
Hemoglobin SC	17	11	0.387
Hemoglobin SB	8	5	0.551
BMI (kg/m^2)	24.9 ± 5.6	23.1 ± 3.8	0.012
HU Therapy (n=)	98	43	0.785
ACS (n=)	144	43	
WBC ($10^{-3}/mm^3$)	9.7 ± 3.6	9.7 ± 3.6	0.97
Hemoglobin (g/dl)	8.8 ± 1.7	8.8 ± 1.8	0.98
Creatinine (mg/dl)	1.10 ± 1.28	0.97 ± 1.23	0.44
HgF (g/dl)	7.6 ± 6.6	6.3 ± 5.2	0.27
LDH (U/L)	422 ± 210	356 ± 162	0.02
Bilirubin (mg/dL)	3.1 ± 2.5	2.6 ± 1.6	0.13
AST (U/L)	48.7 ± 50.2	39.8 ± 17.7	0.047
Peak TRV (≥ 2.5 m/s)	32.9%	54.7%	0.0014
Peak TRV (≥ 3.0 m/s)	8.3%	10.7%	0.624
Peak TRV (m/s)	1.88 ± 0.88	2.39 ± 0.47	<0.001

Supplemental Table 2. Cox Regression Model in the Training Cohort.

	Hazard Ratio (95% CI)	P Value
Univariate Analysis		
Cluster 1	7.385 (1.653, 33.000)	8.85E-03
Tricuspid Regurgitation Velocity (≥ 2.5 m/s)	2.598 (0.873, 7.73)	0.086
White Blood Cell ($10^{-3}/mm^3$)	1.145 (1.012, 1.295)	0.031
Hemoglobin F (%)	0.834 (0.709, 0.981)	0.028
Age (years)	1.009 (0.966, 1.053)	0.691
Gender (Male)	1.167 (0.409, 3.329)	0.772
History of Acute Chest Syndrome	2.532 (0.331, 19.36)	0.371
Absence of Hydroxyurea Use	1.051 (0.369, 2.998)	0.925
Genotype (SS and S-Bthal vs SC)	0.232 (0.185, 10.820)	0.248

Supplemental Table 3. Cox Regression Model in the Testing Cohort.

	Hazard Ratio (95% CI)	P Value
Univariate Analysis		
Cluster 1	3.957 (1.089, 14.380)	0.037
Age (years)	1.057 (1.016, 1.099)	5.62E-03
Gender (male)	4.600 (1.260, 16.790)	0.021
White Blood Cell ($10^{-3}/mm^3$)	1.175 (0.923, 1.495)	0.192
Tricuspid Regurgitation Velocity (≥ 2.5m/s)	1.748 (0.566, 5.401)	0.332

Supplemental Table 4. 31-Gene Signature for Survival.

Gene Symbols (Training cohort)	Gene Names	Fold-Change (Training)	Adjusted *P* values	Fold-Change (Testing)	Adjusted *P* values
SELENBP1	selenium binding protein 1	2.74	2.02E-09	2.10	0.006002928
SLC4A1	solute carrier family 4, member 1	2.89	8.06E-08	2.01	0.008139662
EPB42	erythrocyte membrane protein band 4.2	2.93	2.35E-09	1.92	0.002667278
ALAS2	5'-aminolevulinate synthase 2	2.76	4.50E-07	1.91	0.031355692
GYPA	glycophorin A	2.83	5.45E-13	1.90	0.009538749
FECH	ferrochelatase	2.57	2.07E-09	1.79	0.004552921
PHOSPHO1	phosphatase, orphan 1	1.74	2.33E-06	1.76	0.000814154
RUNDC3A	RUN domain containing 3A	2.62	5.31E-09	1.74	6.50E-05
CLIC2	chloride intracellular channel 2	2.54	5.59E-13	1.72	0.000504417
SNCA	synuclein, alpha	2.04	5.61E-10	1.70	0.008835963
HBM	hemoglobin, mu	3.32	5.02E-11	1.67	0.000720432
SLC25A37	solute carrier family 25, member 37	2.13	9.07E-09	1.67	0.007886577
HEMGN	hemogen	2.31	2.10E-11	1.65	0.003736576
BLVRB	biliverdin reductase B	2.01	3.09E-14	1.63	1.73E-05
TMOD1	tropomodulin 1	2.18	1.29E-09	1.62	0.003152281
BPGM	2,3-bisphosphoglycerate mutase	2.53	7.65E-10	1.58	0.018371109
SPTB	spectrin, beta, erythrocytic	1.67	1.51E-07	1.57	0.00100284
FAM46C	family with sequence 46, member C	2.12	1.14E-08	1.57	0.00772834
SLC7A5	solute carrier family 7, member 5	1.93	9.22E-08	1.56	0.000901199
SLC25A39	solute carrier family 25, member 39	2.07	5.67E-09	1.55	0.003119367
SLC1A5	solute carrier family 1, member 5	1.62	2.14E-07	1.54	3.41E-05
BSG	basigin (Ok blood group)	1.63	5.89E-10	1.54	1.01E-05
SOX6	sex determining region Y-box 6	1.65	5.65E-13	1.54	0.000211996
BCL2L1	BCL2-like 1	1.84	2.28E-08	1.53	0.006323025
ANK1	ankyrin 1, erythrocytic	2.08	5.55E-08	1.52	0.006626359
E2F2	E2F transcription factor 2	1.74	8.32E-08	1.52	0.000183178
PDZK1IP1	PDZK1 interacting protein 1	1.57	2.57E-08	1.50	0.010226558
DOCK9	dedicator of cytokinesis 9	0.64	3.35E-20	0.65	8.18E-11
LEF1	lymphoid enhancer-binding factor 1	0.61	6.14E-13	0.61	1.67E-08
NELL2	neural EGFL like 2	0.66	3.08E-10	0.55	8.49E-09
GOLGA8B	golgin A8 family, member B	0.66	1.57E-18	0.53	5.22E-08

Supplemental Table 5. Composite Risk Score Evaluation. The table reflects cox hazard ratios and *P* values for each category of risk score stratified by clinical variables alone, genomic clustering alone, or a combined risk score. High risk is defined as ≥4 vs low risk defined as <4 in the combined risk score. High risk is defined as ≥3 vs low risk defined as <3 in the risk score derived by clinical risk factors alone.

	Combined Risk Score (Clinical + Genomics)	Risk score (Clinical only)	Risk score (Genomics Cluster only)
Training Cohort			
Cox Hazard Ratio	8.27	7.53	7.4
P value	1.18e-3	8.22E-03	8.85E-03
Testing Cohort			
Cox Hazard Ratio	4.84	1.888	4.0
P value	0.049	0.422	0.037

Supplemental Table 6. RT-qPCR was used to verify the relative gene expression of ten candidate genes standardized to *RPS11* gene.

	Normalized gene expression relative to controls ($2^{-(ddCt)}$)			Standard curve properties	
	SCD (n = 8)	Controls (n = 6)	*P* value	Slope	R^2
		Plate 1			
LEF1	0.55 (0.36-0.85)	1 (0.79-1.27)	0.007	-3.3165	0.9987
NELL2	0.56 (0.38-0.83)	1 (0.74-1.35)	0.009	-3.325	0.9988
SLC4A1	97 (30.6-307.4)	1 (0.1-8.1)	0.002	-3.3098	0.9991
ALAS2	87.6 (29.6-259)	1 (0.1-10.4)	0.004	-3.3146	0.9987
SELENBP1	91.4 (28.9-288.7)	1 (0.1-8.5)	0.002	-3.3096	0.9975
RPS11				-3.3286	0.9981
		Plate 2			
DOCK9	1.05 (0.65-1.71)	1 (0.72-1.39)	0.815	-3.2837	0.9942
GOLGA8B	0.45 (0.25-0.8)	1 (0.6-1.68)	0.018	-3.346	0.9994
EPB42	90.5 (27.9-293.4)	1 (0.1-7.9)	0.002	-3.1924	0.9991
GYPA	173.7 (46.7-647)	1 (0.2-6)	0.0002	-3.3594	0.9989
FECH	29.5 (9.5-91.4)	1 (0.4-2.4)	3.88E-05	-3.3205	0.9994
RPS11				-3.3801	0.9993

P value is calculated based on the delta Ct of SCD and controls using two tail student's t-test with unequal variance.

Supplemental Table 7. Primer sequences (forward, F and reverse, R) used for validation by RT-qPCR.

Name	Primer sequence	Product Size (bp)
RPS11_F	GTCCAGATCGGTGACATCGT	78
RPS11_R	GACCTTGAGCACGTTGAAGC	
FECH_F	GGATTTCGGTACGTCCATCCT	95
FECH_R	TGTGGATACTGTGTGAAAGCAAT	
GYPA_F	ACAACTTGCCCATCATTTCTCTG	110
GYPA_R	TCAGTCGGCGAATACCGTAAG	
EPB42_F	CCTCAGTCAGCCTCCAGAAC	134
EPB42_R	GGCACACATGGTGTTTTCAG	
GOLGA8B_F	CCAAATAATGTGGCTAATAGTGG	153
GOLGA8B_R	CAG TAT TCA TAT TTT AAA ATG TTT TAA	
DOCK9_F	TCAAGGAGCCATCAGGCAAG	79
DOCK9_R	TGATCCGCTGGACGTTTTCA	
SELENBP1_F	TCATCTCCTCTCGCATCTATGTG	109
SELENBP1_R	AAGGCCAGTTCGCACTTGG	
ALAS2_F	CTGCCAGGGTGCGAGATT	80
ALAS2_R	TTGGCTGCTCCACTGTTACG	
SLC4A1_F	AAGAAAAGGCCTTGGTTGGT	160
SLC4A1_R	GGAAGGGTGCTAGCAGAGTG	
NELL2_F	CGTGACATCCTGAACCCTGG	73
NELL2_R	TGCACCAGTGAAGCTAGAGG	
LEF1_F	CATTCCCAACGTGCAAAGCC	196
LEF1_R	GCAGTAGACGAAAGAGGGGT	

REFERENCES

1. Desai AA, Zhou T, Ahmad H, et al. A Novel Molecular Signature for Elevated Tricuspid Regurgitation Velocity in Sickle Cell Disease. *Am J Respir Crit Care Med.* 2012.
2. Maienschein-Cline M, Lei Z, Gardeux V, et al. ARTS: automated randomization of multiple traits for study design. *Bioinformatics.* 2014;30(11):1637-1639.
3. Affymetrix. Exon array background correction. *Affymetrix Whitepaper.* 2005.
4. Asosingh K, Aldred MA, Vasanji A, et al. Circulating angiogenic precursors in idiopathic pulmonary arterial hypertension. *Am J Pathol.* 2008;172(3):615-627.
5. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics.* 2007;8(1):118-127.
6. Dennis G, Jr., Sherman BT, Hosack DA, et al. DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol.* 2003;4(5):P3.
7. Monti S, Tamayo P, Mesirov J, Golub T. Consensus Clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. *Machine Learning Journal.* 2003;52(1-2):91-118.
8. Yang X, Li H, Regan K, Li J, Huang Y, Lussier YA. Towards Mechanism Classifiers: Expression-anchored Gene Ontology Signature Predicts Clinical Outcome in Lung Adenocarcinoma Patients. *AMIA Annu Symp Proc.* 2012;2012:1040-1049.
9. Yang X, Regan K, Huang Y, et al. Single sample expression-anchored mechanisms predict survival in head and neck cancer. *PLoS Comput Biol.* 2012;8(1):e1002350.
10. Livak KJ, Schmittgen TD. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods.* 2001;25(4):402-408.
11. Smith CM, 2nd, Hebbel RP, Tukey DP, Clawson CC, White JG, Vercellotti GM. Pluronic F-68 reduces the endothelial adherence and improves the rheology of liganded sickle erythrocytes. *Blood.* 1987;69(6):1631-1636.