Supplemental Text and Figures

Table S1, related to Figure 1. Demographic and clinical characteristics and therapeutic exposures for cases and controls in training and test sets

	Training Set			Test Set		
	t-MDS/AML cases (n=18)	Controls (n=37)	p-value	t-MDS cases (n=16)	Controls (n=20)	p-value
Primary diagnosis						
HL	1(5.5%)	3 (8.1%)		9 (56%)	11 (55%)	
NHL	17 (94.5%)	34 (91.9%)		7 (44%)	9 (45%)	
Mean age at diagnosis	48.0	47.2	0.76	46.5	43.7	0.14
Male sex	8 (44.4%)	26 (70.3%)	0.08	13(81%)	8 (40%)	0.01
Race/ ethnicity						
Caucasians	10 (55.6%)	24 (64.9%)		11 (69%)	13 (65%)	
Hispanics	7 (38.9%)	9 (24.3%)		4 (25%)	6 (30%)	
African-Americans	1 (5.5%)	3 (8.1%)		1 (6%)	1 (5%)	
Asians/Other	0	1 (2.7%)		0	0	
Mean age at HCT	49.4	49.4		49.8	45.0	0.02
Stem cell source			0.60			1.00
Peripheral stem cells	17 (94%)	36 (97%)		15(94%)	20 (100%)	
Bone marrow+PBSC	1 (6%)	1 (3%)		1 (6%)	0	
Tandem HCTs	1	0	0.25	1 (6%)	4 (20%)	0.25
Number of stem cell collections						
Mean (SD)	7.1 (4.6)	5.1 (3.6)	0.08	7.2 (4.3)	6.2 (4.3)	0.58
Median (range)	6.0 (1-17)	4.0 (1-15)		6.5 (2-17)	4.5 (1-17)	
Stem cell mobilization (yes)						
G-CSF	18 (100%)	37 (100%)	--	16 (100%)	13 (65%)	0.01
Etoposide	2 (11%)	0 (0%)	0.08	3 (19%)	2 (10%)	0.67
Cyclophosphamide	7(39%)	7 (20%)	0.09	3 (19%)	6 (30%)	0.68
Other (taxol)	2 (11%)	0	0.11	0	0	
Stem cell dose infused						
Mean (SD)	5.0 (3.7)	6.6 (5.2)	0.16	5.2 (3.7)	6.0 (2.9)	0.46
Median (range)	3.5 (0.9-16.2)	5.2 (2-26.9)		4.2 (1.9-16.1)	5.0 (2.3-14.2)	
HCT conditioning						
Cyclophosphamide						
Yes	16 (89%)	34 (92%)	1.00	14 (88%)	15 (75%)	0.39
Mean (SD)	3818 (892)	4137 (460)	0.09	4314 (437)	4129 (403)	0.13
Etoposide						
Yes	18 (100%)	37 (100%)		15 (94%)	18 (90%)	1.00
Mean (SD)	2333 (750)	2318 (607)	0.99	2469 (526)	2039 (782)	0.13
Total body irradiation						
Yes	11 (61%)	19 (51%)	0.58	8 (50%)	6 (30%)	0.36
Mean	1200	1200		1230	1200	0.18
Therapeutic exposures (pre-HCT + post-HCT)						
Radiation (Yes)	15(83%)	26 (70%)	0.53	13(81%)	11 (55%)	0.11
Alkylating agent score						
1,2	11 (61%)	15 (41%)	0.34	3 (19%)	7 (35%)	0.31
3, 4, 5	7 (39%)	22 (59%)		13 (81%)	13 (65%)	
Topoisomerase II inhibitor score						
2	5 (28%)	9 (24%)	0.72	4 (25%)	8 (40%)	0.40
3, 4	13 (72%)	28 (76%)		12 (75%)	12 (60%)	

Table S2, related to Figure 1. Clinical characteristics of t-MDS/AML after autologous HCT for lymphoma for the training set

Serial No.	*Primary Diagnosis*	*Age at HCT*	*Sex*	*Race/ ethnicity*	*Stem cell mobilization*	*Stem cell dose infused*	*Conditioning regimen*	*Time from aHCT to t-MDS/AML (days)*	*Cytogenetics*	*Morphology*	*WBC*	*HB*	*PLT*	*Cytopenia*
8	HL	32	M	H	GCSF, CTX,VP16	3.04	Cytoxan, VP16, Melphan, TBI	913	46,XY [20 of 20 cells]	MDS	2.4	12.1	189	Leukopenia
11	NHL	50	M	W	GCSF	4.59	BCNU, Cytoxan, VP16	178	46,XY,del(7)(q22q36) [2 of 22 cells] 46,XY,t(7;16)(p13;p13.1) [1] 46,XY,?t(Y;16)(p11.2;p11.2), t(2;6)(p23;p21.1)	RAEB	2.4	13.9	132	Leukopenia; Thrombocytopenia
24	NHL	56	M	W	GCSF		Cytoxan,VP16, TBI	1351	45,XY,inv(6)(p11.2p25), -7, t(12;14)(p11.2;p11.2)[3 of 30 cells] 46,XY,t(2;11)(q37;q13) [1]	MDS	3.7	10.5	258	Leukopenia; Anemia
25	NHL	54	M	W	GCSF	7.69	Cytoxan, VP16, TBI	1076	45,X,-Y [5 of 20 cells] 47,XY,+8 [4 of 20 cells] 46,XY,del(20)(q11.2q13.1) [2 of 20 cells]	MDS	6.4	13.2	111	Thrombocytopenia
61	NHL	53	F	W	GCSF	5.77	BCNU, Cytoxan, VP16	1912	46,XX,del(20)(q11.2q13.1) [6 of 30 cells]	MDS	3.2	12.2	183	Leukopenia
62	NHL	49	M	H	GCSF	4.35	Cytoxan, VP16, TBI	1813	46,XY,der(4)del(4)(q21q25)t(4;22)(q25;q11.1), t(7;12)(q34;q15),del(13)(q12q15),der(22)t(4;22) [3 of 30 cells] 46,XY,t(1;11)(q21;q23) [1]	MDS	2.9	13.9	166	Leukopenia
101	NHL	50	F	W	GCSF, CTX,	2.49	Cytoxan,VP16, TBI	1171	46,XX,t(7;11)(p13;q23)[20]	AML	18	10.5	10	Leukopenia,, Pancytopenia
111	NHL	51	F	B	GCSF, CTX,		Cytoxan, VP16, TBI	708	45,XX,-7 [20 of 20 cells]	AML	0.3	7.1	12	Pancytopenia
120	NHL	50	F	W	GCSF	2.88	Cytoxan, VP16, TBI	1098	Nonclonal Aberration: 90, XXX, -X, +3 ,+4, -5, -5, +6, -9, -9, -10, +11, +12, +13, +14, -15, -16, -18 [1]	MDS	3.3	12	47	Leukopenia Thrombocytopenia
125	NHL	52	F	W	GCSF	3.75	Cytoxan,VP16 TBI	1514	46,XX,del(20)(q11.2q13.1) [10 of 30 cells]	MDS	7.4	12.7	206	Normal
137	NHL	65	F	H	GCSF, CTX,		BCNU, Arac, VP16, Melphalan	178	46,XX,t(9;11)(p22;q23),add(19)(p13.3) [30 of 30 cells]	AML	32.3	9.4	17	Anemia, Thrombocytopenia
155	NHL	42	M	W	GCSF, CTX,	2.7	Cytoxan,VP16, TBI	239	46,XY,t(2;12)(p15;q13),del(13)(q12q21),t(20;21)(p11.2;q22) [2 of 30 cells]	MDS	1.9	10.3	33	Pancytopenia
168	NHL	28	F	W	GCSF	10.8	Cytoxan,VP16, Rituxan IN111, Y90	728	46,XX,del(20)(q11.2q13.1) [4 of 20 cells]	MDS	6.2	13.6	262	Normal
231	NHL	53	F	H	GCSF	5.13	Cytoxan,VP16, Rituxan IN111, Y90	204	46,XX,del(13)(q12q22) [9 of 21 cells]	NA	3.8	12.2	264	Normal
5005	NHL	67	M	W	GCSF	3.34	BCNU, Cytoxan, VP16	926	45,XY,-7 [8 of 20 cells] 47,XY,+15 [2 of 20 cells]	MDS	3.7	12.5	149	Normal
5036	NHL	58	F	H	GCSF, CTX, Taxol	2.68	Cytoxan, VP16, Rituxan, TBI	524	46,XX [20 of 20 cells]	MDS	2.9	10.7	29	Pancytopenia
5040	NHL	29	F	H	GCSF, CTX,VP16		BCNU, Ara-C, VP16, Melphalan	1191	46,XX,t(9;11)(p22;q23) [19 of 20 cells]	AML	4.5	12.2	26	Thrombocytopenia
5042	NHL	52	M	H	GCSF	8.13	Cytoxan, VP16, TBI	1174	42-46,XY, t(1;8)(p34.2;q24.3),-3, -5,-7, ?ins(11;?12)(q23;q13q21.2), add(17)(p11.2)	RAEB	1.4	8.6	40	Pancytopenia

Abbreviations: M=male; F=Female; W=white; H=Hispanic; B=black

Table S3, related to Figure 1 and Table 1. Gene sets, canonical pathways and Gene Ontology categories enriched in genes differentially expressed in PBSC CD34+ cells between t-MDS/AML cases and controls (provided as a separate Excel file)

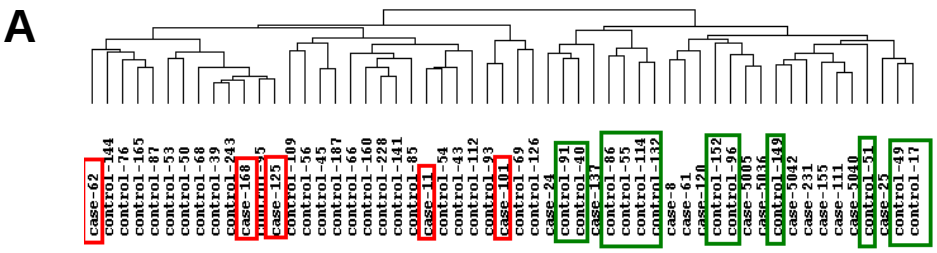
Table S4, related to Figure 1. GSEA analysis of genes differentially expressed in PBSC from cases and controls after removing the 2 cases with transient t-MDS/AML and their controls

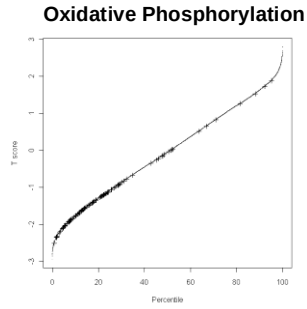
Increased in t-MDS/AML PBSC

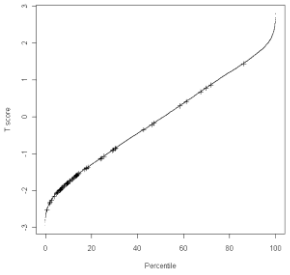
NAME	NES	FDR q-val
HSA04080_NEUROACTIVE_LIGAND_RECEPTOR_INTERACTION	2.650688	<0.001
GPCRDB_CLASS_A_RHODOPSIN_LIKE	2.611216	<0.001
GPCRS_CLASS_A_RHODOPSIN_LIKE	2.443164	<0.001
HSA01430_CELL_COMMUNICATION	2.365203	0.002867
MOREAUX_TACI_HI_VS_LOW_UP	2.317167	0.003502
PEPTIDE_GPCRS	2.313133	0.002918
HUMAN_TISSUE_LIVER	2.285419	0.002501
CARIES_PULP_DN	2.230311	0.003687
MONOAMINE_GPCRS	2.170755	0.00667
HALMOS_CEBP_DN	2.043267	0.017537
GAMMA_HEXACHLOROCYCLOHEXANE_DEGRADATION	1.97882	0.025142
UVC_HIGH_ALL_UP	1.957079	0.02656
IGF_VS_PDGF_UP	1.95353	0.025448
GH_EXOGENOUS_MIDDLE_UP	1.935157	0.026592
VEGF_HUVEC_30MIN_UP	1.872544	0.040946
BOQUEST_CD31PLUS_VS_CD31MINUS_DN	1.845108	0.051132
GERY_CEBP_TARGETS	1.828249	0.053361
LEE_MYC_DN	1.824483	0.05173
VEGF_MMMEC_6HRS_UP	1.818545	0.04997
STAEGE_EFTS_UP	1.811746	0.050798
GH_EXOGENOUS_ANY_UP	1.794856	0.054716
ADIP_DIFF_CLUSTER2	1.79459	0.052229
LVAD_HEARTFAILURE_UP	1.733005	0.077574
SCHURINGA_STAT5A_UP	1.728285	0.075347
VEGF_MMMEC_12HRS_UP	1.720316	0.078354
MOREAUX_TACI_HI_IN_BMPC	1.717426	0.076952
FERRANDO_T_CELL_DIFFERENTIATION_PATHWAY	1.711202	0.07813
YAO_P4_KO_VS_WT_UP	1.708963	0.075973
STRIATED_MUSCLE_CONTRACTION	1.704732	0.076221
HSA05217_BASAL_CELL_CARCINOMA	1.697705	0.079466
HOX_GENES	1.692072	0.079003
CMV_HCMV_6HRS_DN	1.684333	0.081244
HINATA_NFKB_DN	1.672483	0.084999
EMT_DN	1.667348	0.08633
NUCLEAR_RECEPTORS	1.66686	0.084048
CMV_HCMV_TIMECOURSE_8HRS_DN	1.641662	0.096376
RORIE_ES_PNET_DN	1.623858	0.1069
TSADAC_PANC50_UP	1.611256	0.114336
KANG_TERT_DN	1.598205	0.122511
HSA04742_TASTE_TRANSDUCTION	1.595793	0.122301
FRASOR_ER_UP	1.577758	0.135234
HSA00602_GLYCOSPHINGOLIPID_BIOSYNTHESIS_NEO_LACTOSERIES	1.573269	0.134457
IDX_TSA_UP_CLUSTER1	1.556179	0.1491
JECHLINGER_EMT_DN	1.554442	0.147228
TAKEDA_NUP8_HOXA9_6H_UP	1.551692	0.146235
HSA04340_HEDGEHOG_SIGNALING_PATHWAY	1.550069	0.144326
GH_EXOGENOUS_LATE_UP	1.543701	0.145139
HSA04512_ECM_RECEPTOR_INTERACTION	1.531886	0.154135
CALCIUM_REGULATION_IN_CARDIAC_CELLS	1.530181	0.153304
BRENTANI_TRANSCRIPTION_FACTORS	1.51914	0.16253

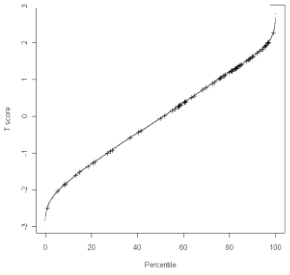
Increased in control PBSC

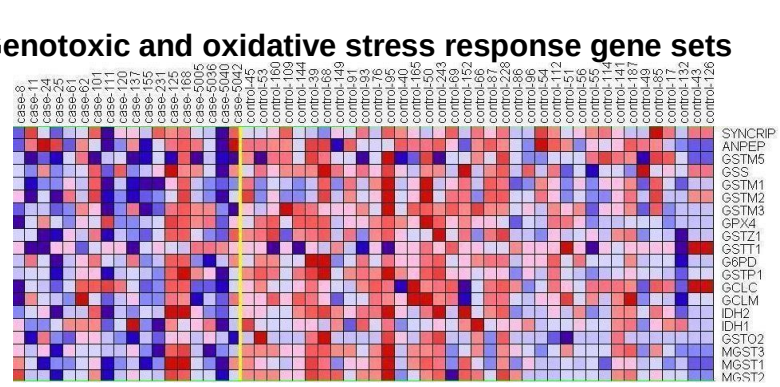
NAME	NES	FDR q-val
ELECTRON_TRANSPORT_CHAIN	-3.15933	<0.001
MOOTHA_VOXPHOS	-3.0799	<0.001
HSA00190_OXIDATIVE_PHOSPHORYLATION	-3.04706	<0.001
HUMAN_MITODB_6_2002	-2.99652	<0.001
MITOCHONDRIA	-2.99255	<0.001
PROTEASOMEPATHWAY	-2.88423	<0.001
RIBOSOMAL_PROTEINS	-2.87936	<0.001
HSA03010_RIBOSOME	-2.8533	<0.001
TARTE_PLASMA_BLASTIC	-2.83713	<0.001
CANCER_UNDIFFERENTIATED_META_UP	-2.75819	<0.001
PENG_RAPAMYCIN_DN	-2.71085	<0.001
CANTHARIDIN_DN	-2.71076	<0.001
MOREAUX_TACI_HI_IN_PPC_UP	-2.69927	<0.001
MOREAUX_TACI_HI_VS_LOW_DN	-2.69619	<0.001
OXIDATIVE_PHOSPHORYLATION	-2.6729	<0.001
HCC_SURVIVAL_GOOD_VS_POOR_DN	-2.66091	<0.001
TCA	-2.65011	<0.001
UVB_SCC_UP	-2.63461	<0.001
HEARTFAILURE_ATRIA_DN	-2.59439	<0.001
PENG_GLUTAMINE_DN	-2.57008	<0.001
BRCA1_OVEREXP_DN	-2.55092	<0.001
TRNA_SYNTHETASES	-2.53694	<0.001
ROME_INSULIN_2F_UP	-2.52322	<0.001
KREBS_TCA_CYCLE	-2.52253	<0.001
JAIN_NEMO_DIFF	-2.49564	<0.001
IDX_TSA_UP_CLUSTER3	-2.49481	<0.001
PGC	-2.48419	<0.001
CMV_IE86_UP	-2.47819	<0.001
RUTELLA_HEPATGFSNDCS_UP	-2.47462	<0.001
PROTEASOME_DEGRADATION	-2.45868	<0.001
HSA03050_PROTEASOME	-2.45318	<0.001
CHANG_SERUM_RESPONSE_UP	-2.44659	<0.001
AMINOACYL_TRNA_BIOSYNTHESIS	-2.44393	<0.001
DNA_REPLICATION_REACTOME	-2.44173	<0.001
PRMT5_KD_UP	-2.44159	<0.001
SERUM_FIBROBLAST_CORE_UP	-2.43556	<0.001
WIELAND_HEPATITIS_B_INDUCED	-2.41563	<0.001
HSC_LATEPROGENITORS_SHARED	-2.41158	<0.001
HSC_LATEPROGENITORS_ADULT	-2.41153	<0.001
UVB_NHEK2_UP	-2.4094	<0.001
HDACI_COLON_CUR24HRS_UP	-2.40287	<0.001
HSC_INTERMEDIATEPROGENITORS_SHARED	-2.39969	<0.001
HSC_INTERMEDIATEPROGENITORS_ADULT	-2.38435	<0.001
IDX_TSA_UP_CLUSTER5	-2.37883	<0.001
RADAEVA_IFNA_UP	-2.37015	<0.001
FERRANDO_MLL_T_ALL_DN	-2.36861	2.55E-04
RCC_NL_UP	-2.36775	2.50E-04
HSC_LATEPROGENITORS_FETAL	-2.36063	2.45E-04
SHIPP_FL_VS_DLBCL_DN	-2.35459	2.40E-04
HSA00970_AMINOACYL_TRNA_BIOSYNTHESIS	-2.35246	2.35E-04

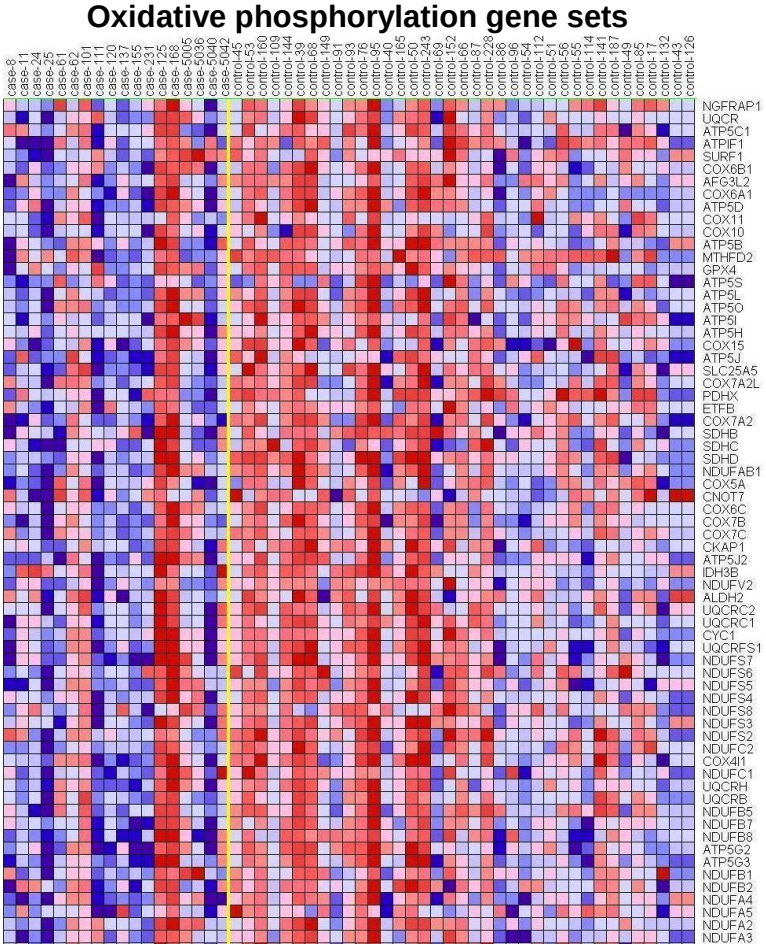

A
case-62
control-144
control-76
control-165
control-87
control-53
control-50
control-68
control-39
control-243
case-168
control-5
case-125
control-109
control-56
control-45
control-187
control-66
control-160
control-228
control-141
control-85
case-11
control-54
control-43
control-112
control-93
case-101
control-69
control-126
case-24
control-91
control-40
case-137
control-86
control-55
control-114
control-132
case-8
case-61
case-120
control-152
control-96
case-5005
case-5036
control-149
case-5042
case-231
case-155
case-111
case-5040
control-51
case-25
control-49
control-17


B
Oxidative Phosphorylation
T score
Percentile


Ribosomal
T score
Percentile


Cell Communication
T score
Percentile


C
Genotoxic and oxidative stress response gene sets
SYNCRIP
ANPEP
GSTM5
GSS
GSTM1
GSTM2
GSTM3
GPX4
GSTZ1
GSTT1
G6PD
GSTP1
GCLC
GCLM
IDH2
IDH1
GSTO2
MGST3
MGST1
MGST2


Oxidative phosphorylation gene sets
NGFRAP1
UQCR
ATP5C1
ATPIF1
SURF1
COX6B1
AFG3L2
COX6A1
ATP5D
COX11
COX10
ATP5B
MTHFD2
GPX4
ATP5S
ATP5L
ATP5O
ATP5J
ATP5H
COX15
ATP5J
SLC25A5
COX7A2L
PDHX
ETFB
COX7A2
SDHB
SDHC
SDHD
NDUFAB1
COX5A
CNOT7
COX6C
COX7B
COX7C
CKAP1
ATP5J2
IDH3B
NDUFV2
ALDH2
UQCRC2
UQCRC1
CYC1
UQCRFS1
NDUFS7
NDUFS6
NDUFS5
NDUFS4
NDUFS8
NDUFS3
NDUFS2
NDUFC2
COX4I1
NDUFC1
UQCRH
UQCRB
NDUFB5
NDUFB7
NDUFB8
ATP5G2
ATP5G3
NDUFB1
NDUFB2
NDUFA4
NDUFA5
NDUFA2
NDUFA3

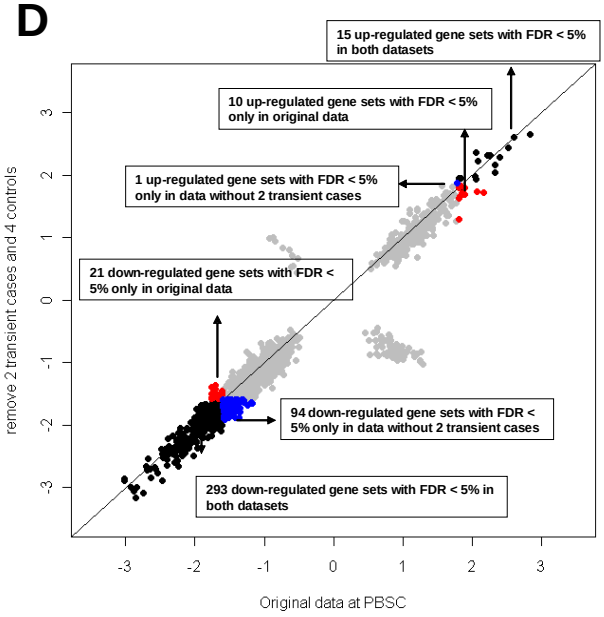

D
15 up-regulated gene sets with FDR < 5% in both datasets
10 up-regulated gene sets with FDR < 5% only in original data
1 up-regulated gene sets with FDR < 5% only in data without 2 transient cases
21 down-regulated gene sets with FDR < 5% only in original data
94 down-regulated gene sets with FDR < 5% only in data without 2 transient cases
293 down-regulated gene sets with FDR < 5% in both datasets
remove 2 transient cases and 4 controls
Original data at PBSC

Figure S1

Figure S1, related to Figure 1. (A) Hierarchical clustering for PBSC samples with all genes after filtering. Samples clustered into two large groups. In one cluster 5 cases clustered with 25 controls. In the other cluster 12 cases clustered with 11 controls. **(B) Altered gene expression in PBSC CD34+ cells from t-MDS/AML cases compared with controls.** Enrichment of oxidative phosphorylation, ribosomal and cell communication genes in t-MDS/AML PBSC CD34+ cells compared with controls is shown. **(C) Expression of leading edge genes in cases and controls.** Expression of leading edge genes for genotoxic and oxidative stress response gene sets (top), and oxidative phosphorylation gene sets (bottom) in PBSC CD34+ cells from individual cases and controls is shown. **(D) Gene expression in PBSC from t-MDS/AML cases after removal of 2 cases with transient disease and their matching controls compared with the original PBSC analysis**. 339 gene sets were significantly altered in the original analysis with FDR < 5%; and 403 gene sets were significantly altered in the analysis performed after removal of the 2 cases with transient t-MDS/AML with FDR <5%. The majority (308 gene sets) were common to both analyses.

Table S5, related to Figure 2 and Table 2. Gene sets, canonical pathways and Gene Ontology categories enriched in genes differentially expressed in BM CD34+ cells between t-MDS/AML cases and controls (provided as a separate Excel file)

Table S6, related to Figure 2. Gene sets enriched in differentially expressed gene between normal BM CD34+ cells and CD34+ cells from t-MDS/AML cases or controls (provided as a separate Excel file)

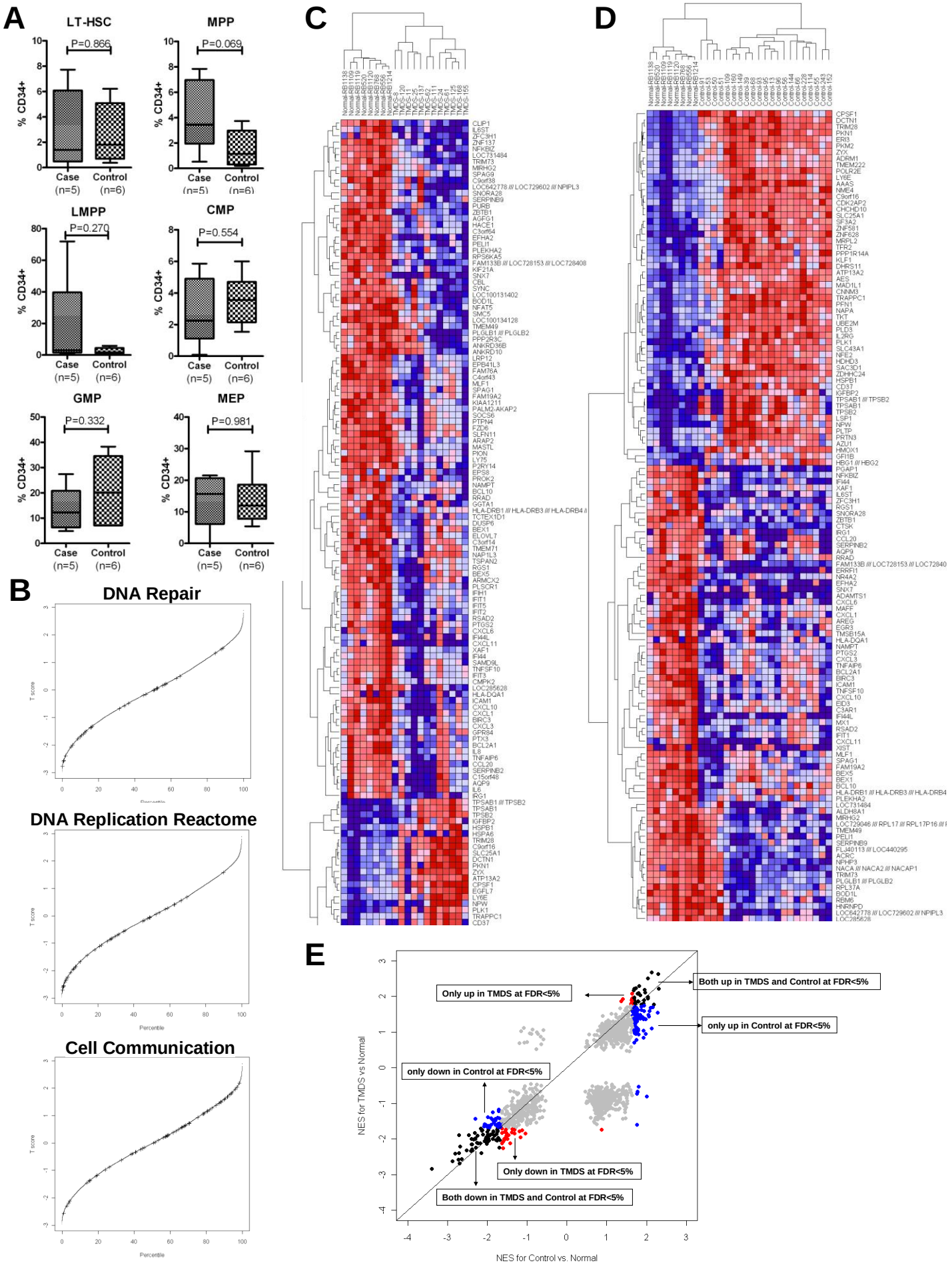

A
LT-HSC
P=0.866
MPP
P=0.069
LMPP
P=0.270
CMP
P=0.554
GMP
P=0.332
MEP
P=0.981
% CD34+
Case
Control
(n=5)
(n=6)
B
DNA Repair
DNA Replication Reactome
Cell Communication
T score
Percentile
C
D
E
Only up in TMDS at FDR<5%
Both up in TMDS and Control at FDR<5%
only up in Control at FDR<5%
only down in Control at FDR<5%
Only down in TMDS at FDR<5%
Both down in TMDS and Control at FDR<5%
NES for TMDS vs Normal
NES for Control vs. Normal

Figure S2, related to Figure 2. (A) Lack of difference in frequency of subpopulations of BM CD34+ cells from cases at the time of development of t-MDS/AML and controls. BM cells collected from cases at the time of development of t-MDS/AML and from controls at corresponding time were labeled with lineage markers (CD11b, CD14, Glycophorin A, CD2, CD7, CD10 and CD19), CD34, CD38, CD123, CD90 and CD45RA. CD34+ cells subpopulations including hematopoietic stem cells (HSC; Lin-CD34+CD38-CD45RA-CD90+), multipotent progenitors (MPP; Lin-CD34+CD38-CD45RA-CD90-), lymphoid-primed multipotent progenitors (LMPP; Lin-CD34+CD38-CD45RA+CD90-), common myeloid progenitors (CMP; Lin-CD34+CD38+CD45RA-CD123+), granulocyte-macrophage progenitors (GMP; Lin-CD34+CD38+CD45RA+CD123+) and megakaryocyte-erythroid progenitors (MEP; Lin-CD34+CD38+CD45RA-CD123-) were analyzed using multi-color flow cytometry. No significant difference in frequency of different CD34+ subpopulations was observed between cases and controls. Results represent the median, inter-quartile range and range of values. **(B) Altered gene expression in BM CD34+ cells from t-MDS/AML cases compared with controls.** Enrichment of DNA Repair, Cell communication, and DNA replication reactome genes in BM CD34+ cells at time of t-MDS/AML compared with controls. **(C) Gene expression in BM CD34+ cells from t-MDS/AML patients compared to normal BM CD34+ cells.** Gene expression in 8 normal samples and 12 t-MDS samples were compared using t-test: The heat map represents genes differentially expressed between the two groups using a threshold of FDR<5% and abs(logFC)>2 (n=133 genes). **(D) Gene expression in BM CD34+ cells from control NHL/HL patients that did not develop t-MDS/AML after aHCT compared to normal BM CD34+ cells.** Gene expression in 8 normal samples and 21 t-MDS samples were compared using t-test: The heat map represents genes differentially expressed between the two groups using a threshold of FDR<5% and abs(logFC)>2 (n=128 genes). **(E) Gene expression in BM CD34+ cells from NHL/HL patients that did or did not develop t-MDS/AML compared to normal BM CD34+ cells.** Normalized enrichment scores (NES) from GSEA analyses for altered gene sets in BM CD34+ cells from patients at time of t-MDS/AML compared with normal BM CD34+ cells were plotted against NES for altered gene sets in BM CD34+ cells from controls that did not develop t-MDS/AML after aHCT. Gene sets that were significantly upregulated or downregulated in both t-MDS/AML and control samples with FDR <5% are highlighted in black. Those altered only in t-MDS/AML samples in red; and those altered only in controls in blue.

Table S7, related to Figure 3. GSEA analysis of genes showing significant increase or decrease in expression over time from PBSC collection to development of t-MDS/AML in cases compared with controls over a similar time period

Enriched in Cases			Enriched in Controls		
Gene set	**NES**	**FDR q-val**	**Gene set**	**NES**	**FDR q-val**
HSA03010_RIBOSOME	2.88	<0.001	UVC_XPCS_ALL_DN	-2.74	<0.001
HSA00190_OXIDATIVE_PHOSPHORYLATION	2.73	<0.001	UVC_XPCS_4HR_DN	-2.73	<0.001
RIBOSOMAL_PROTEINS	2.71	<0.001	UVC_XPCS_8HR_DN	-2.71	<0.001
MOOTHA_VOXPHOS	2.53	<0.001	UVC_TTD_ALL_DN	-2.56	<0.001
ELECTRON_TRANSPORT_CHAIN	2.50	<0.001	UVC_TTD_4HR_DN	-2.50	<0.001
OXIDATIVE_PHOSPHORYLATION	2.31	<0.001	HADDAD_HSC_CD10_UP	-2.46	<0.001
HEMATOP_STEM_ALL_UP	2.16	0.001	UVB_NHEK1_DN	-2.46	<0.001
HUMAN_MITODB_6_2002	2.16	0.001	HADDAD_HPCLYMPHO_ENRICHED	-2.43	<0.001
LE_MYELIN_DN	2.07	0.007	GOLDRATH_CELLCYCLE	-2.33	<0.001
HOUSTIS_ROS	2.06	0.007	DOX_RESIST_GASTRIC_UP	-2.30	<0.001
NAKAJIMA_MCSMBP_MAST	2.05	0.008	CHEN_HOXA5_TARGETS_UP	-2.30	<0.001
MITOCHONDRIA	2.03	0.010	UVC_TTD-XPCS_COMMON_DN	-2.27	<0.001
TRNA_SYNTHETASES	2.00	0.011	RAC1PATHWAY	-2.25	<0.001
HSA03020_RNA_POLYMERASE	1.99	0.011	GLEEVECPATHWAY	-2.23	0.001
BRUNO_IL3_DN	1.98	0.014	ATMPATHWAY	-2.21	0.002
HYPOPHYSECTOMY_RAT_UP	1.95	0.016	SERUM_FIBROBLAST_CELLCYCLE	-2.17	0.003
UVB_NHEK2_UP	1.91	0.023	HSA05213_ENDOMETRIAL_CANCER	-2.14	0.004
IRS_KO_ADIP_UP	1.91	0.022	ZHAN_MM_CD138_PR_VS_REST	-2.12	0.003
HSA00480_GLUTATHIONE_METABOLISM	1.91	0.022	TARTE_BCELL	-2.08	0.004
HEARTFAILURE_VENTRICLE_DN	1.89	0.025	YU_CMYC_DN	-2.07	0.004
POMEROY_DESMOPLASIC_VS_CLASSIC_MD_UP	1.89	0.025	TAKEDA_NUP8_HOXA9_16D_DN	-2.07	0.005
INNEREAR_UP	1.87	0.027	REOVIRUS_HEK293_UP	-2.07	0.004
IDX_TSA_UP_CLUSTER5	1.86	0.030	UVB_NHEK1_C6	-2.07	0.004
AGED_MOUSE_HYPOTH_UP	1.84	0.035	UVC_HIGH_ALL_DN	-2.05	0.005
ATP_SYNTHESIS	1.83	0.034	TNFR2PATHWAY	-2.05	0.004
PENG_RAPAMYCIN_DN	1.83	0.033	NOVA2_KO_SPLICING	-2.04	0.004
IFN_BETA_GLIOMA_DN	1.83	0.032	FSH_OVARY_MCV152_DN	-2.02	0.006
HSA00710_CARBON_FIXATION	1.81	0.042	UVB_NHEK3_C5	-1.98	0.010
PASSERINI_OXIDATION	1.79	0.046	SIG_PIP3_SIGNALING_IN_B_LYMPHOCYTES	-1.97	0.010
GLUTATHIONE_METABOLISM	1.79	0.046	PTENPATHWAY	-1.97	0.011
UVB_NHEK1_C1	1.78	0.048	HSA05212_PANCREATIC_CANCER	-1.97	0.012
PGC	1.76	0.056	HSC_MATURE_SHARED	-1.97	0.012
HSA00100_BIOSYNTHESIS_OF_STEROIDS	1.75	0.064	HDACI_COLON_SUL48HRS_UP	-1.96	0.011
ELECTRON_TRANSPORT	1.75	0.065	STRESS_GENOTOXIC_SPECIFIC_DN	-1.95	0.012
FRUCTOSE_AND_MANNOSE_METABOLISM	1.74	0.064	LINDSTEDT_DEND_UP	-1.95	0.012
MYC_TARGETS	1.74	0.063	AKTPATHWAY	-1.95	0.012
GLYCOLYSIS_AND_GLUCONEOGENESIS	1.72	0.078	UVC_TTD_8HR_DN	-1.93	0.013
HSA00620_PYRUVATE_METABOLISM	1.72	0.076	TCRPATHWAY	-1.93	0.014
PHOTOSYNTHESIS	1.71	0.077	CMV_HCMV_TIMECOURSE_4HRS_DN	-1.93	0.014
ADIP_VS_FIBRO_UP	1.71	0.075	HSA05210_COLORECTAL_CANCER	-1.92	0.015
ALCALAY_AML_NPMC_UP	1.70	0.084	ET743_SARCOMA_24HRS_DN	-1.90	0.018
TYPE_III_SECRETION_SYSTEM	1.70	0.082	CMV_HCMV_TIMECOURSE_16HRS_DN	-1.90	0.019
SANA_IFNG_ENDOTHELIAL_DN	1.69	0.084	IGF1RPATHWAY	-1.90	0.018
FLAGELLAR_ASSEMBLY	1.69	0.083	APOPTOSIS	-1.89	0.019
MUNSHI_MM_UP	1.68	0.085	IFNA_HCMV_6HRS_UP	-1.89	0.018
ADIP_VS_PREADIP_UP	1.68	0.084	ROTH_HTERT_DIFF	-1.89	0.018
ROS_MOUSE_AORTA_UP	1.68	0.084	HSA04664_FC_EPSILON_RI_SIGNALING_PATHWAY	-1.88	0.019
ELECTRON_TRANSPORTER_ACTIVITY	1.68	0.083	HSA04662_B_CELL_RECEPTOR_SIGNALING_PATHWAY	-1.88	0.019
HIPPOCAMPUS_DEVELOPMENT_PRENATAL	1.66	0.089	CROONQUIST_IL6_RAS_DN	-1.88	0.019
HSA00052_GALACTOSE_METABOLISM	1.66	0.092	IL2RBPATHWAY	-1.86	0.020

Ninety-five gene sets were enriched in cases with FDR < 25%, and 288 gene sets were enriched in controls. Abbreviations: GSEA=gene set enrichment analysis; NES=normalized enrichment score; FDR=false discovery rate

Table S8, related to Figure 4. Clinical characteristics of t-MDS/AML cases in test sets and comparison of clinical and demographic characteristics and therapeutic exposures of t-MDS/AML cases in the training and test sets (provided as a separate Excel file)

Table S9, related to Figure 4. Gene sets enriched in PBSC CD34+ cells from t-MDS/AML cases in test set

Enriched in case			Enriched in controls		
Gene set	**NES**	**FDR q-val**	**Gene set**	**NES**	**FDR q-val**
HSA04080_NEUROACTIVE_LIGAND_RECEPTOR_INTERACTION	3.89	<0.001	MOREAUX_TACI_HI_VS_LOW_DN	-4.85	<0.001
GPCRDB_CLASS_A_RHODOPSIN_LIKE	3.35	<0.001	FLECHNER_KIDNEY_TRANSPLANT_WELL_UP	-4.68	<0.001
MOREAUX_TACI_HI_VS_LOW_UP	3.02	<0.001	TARTE_PLASMA_BLASTIC	-4.40	<0.001
PEPTIDE_GPCRS	2.86	<0.001	UVC_XPCS_8HR_DN	-4.35	<0.001
HSA01430_CELL_COMMUNICATION	2.86	<0.001	HCC_SURVIVAL_GOOD_VS_POOR_DN	-4.29	<0.001
GPCRS_CLASS_A_RHODOPSIN_LIKE	2.83	<0.001	UVC_TTD_4HR_DN	-4.23	<0.001
MONOAMINE_GPCRS	2.62	<0.001	ET743_SARCOMA_DN	-4.15	<0.001
GH_EXOGENOUS_MIDDLE_UP	2.49	<0.001	HSA03010_RIBOSOME	-4.06	<0.001
IGF_VS_PDGF_UP	2.47	<0.001	BRCA1_OVEREXP_UP	-4.01	<0.001
HSA04950_MATURITY_ONSET_DIABETES_OF_THE_YOUNG	2.47	<0.001	UVC_XPCS_ALL_DN	-4.00	<0.001
HSA04020_CALCIUM_SIGNALING_PATHWAY	2.46	<0.001	UVC_TTD_ALL_DN	-3.95	<0.001
GPCRDB_OTHER	2.38	<0.001	HUMAN_MITODB_6_2002	-3.93	<0.001
GH_EXOGENOUS_ANY_UP	2.37	<0.001	ET743_SARCOMA_72HRS_DN	-3.90	<0.001
HSA04742_TASTE_TRANSDUCTION	2.35	0.001	PENG_RAPAMYCIN_DN	-3.88	<0.001
HUMAN_TISSUE_LIVER	2.27	0.002	HSC_LATEPROGENITORS_SHARED	-3.87	<0.001
HSA04340_HEDGEHOG_SIGNALING_PATHWAY	2.15	0.006	UVB_NHEK1_DN	-3.86	<0.001
LEE_ACOX1_DN	2.15	0.005	ROME_INSULIN_2F_UP	-3.86	<0.001
LEE_CIP_DN	2.09	0.010	MITOCHONDRIA	-3.85	<0.001
GAMMA_HEXACHLOROCYCLOHEXANE_DEGRADATION	2.01	0.018	RCC_NL_UP	-3.85	<0.001
HSA05217_BASAL_CELL_CARCINOMA	2.01	0.018	ET743_SARCOMA_48HRS_DN	-3.84	<0.001
REFRACTORY_GASTRIC_UP	1.95	0.029	UVB_NHEK3_ALL	-3.84	<0.001
HSA00591_LINOLEIC_ACID_METABOLISM	1.94	0.030	RIBOSOMAL_PROTEINS	-3.77	<0.001
RORIE_ES_PNET_DN	1.94	0.029	HSC_LATEPROGENITORS_ADULT	-3.73	<0.001
MOREAUX_TACI_HI_IN_BMPC	1.89	0.041	STEMCELL_COMMON_UP	-3.73	<0.001
BCNU_GLIOMA_MGMT_48HRS_DN	1.87	0.044	FLECHNER_KIDNEY_TRANSPLANT_WELL_PBL_UP	-3.72	<0.001
WNT_TARGETS	1.84	0.053	PENG_LEUCINE_DN	-3.65	<0.001
CELL_ADHESION	1.82	0.061	HSC_LATEPROGENITORS_FETAL	-3.61	<0.001
ST_WNT_CA2_CYCLIC_GMP_PATHWAY	1.81	0.059	UVB_SCC_UP	-3.61	<0.001
STAEGE_EFTS_UP	1.80	0.064	UVC_XPCS_4HR_DN	-3.61	<0.001
GPCRS_CLASS_B_SECRETIN_LIKE	1.78	0.072	MRNA_PROCESSING_REACTOME	-3.60	<0.001
GPCRDB_CLASS_B_SECRETIN_LIKE	1.78	0.070	PENG_GLUTAMINE_DN	-3.55	<0.001
HSA00602_GLYCOSPHINGOLIPID_BIOSYNTHESIS_NEO_LACTOSERIES	1.77	0.072	MOREAUX_TACI_HI_IN_PPC_UP	-3.55	<0.001
HSA04512_ECM_RECEPTOR_INTERACTION	1.76	0.074	ET743_SARCOMA_24HRS_DN	-3.52	<0.001
NUCLEAR_RECEPTORS	1.76	0.072	SERUM_FIBROBLAST_CORE_UP	-3.52	<0.001
TESTIS_EXPRESSED_GENES	1.75	0.075	LEE_TCELLS10_UP	-3.50	<0.001
HINATA_NFKB_DN	1.74	0.078	RUTELLA_HEPATGFSNDCS_UP	-3.48	<0.001
HTERT_DN	1.73	0.080	LEE_TCELLS8_UP	-3.47	<0.001
HALMOS_CEBP_DN	1.73	0.079	UVC_HIGH_ALL_DN	-3.46	<0.001
HSIAO_LIVER_SPECIFIC_GENES	1.72	0.081	HSC_EARLYPROGENITORS_FETAL	-3.45	<0.001
CELL_ADHESION_MOLECULE_ACTIVITY	1.72	0.079	HDACI_COLON_SUL_UP	-3.45	<0.001
STRIATED_MUSCLE_CONTRACTION	1.72	0.078	UVC_TTD-XPCS_COMMON_DN	-3.40	<0.001
HCC_SURVIVAL_GOOD_VS_POOR_UP	1.70	0.088	VERHAAK_AML_NPM1_MUT_VS_WT_UP	-3.40	<0.001
HUMAN_TISSUE_PLACENTA	1.68	0.100	GOLDRATH_HP	-3.40	<0.001
DAC_PANC_UP	1.67	0.101	PRMT5_KD_UP	-3.39	<0.001
TYROSINE_METABOLISM	1.67	0.102	BASSO_REGULATORY_HUBS	-3.39	<0.001
INTRINSICPATHWAY	1.65	0.111	REOVIRUS_HEK293_UP	-3.39	<0.001
SCHURINGA_STAT5A_UP	1.64	0.118	LEE_TCELLS1_UP	-3.38	<0.001
JNK_UP	1.64	0.118	BYSTRYKH_HSC_CIS_GLOCUS	-3.37	<0.001
PASSERINI_GROWTH	1.64	0.117	PGC	-3.37	<0.001
GH_EXOGENOUS_ALL_UP	1.61	0.130	FLECHNER_KIDNEY_TRANSPLANT_REJECTION_DN	-3.36	<0.001

Abbreviations: NES=normalized enrichment score; FDR=false discovery rate

Table S10, related to Figure 4. Thirty eight gene signature associated with t-MDS/AML

Gene ID	Description
NR4A2	nuclear receptor subfamily 4, group A, member 2
FOS	v-fos FBJ murine osteosarcoma viral oncogene homolog
EGR1	early growth response 1
CARD6	caspase recruitment domain family, member 6
PEX11B	peroxisomal biogenesis factor 11 beta
EGR3	early growth response 3
EGR4	early growth response 4
MRPL15	mitochondrial ribosomal protein L15
SLC7A11	solute carrier family 7, (cationic amino acid transporter, y+ system) member 11
REEP1	receptor accessory protein 1
FOSB	FBJ murine osteosarcoma viral oncogene homolog B
GOLGA5	golgi autoantigen, golgin subfamily a, 5
ACTL6A	actin-like 6A
GOLPH3L	golgi phosphoprotein 3-like
CCDC99	coiled-coil domain containing 99
SMAD7	SMAD family member 7
SHMT2	serine hydroxymethyltransferase 2 (mitochondrial)
LRPPRC	leucine-rich PPR-motif containing
CDCA4	cell division cycle associated 4
PDIA4	protein disulfide isomerase family A, member 4
GOT1	glutamic-oxaloacetic transaminase 1, soluble (aspartate aminotransferase 1)
RTN3	reticulon 3
KLF2	Kruppel-like factor 2 (lung)
JUN	jun oncogene
STK17B	serine/threonine kinase 17b
PSMC2	proteasome (prosome, macropain) 26S subunit, ATPase, 2
LRBA	LPS-responsive vesicle trafficking, beach and anchor containing
XPOT	exportin, tRNA (nuclear export receptor for tRNAs)
ZYG11B	zyg-11 homolog B (C. elegans)
ZNF137	zinc finger protein 137
GEM	GTP binding protein overexpressed in skeletal muscle
PGRMC2	progesterone receptor membrane component 2
ARL6IP6	ADP-ribosylation-like factor 6 interacting protein 6
SLC2A3P1	Solute carrier family 2 (facilitated glucose transporter), member 3 pseudogene 1
NR4A3	nuclear receptor subfamily 4, group A, member 3
RGS2	regulator of G-protein signaling 2, 24kDa
NRIP3	nuclear receptor interacting protein 3
SLC26A2	solute carrier family 26 (sulfate transporter), member 2

Table S11, related to Figure 4. Cytogenetic analysis of t-MDS/AML cases at pre-HCT time point (provided as a separate Excel file)

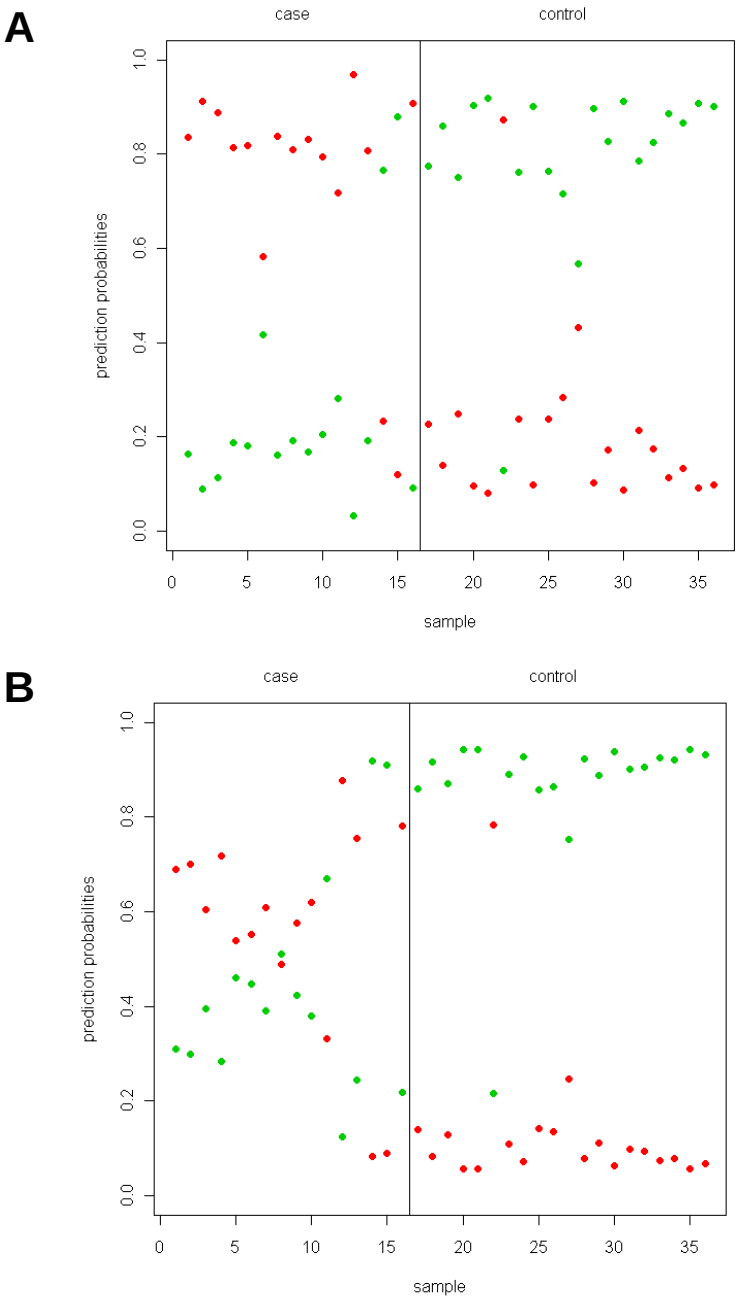


Figure S3, related to Figure 4. (A) Prediction probabilities for t-MDS/AML versus control status in test set using the 38-gene signature. We applied the 38-gene signature derived from the training set to the test set. The posterior probabilities of being t-MDS/AML (red) or control (green) are shown. The signature successfully predicted disease status for the majority of the subjects. All controls were predicted correctly, except one subject. There are 2 misclassifications in predicting t-MDS/AML. Most subjects were classified with high prediction probability. **(B) Prediction probabilities for t-MDS/AML versus control status in test set using the 31-gene signature.** The 2 cases with transient t-MDS/AML and their controls were removed from the training sets for the analysis and a 31-gene signature selected with lowest prediction error in cross-validation. Application of the 31-gene signature to the test set misclassified 4/16 cases as controls and 1/20 controls as cases in the test set.

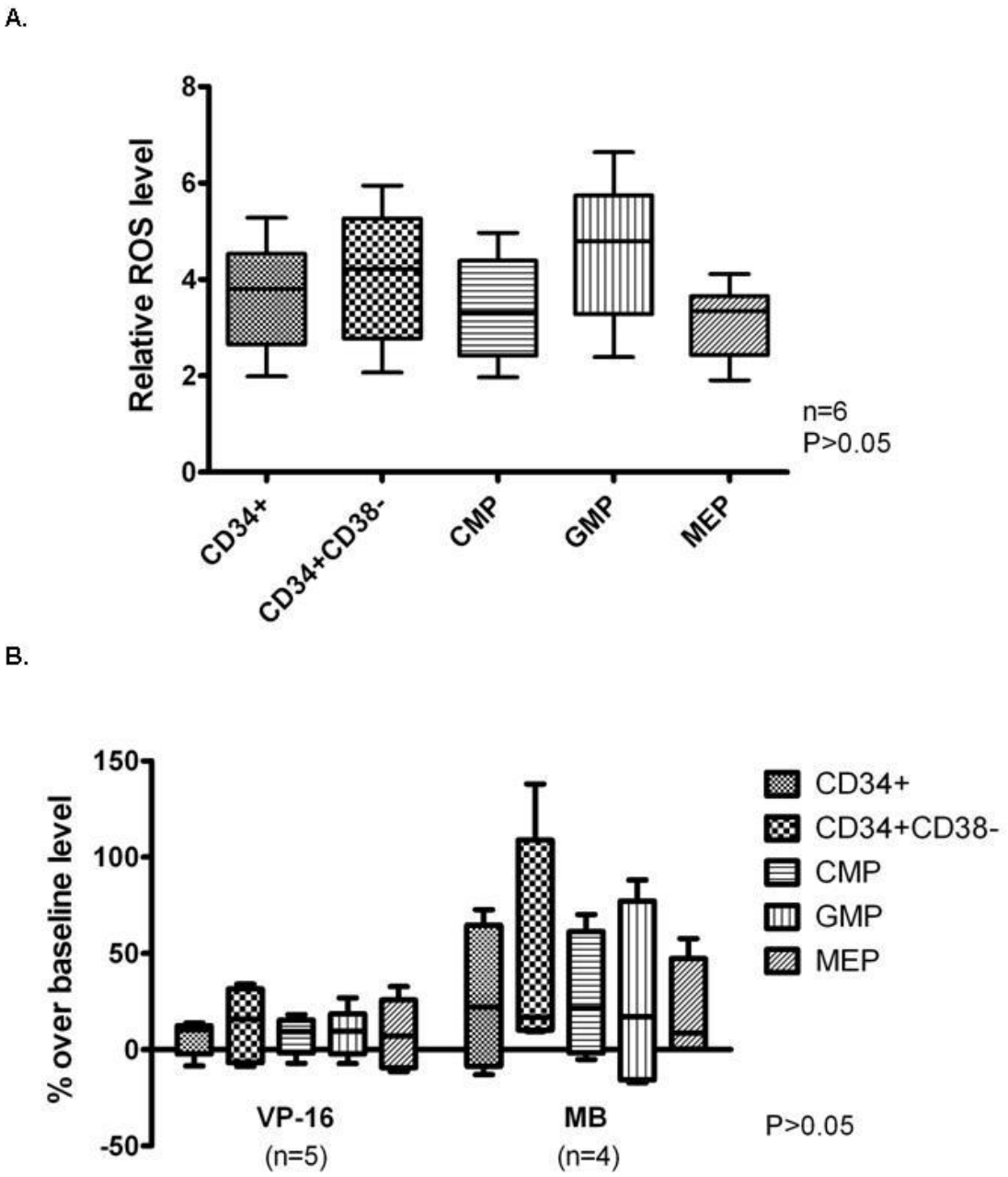


Figure S4, related to Figure 5. Lack of difference in ROS levels in different subpopulations of CD34+ cells obtained from normal PBSC samples. Normal PBSC MNCs were treated with or without etoposide (VP-16) and methylene blue with visible light (MB) and stained with lineage markers (CD11b, CD14, Glycophorin A, CD2, CD7, CD10 and CD19), CD34, CD38, CD123, CD45RA and ROS indicator C-H2DCFDA. ROS levels in different CD34 subpopulations including HSC (Lin-CD34+CD38-), CMP (Lin-CD34+CD38+CD45RA-CD123+), GMP (Lin-CD34+CD38+CD45RA+CD123+) and MEP (Lin-CD34+CD38+CD45RA-CD123-) as well as total CD34+ cells were analyzed using multi-color flow cytometry. (**A**) **Baseline ROS levels in different CD34 subpopulations from normal PBSC samples**. There was no significant difference among different CD34 subpopulations. Results represent median, interquartile range and range of values. (**B**) **ROS levels in different CD34 subpopulations of normal PBSC measured 2 hours after exposure to etoposide (VP-16) and methylene blue with visible light (MB)**. No significant difference among different CD34 subpopulations was observed. Results represent median, interquartile range and range of values.

Supplemental Experimental Procedures

Statistical analysis

For quality control of microarray data, images of the individual arrays were screened for experimental error; Affymetrix MAS 5 report was checked for background expression, scale factors and percent of present calls; and RNA degradation was examined by beta-actin 3/5 and GAPDH 3/5 ratios using Affymetrix internal controls. No obvious batch effect was observed. Data were normalized using robust multiarray averages with consideration of GC content (GCRMA), where only the probes with present call were used to estimate the background, with subsequent applications of quantile normalization and median polishing. The normalization was carried out separately for PBSC samples obtained pre-aHCT and for BM samples collected at the time of t-MDS/AML/AML or at comparable time points after aHCT for controls. The Affymetrix annotation file was used to map probesets to genes. Expression of genes represented by multiple probesets was set as the median of the probesets.

For Gene set enrichment analysis (GSEA), the pre-ranked gene list was used to test all the 1383 gene sets (with size of [15, 500]) in the C-2 category of the GSEA Molecular Signatures Database, representing curated gene sets collected from various sources including online pathway databases, biomedical literature, and the L2L database of published microarray gene expression data. Where multiple significant gene sets were related to each other, analysis was performed to identify a subset of common enriched genes. Average gene expression was calculated for each set and heatmaps plotted to show the contrasts between cases and controls. Gene Ontology (GO) and pathway analysis was performed, retaining genes with z-scores ≥1.8 or ≤-1.8, and ≥1.5-fold change in OR between cases and controls.

Pre-processing, normalization and filtering procedures for the test set were identical to the training set. PAM was used to derive a prognostic gene signature from the training set to classify patients as case or control. PAM uses the “nearest shrunken centroid” approach and 10-fold cross-validation to select a parsimonious gene expression signature that can classify samples with minimal misclassification. PAM was applied to genes common to both datasets.