Supplementary Table 1. The characteristics of patients from MDACC

Feature	Cohort (n=75)	Cohort (n=55)
Age (y)		
Median	65	65
Range	32 - 82	32 - 76
Mean	64.2	63.6
Gender		
Female	27	17
Male	48	38
Race		
Caucasian	71	53
Hispanic	1	0
Black	2	2
Asian	1	0
TNM Stage		
I	34	24
II	19	14
III	22	17
Histology type		
Adenocarcinomas	48	30
Squamous cell carcinoma	27	25
Smoking history		
No	7	2
Yes	68	53
Adjuvant chemotherapy		
No	28	21
Yes	47	34
Follow-up times (y)		
Median	2.76	2.74
Range	0.24 – 6.9	0.24 – 6.9
Mean	2.79	2.8

Abbreviations: TNM, tumor size, node involvement, metastasis status.

Supplementary Table 2. The summary of all the training and testing sets, the P-values of log-rank tests between the predicted high and low risk groups for all the prediction models derived and validated in this study. Six prediction models were developed and tested in this project. Model 3 and 5 are prediction models developed from FFPE samples based on 1400 genes and 59 genes respectively. Models 1, 2, 4 and 6 were used to demonstrate the robustness of the genes derived from FFPE samples. 1012 features in Model 2 and 4 are a subset of 1400 robust genes and the difference is due to the microarray platform difference between the MDACC study (Affymetrix U133 plus 2.0) and the consortium study (Affymetrix U133 A).

Figure	Training	Testing	P-value	Model (# of features in the model)
3A	FFPE training (N=25)	FFPE testing (N=30)	0.013	1 (1400)
3B	Consortium training (N=254)	Consortium testing (N=157)	0.00014	2 (1012)
3C	FFPE (N = 55)	Consortium (N=442)	5.4E-7	3 (1400)
3D	Consortium (N=442)	FFPE (N = 55)	0.068	4 (1012)
3E	FFPE (N = 55)	Consortium stage I (N=215)	0.036	3 (1400)
3F	FFPE (N = 55)	Consortium stage II (N=82)	0.022	3 (1400)
3G	FFPE (N = 55)	Consortium stage III (N=64)	0.021	3 (1400)
5A	FFPE (N = 55)	*Bhattacharjee et al* (N=117)	0.016	5 (59)
5B	FFPE (N = 55)	*Bhattacharjee et al* stage I (N=70)	0.039	5 (59)
5C	FFPE (N = 55)	*Bild et al* (N=111)	0.02	5 (59)
5D	FFPE (N = 55)	*Bild et al* stage I(N=62)	0.028	5 (59)
S2A	FFPE (N = 55)	MSKCC (N= 93)	0.0093	3 (1400)
S2B	FFPE (N = 55)	DFCI (N= 64)	0.0076	3 (1400)
S2C	FFPE (N = 55)	MI (N= 133)	0.0011	3 (1400)
S2D	FFPE (N = 55)	HLM (N= 69)	0.40	3 (1400)
S3A	FFPE (N = 55)	Consortium with chemo (N=89)	0.015	3 (1400)
S3B	FFPE (N = 55))	Consortium without chemo (N=233)	0.00062	3 (1400)
S4A	Consortium (N=442)	*Bhattacharjee et al* (N=117)	0.0052	6 (59)
S4B	Consortium (N=442)	*Bhattacharjee et al* stage I (N=70)	0.087	6 (59)
S4C	Consortium (N=442)	*Bild et al* (N=111)	0.17	6 (59)
S4D	Consortium (N=442)	*Bild et al* (N=62)	0.18	6 (59)

Supplementary Table 3. The association between patients' characteristics and RGS groups defined from unsupervised clustering analysis for MDACC patients. P-values for age is based on Wilcoxon test and for other categorical variables are based on Fisher's exact test.

Characteristics	Group 1	Group 2	p-value
Age (median)	65	65	0.47
Male	21 (75%)	17 (63%)	0.57
Current Smoker	17 (61%)	12 (44%)	0.22
Stage			
I	10 (36%)	14 (52%)	0.51
II	8 (28%)	6 (22%)	
III	10 (36%)	7 (26%)	
Histology			
Adenocarcinoma	5 (18%)	25 (93%)	1.3E-8
Squamous Cell Carcinoma	23 (82%)	2 (7%)	
Adjuvant Chemotherapy	14 (50%)	20 (74%)	0.10

Supplementary Table 4. The top gene sets enriched for RGS group 1 and group2 based on gene set enrichment analysis.

Enriched in group 1

	GS follow link to MSigDB	GS DETAILS	SIZE	ES	NES	NOM p-val	FDR q-val	FWER p-val	RANK AT MAX	LEADING EDGE
1	BRCA_ER_NEG	Details ...	82	0.55	2.12	0.000	0.003	0.003	243	tags=49%, list=20%, signal=57%
2	PENG_LEUCINE_DN	Details ...	23	0.67	2.00	0.000	0.007	0.015	240	tags=52%, list=19%, signal=63%
3	LI_FETAL_VS_WT_KIDNEY_DN	Details ...	15	0.76	1.99	0.000	0.005	0.016	172	tags=67%, list=14%, signal=76%
4	PRMT5_KD_UP	Details ...	31	0.62	1.94	0.000	0.009	0.036	159	tags=42%, list=13%, signal=47%
5	HOFFMANN_BIVSBII_BI_TABLE2	Details ...	19	0.66	1.87	0.000	0.017	0.085	161	tags=47%, list=13%, signal=54%
6	ELONGINA_KO_DN	Details ...	28	0.61	1.85	0.000	0.019	0.112	141	tags=39%, list=11%, signal=43%
7	TARTE_PLASMA_BLASTIC	Details ...	50	0.52	1.82	0.000	0.027	0.177	338	tags=56%, list=27%, signal=74%
8	CIS_XPC_UP	Details ...	16	0.66	1.80	0.000	0.030	0.225	81	tags=38%, list=7%, signal=40%
9	PENG_GLUTAMINE_DN	Details ...	39	0.54	1.79	0.000	0.028	0.233	257	tags=46%, list=21%, signal=56%
10	HSA04110_CELL_CYCLE	Details ...	15	0.68	1.77	0.002	0.032	0.286	252	tags=53%, list=20%, signal=66%
11	STEMCELL_NEURAL_UP	Details ...	314	0.38	1.73	0.000	0.044	0.408	344	tags=40%, list=28%, signal=41%
12	STEMCELL_EMBRYONIC_UP	Details ...	226	0.39	1.72	0.000	0.046	0.449	313	tags=36%, list=25%, signal=40%
13	HCC_SURVIVAL_GOOD_VS_POOR_DN	Details ...	40	0.49	1.62	0.006	0.107	0.778	275	tags=43%, list=22%, signal=53%
14	SERUM_FIBROBLAST_CORE_UP	Details ...	22	0.54	1.60	0.021	0.119	0.838	243	tags=55%, list=20%, signal=67%
15	MOREAUX_TACI_HI_VS_LOW_DN	Details ...	24	0.54	1.57	0.029	0.146	0.904	252	tags=50%, list=20%, signal=61%

Enriched in group 2

	GS follow link to MSigDB	GS DETAILS	SIZE	ES	NES	NOM p-val	FDR q-val	FWER p-val	RANK AT MAX	LEADING EDGE
1	LI_FETAL_VS_WT_KIDNEY_UP	Details ...	24	-0.74	-2.25	0.000	0.000	0.000	138	tags=46%, list=11%, signal=51%
2	BRCA_ER_POS	Details ...	53	-0.53	-1.94	0.000	0.035	0.062	197	tags=40%, list=16%, signal=45%
3	TAKEDA_NUP8_HOXA9_8D_DN	Details ...	19	-0.65	-1.88	0.004	0.054	0.140	116	tags=37%, list=9%, signal=40%
4	SERUM_FIBROBLAST_CORE_DN	Details ...	27	-0.57	-1.73	0.012	0.173	0.482	188	tags=41%, list=15%, signal=47%
5	PENG_RAPAMYCIN_UP	Details ...	19	-0.60	-1.71	0.006	0.168	0.548	388	tags=68%, list=31%, signal=98%
6	UEDA_MOUSE_LIVER	Details ...	15	-0.63	-1.70	0.013	0.153	0.581	310	tags=53%, list=25%, signal=70%
7	AGEING_KIDNEY_UP	Details ...	42	-0.48	-1.64	0.000	0.210	0.742	181	tags=31%, list=15%, signal=35%
8	HSC_MATURE_FETAL	Details ...	38	-0.47	-1.60	0.019	0.266	0.860	210	tags=29%, list=17%, signal=34%
9	IRITANI_ADPROX_VASC	Details ...	16	-0.59	-1.59	0.036	0.243	0.869	242	tags=50%, list=19%, signal=61%
10	TAKEDA_NUP8_HOXA9_3D_UP	Details ...	20	-0.54	-1.59	0.029	0.226	0.874	121	tags=30%, list=10%, signal=33%
11	HSC_MATURE_SHARED	Details ...	31	-0.49	-1.59	0.013	0.211	0.882	208	tags=29%, list=17%, signal=34%
12	TPA_SENS_MIDDLE_DN	Details ...	26	-0.51	-1.57	0.022	0.213	0.904	235	tags=42%, list=19%, signal=51%

Supplementary Table 5. 59-genes and the association with survival in both FFPE and Consortium set.

Acession	Symbol	Hazard Ratio from FFPE set	p-value from FFPE set	Hazard Ratio from consortium set	p-value from consortium set
NM_002107	H3F3A	0.30	0.034	0.57	0.019
NM_031263	HNRNPK	3.23	0.018	1.73	0.019
NM_002156	HSPD1	2.23	0.010	1.82	0.000
NM_003472	DEK	3.89	0.005	1.43	0.004
NM_016587	CBX3	3.17	0.002	1.78	0.000
NM_173638	NBPF15	0.19	0.008	0.53	0.001
NM_001677	ATP1B1	0.61	0.023	0.87	0.028
NM_005003	NDUFAB1	1.99	0.017	1.57	0.008
NM_005124	NUP153	4.93	0.012	1.52	0.006
NM_004390	CTSH	0.57	0.017	0.69	0.000
NM_014736	KIAA0101	2.04	0.011	1.47	0.000
NM_000935	PLOD2	1.82	0.019	1.22	0.000
NM_012322	LSM5	4.64	0.000	1.90	0.000
NM_002485	NBN	3.39	0.030	1.45	0.012
NM_002453	MTIF2	2.11	0.046	1.42	0.021
NM_002789	PSMA4	3.34	0.022	1.31	0.049
NM_004607	TBCA	3.56	0.026	1.60	0.004
NM_006660	CLPX	4.05	0.017	1.74	0.000
NM_002137	HNRNPA2B1	2.25	0.023	1.83	0.009
NM_001918	DBT	0.25	0.031	0.65	0.001
NM_003096	SNRPG	2.35	0.019	1.67	0.000
NM_003090	SNRPA1	1.95	0.035	1.52	0.003
NM_030881	DDX17	0.17	0.001	0.48	0.000
NM_007208	MRPL3	2.43	0.015	1.42	0.006
NM_002129	HMGB2	1.89	0.025	1.33	0.001
NM_018947	CYCS	2.58	0.027	2.04	0.000
NM_005596	NFIB	0.49	0.013	0.75	0.001
NM_007100	ATP5I	2.20	0.037	1.52	0.016
NM_015149	RGL1	0.44	0.023	0.69	0.000
NM_170662	CBLB	0.42	0.004	0.79	0.036
AL136621	ZMYM2	0.44	0.009	0.56	0.016
NM_006082	TUBA1B	2.44	0.047	1.63	0.005
NM_000712	BLVRA	0.32	0.040	0.73	0.017
NM_033551	LARP1	4.72	0.018	1.82	0.001
NM_015335	MED13L	0.41	0.036	0.78	0.037
AK057191	IDS	0.42	0.032	0.77	0.015
NM_002076	GNS	0.44	0.017	0.68	0.009
NM_005177	ATP6V0A1	0.27	0.015	0.60	0.001
NM_015962	FCF1	3.18	0.007	1.34	0.043
NM_033450	ABCC10	0.28	0.025	0.66	0.001
NM_198843	SFTPB	0.68	0.013	0.85	0.000
NM_001037637	BTF3	2.74	0.047	1.39	0.020
NM_148923	CYB5A	0.49	0.010	0.79	0.000

NM_016061	**YPEL5**	0.23	0.013	0.66	0.004
NM_016021	**UBE2J1**	0.26	0.041	0.76	0.044
NM_014056	**HIGD1A**	3.24	0.009	1.36	0.026
NM_016359	**NUSAP1**	2.00	0.037	1.31	0.000
NM_021825	**CCDC90B**	3.17	0.018	1.28	0.010
NM_014167	**CCDC59**	2.55	0.039	1.67	0.000
NM_013341	**OLA1**	5.07	0.001	1.58	0.000
NM_030793	**FBXO38**	0.39	0.021	0.72	0.011
NM_003677	**DENR**	3.55	0.011	1.34	0.041
NM_144567	**ANGEL2**	0.18	0.004	0.69	0.011
CR601845	**N4BP2L2**	0.40	0.009	0.73	0.003
NM_182746	**MCM4**	2.72	0.001	1.37	0.000
NM_001991	**EZH1**	0.31	0.034	0.45	0.000
NM_015349	**KIAA0240**	0.57	0.049	0.54	0.000
NM_001040455	**SIDT2**	0.41	0.015	0.74	0.031
NM_001012339	**DNAJC21**	0.20	0.034	0.27	0.001

Supplementary Table 6. The association between patients' characteristics and 59-gene signature and survival time for stage I patients from Bild et al study and Bhattacharjee et al study based on multivariate Cox regression model. In the left panel, 59-gene signature scores were calculated from the prediction model built from MDACC FFPE samples; in the right panel, 59-gene signature scores were calculated from the prediction model built from the consortium frozen samples.

Variables	MDACC FFPE samples used as the training set		Consortium frozen samples used as the training set	
	HR (95% CI)	**p-value**	**HR (95% CI)**	**p-value**
Bild et al dataset as the testing set				
Risk Scores	1.55 (1.05, 2.30)	0.027	3.43 (0.99,11.87)	0.051
Gender (Male vs. Female)	1.30 (0.65, 2.63)	0.46	1.32 (0.65, 2.68)	0.45
Smoking	1.00 (0.99, 1.01)	0.92	1.00 (0.99, 1.01)	0.82
Bhattacharjee et al dataset as the testing set				
Risk Scores	1.36 (1.00, 1.84)	0.047	3.05 (0.94, 9.89)	0.064
Histology	0.66 (0.27, 1.60)	0.36	0.66 (0.26, 1.65)	0.37

Supplementary Table 7. The risk formula for 59-gene signature developed from FFPE samples. The table gives the U matrix for the singular value decomposition (SVD) analysis and the corresponding coefficients are -2.45, 0.75 and 0.43.

	Component 1	Component 2	Component 3
200080_s_at	0.05	-0.02	-0.08
200807_s_at	-0.13	0.03	-0.15
200934_at	-0.10	-0.05	-0.03
201091_s_at	-0.11	0.06	-0.06
201103_x_at	0.08	-0.05	0.01
201242_s_at	0.15	0.13	-0.23
202077_at	-0.05	-0.02	0.02
202295_s_at	0.29	0.12	0.26
202503_s_at	-0.24	0.30	0.08
202620_s_at	-0.15	0.28	0.66
202904_s_at	-0.18	-0.11	-0.06
203396_at	-0.11	0.06	-0.04
204809_at	-0.14	0.22	-0.10
205292_s_at	-0.04	-0.15	-0.06
205370_x_at	0.06	-0.08	-0.01
205644_s_at	-0.13	0.07	-0.03
206055_s_at	-0.15	0.17	0.13
208718_at	0.11	0.02	0.13
208808_s_at	-0.14	-0.05	0.07
209289_at	0.16	-0.19	0.20
209492_x_at	-0.04	-0.03	-0.11
209568_s_at	0.14	-0.25	0.06
209682_at	0.11	-0.15	0.30
210282_at	0.08	0.22	-0.03
212208_at	0.06	0.02	-0.02
212221_x_at	0.08	-0.10	0.11
212334_at	0.15	0.08	0.14
212383_at	0.10	-0.18	0.01
212499_s_at	-0.07	0.00	-0.04
213485_s_at	0.01	-0.08	-0.09
213936_x_at	0.45	0.56	-0.16
215726_s_at	0.23	0.09	-0.28
217783_s_at	0.10	-0.05	0.00
217845_x_at	-0.04	0.09	-0.02
218039_at	-0.24	-0.09	-0.10
219293_s_at	-0.09	0.01	0.01
221257_x_at	0.18	-0.20	-0.04
221825_at	0.08	-0.08	-0.08
221899_at	0.17	-0.09	0.01
222036_s_at	-0.26	0.06	-0.15
38892_at	0.17	-0.18	0.07
56256_at	0.10	-0.04	0.07