Supplementary Materials:

S1. S-Score performs and comparison with other scoring methods

We first use the eRMS data, discussed in (1) to demonstrate the performance of the *S*-score, where we have applied *S*-score to classify different subtypes of sarcoma based on GSSs of p53 off, Ras on, Shh on, and Rb1 off (1). The classification has been validated by biological experiments and H&E staining. The promising results demonstrated the ability of *S*-score for scoring and clustering based on GSSs. Here we used one of the GSS, p53off, derived by comparing cancer samples with p53 mutation (marked in red in *Figure 1*) *vs.* samples with normal p53 (marked in green in *Figure 1*) to compare the capability of *S*-score against the other three scoring methods, ES-score (2), averaged *Z*-score (3), and Pearson correlation (4,5). The p53off GSS contains 150 up-regulated genes and 176 down-regulated genes (1,6). 112 expression profiles containing 94 embryonal rhabdomyosarcomas (eRMS) with unknown mutation status and 18 normal skeletal muscles (with normal p53) were applied for scoring (7). For the methods of ES-score and *Z*-score that can only process GSSs containing genes with the same direction of fold change (either up or down), up-regulated and down-regulated subsets of the p53off signature (p53off-up and p53off-down) were applied instead of whole signatures. The scores of the four scoring methods were shown in *Figure S1A* in descending order of *S*-score. Control samples (labeled as green), which were utilized to generate p53off signature, have the lowest scores in the methods of *S*-score, ES score in p53off-down, *Z*-score in p53off-down, and correlation. Normal skeletal muscle samples (labeled as black) have scores lower than the control samples in ES-score and *Z*-score of p53off-on. The scoring profiles of eRMS tumor samples, which are labeled as blue, have a similar index ranking pattern in *S*-score, ES-score in p53off-up, *Z*-score in p53off-up, and correlation. However, due to the limitation of ES-score and *Z*-score, the profiles of p53off-up and p53off-down subsets are quite different in both ES-score and *Z*-score methods. The order of the samples, which are sorted by the scores of each method, are shown in *Figure S1B.* The p53off scores of tumors are higher than most of the normal muscles in methods of *S*-score, ES-score in p53off-up, *Z*-score in p53off-up, and correlation. Notice that the distributions of tumor and normal muscles are mixed in p53off-down subset in both ES-score and *Z*-score. This result implicates that p53off-down subset has worse representation of the p53off status while they correctly assign index score to control samples. Clearly, without the integrated ability of both up- and down-regulated genes, it is difficult to obtain a conclusive score that faithfully reflects the expression pattern using ES-score and *Z*-score. With the capacity of integration, *S*-score and correlation-based method accurately project expression pattern to a signature score, with the suppression of the noisy effect of unrepresentative genes in the GSS, if they exist.

Table S1 The data sources of the 31 gene signature sets

Signature set	# of genes	Data source	Study
CCS	210	GSE1692	(8,9)
KRT19	110	*Table S3* in (10)	(10)
EpCam	76	GSE5975	(11)
wound	402	SMD*	(12)
shh	552	GSE10327	(13)
RAF	140	GSE3542	(14)
MEK	83		
Erbb2	59		
EGFR+EGF	100		
Kras addiction	243	GSE15126	(15)
TGFβ	178	*Table S2* in (16)	(16)
c-Met	272	GSE25142	(17)
Acox1	91	GSE1897	(18)
Src	827	http://data. duke. genome.edu	(19)
E2F1	295		
STAT3	300		
p63	255		
p53	737		
Myc	628		
AKT	671		
PI3K	190		
Her2	138		
TNF	82		
IFNα	282		
IFNγ	189		
βCatnin	230		
EGFR	602		
TGFβ.2	35		
PR	155		
ER	160		
Ras	435		

SMD: stanford microarray database

Table S2 The differential expressed gene ontology items of CAC

Gene set	CC *vs.* NL		CC *vs.* IDB	
	Fold change	*P* value	Fold change	*P* value
Biological process				
Mitotic spindle organization	0.70	5.5E-21	0.75	2.3E-04
DNA strand elongation involved in DNA replication	0.67	5.1E-15	0.75	2.4E-03
Chromosome organization	0.65	4.6E-21	0.78	2.1E-05
mRNA metabolic process	0.58	1.0E-27	0.42	2.1E-03
Mitotic prometaphase	0.57	2.2E-19	0.66	2.1E-04
M phase of mitotic cell cycle	0.56	6.8E-18	0.66	3.0E-04
RNA metabolic process	0.53	2.0E-25	0.41	2.2E-03
Regulation of transcription involved in G1/S phase of mitotic cell cycle	0.52	3.3E-12	0.56	6.4E-03
Telomere maintenance via semi-conservative replication	0.50	1.1E-10	0.67	3.0E-03
Spliceosome assembly	0.50	6.7E-17	0.43	8.2E-03
Lamellipodium assembly	0.50	6.3E-22	0.44	1.2E-03
Telomere maintenance via recombination	0.49	5.3E-11	0.66	2.7E-03
Mitotic sister chromatid segregation	0.49	6.4E-10	0.64	3.6E-03
Mitotic cell cycle	0.49	2.4E-20	0.55	1.8E-04
Mitotic cell cycle spindle assembly checkpoint	0.49	2.6E-19	0.49	1.0E-03
CenH3-containing nucleosome assembly at centromere	0.48	9.7E-14	0.66	1.9E-04
DNA-dependent DNA replication initiation	0.47	1.5E-09	0.67	2.8E-03
Regulation of translational initiation	0.47	3.5E-17	0.46	2.5E-03
Cell cycle checkpoint	0.47	2.9E-19	0.54	1.9E-04
M/G1 transition of mitotic cell cycle	0.46	3.2E-15	0.54	8.9E-04
RNA splicing, via transesterification reactions	0.46	1.6E-15	0.43	3.2E-03
Negative regulation of ubiquitin-protein ligase activity involved in mitotic cell cycle	0.44	3.1E-16	0.47	1.3E-03
Anaphase-promoting complex-dependent proteasomal ubiquitin-dependent protein catabolic process	0.43	1.4E-15	0.51	7.9E-04
Membrane protein ectodomain proteolysis	0.43	1.7E-11	0.49	9.9E-04
S phase of mitotic cell cycle	0.43	9.3E-15	0.53	7.8E-04
Translation	0.43	1.4E-15	0.41	5.6E-03
Nucleotide-excision repair, DNA gap filling	0.42	8.7E-10	0.59	3.3E-03
Cell chemotaxis	0.42	5.9E-15	0.41	4.5E-03
Regulation of ubiquitin-protein ligase activity involved in mitotic cell cycle	0.42	9.4E-16	0.50	6.1E-04
Positive regulation of actin filament polymerization	0.42	3.6E-16	0.40	3.0E-03
Cell division	0.41	1.2E-17	0.45	8.8E-04
Double-strand break repair via homologous recombination	0.41	1.5E-13	0.49	2.3E-03
Acute-phase response	–0.60	8.2E-25	-0.50	4.5E-04
Triglyceride metabolic process	–0.69	2.5E-32	-0.63	7.1E-07
Blood coagulation, intrinsic pathway	–0.95	2.0E-34	-0.58	1.0E-03
Complement activation	–1.03	1.8E-26	-0.63	6.1E-03
Triglyceride homeostasis	–1.09	6.9E-38	-0.70	1.1E-04

(*continued*)

© Pioneer Bioscience Publishing Company. All rights reserved.

Table S2 The differential expressed gene ontology items of CAC(*continued*)

Gene set	CC *vs.* NL		CC *vs.* IDB	
	Fold change	*P* value	Fold change	*P* value
Molecular function				
Proton-transporting ATPase activity, rotational mechanism	0.60	4.3E-17	0.74	2.4E-04
Hydrogen ion transporting ATP synthase activity, rotational mechanism	0.57	1.0E-14	0.75	3.1E-04
RNA helicase activity	0.50	8.6E-12	0.56	2.3E-03
ATP-dependent DNA helicase activity	0.48	7.5E-16	0.56	6.1E-04
DNA helicase activity	0.47	5.3E-13	0.67	1.5E-04
Rac GTPase binding	0.46	5.5E-17	0.42	3.5E-03
WW domain binding	0.45	2.9E-19	0.42	1.3E-03
Protein phosphatase type 2A regulator activity	0.44	8.8E-17	0.61	2.4E-05
MHC class I protein binding	0.42	3.6E-09	0.73	2.5E-04
Nuclease activity	0.41	1.5E-13	0.53	7.9E-04
Fatty acid binding	–0.59	5.1E-35	–0.51	3.6E-07
Lipid transporter activity	–0.89	3.4E-30	–0.57	2.3E-03
Cellular component				
Eukaryotic translation initiation factor 3 complex	0.73	7.4E-19	0.72	1.4E-03
U12-type spliceosomal complex	0.55	4.5E-19	0.47	5.6E-03
Spindle microtubule	0.52	2.6E-17	0.53	1.8E-03
MLL1 complex	0.52	4.0E-18	0.61	1.8E-04
Lateral plasma membrane	0.49	2.2E-19	0.46	1.3E-03
Ribonucleoprotein complex	0.46	7.3E-22	0.42	1.1E-03
Condensed chromosome kinetochore	0.45	3.7E-15	0.56	6.7E-04
Spindle pole	0.44	8.5E-17	0.40	6.3E-03
Membrane coat	0.43	1.3E-14	0.59	1.6E-04
Condensed chromosome	0.42	1.5E-11	0.57	1.5E-03
Kinetochore	0.41	6.4E-17	0.53	1.2E-04
Kinesin complex	0.40	1.9E-13	0.44	2.0E-03
Very-low-density lipoprotein particle	–0.92	4.6E-26	–0.57	4.8E-03
High-density lipoprotein particle	–0.99	3.8E-29	–0.68	9.9E-04

CC, cholangiocarcinoma; NL, normal liver; IDB, intrahepatic bile duct

S2. Robustness of S-Score

To double confirm the robustness of the scoring methods, we also performed a simulation with a negative score samples as blue arrow pointed in *Figure S1A*. The simulation was the same as the one of the positive score sample: The mean of added noises set to zero and the strengths (standard deviation) were increased from 0.1 to 2 times the standard deviation of each gene. Each condition was simulated 1,000 times and then calculations were made for the mean shift and standard deviation of each score. The result, shown in *Figure S2*, is similar to the positive score one. *S*-score and *Z*-score have the smallest score shift (all close to zero), regardless of the standard deviation of added noise (*Figure S2A*). The mean shift of correlation method is similar to ES-score in p53off-up and p53off-down that increase with the standard deviation of noise (*Figure S2B*). Among these methods, ES-score with p53off-up gene set has the largest mean shift in two simulations, indicating worst robustness under noisy conditions. The standard deviations of all four test scores were varied at a similar range with those of ES-score with p53off-up genes and correlation slightly lower than other methods.

© Pioneer Bioscience Publishing Company. All rights reserved.

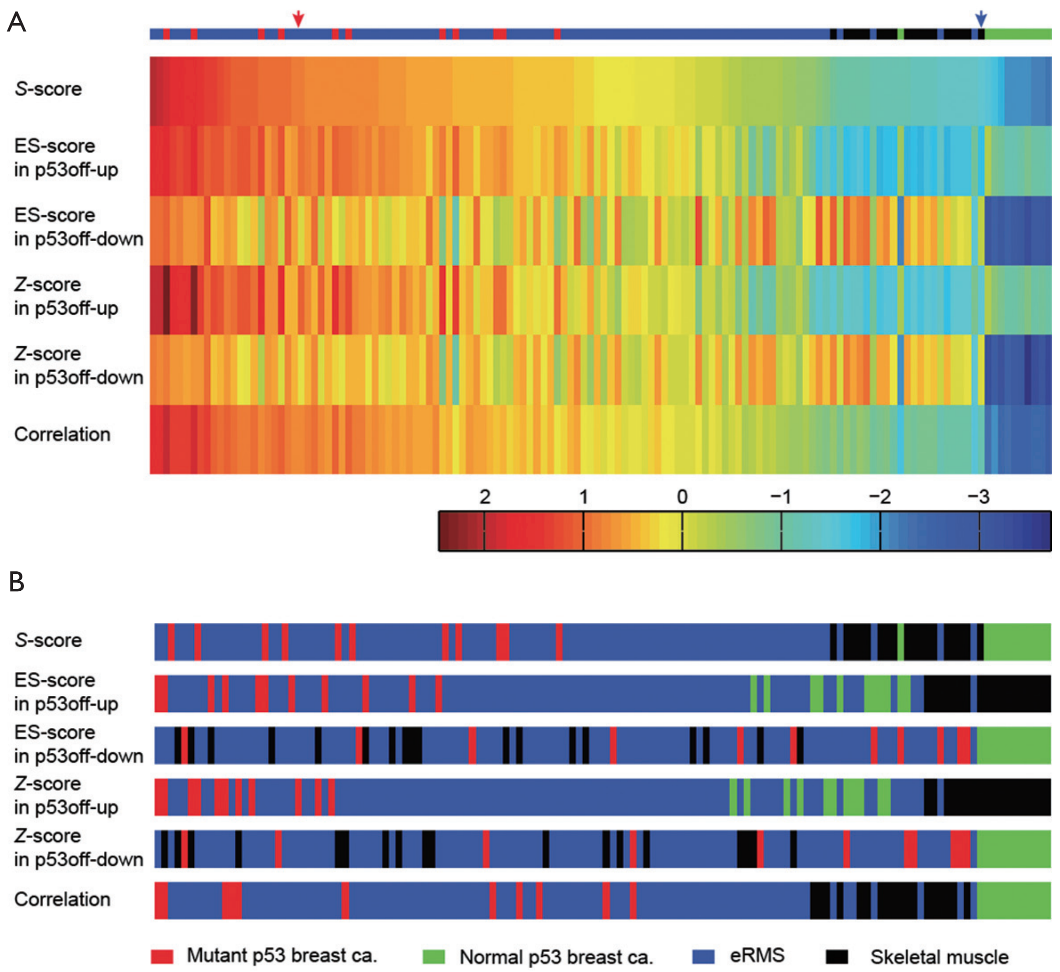

Figure S1 The heatmap of GSS scores. A comparison of scoring methods, *S*-score, ES-score and *Z*-score in p53off-up and p53off-down subsets, and correlation, were performed. A. A heatmap of normalized scores of the methods sorted in descending order of *S*-score; B. A heatmap of sample types sorted by the order of score values of each method. The samples of mutant and normal p53 breast cancers, which were used to derive the p53off GSSs, were labeled as red and green, respectively. eRMS samples were labeled as blue and normal muscles were labeled as black. Two samples, indicated by the arrows, were applied to the simulations of robustness (*Figure 1 & S2*)

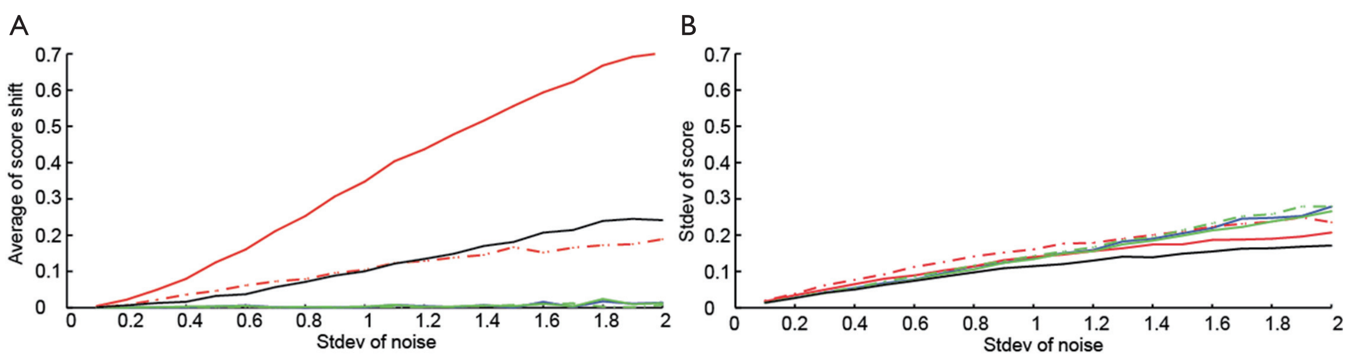

Figure S2 The plots of scoring variation under noise perturbation. One negative score samples as the blue arrows indicate in *Figure S1* were applied for the simulations of robustness under noise disturbance. (A) The mean shifts and (B) standard deviations the negative score (p53 inactive)

© Pioneer Bioscience Publishing Company. All rights reserved.

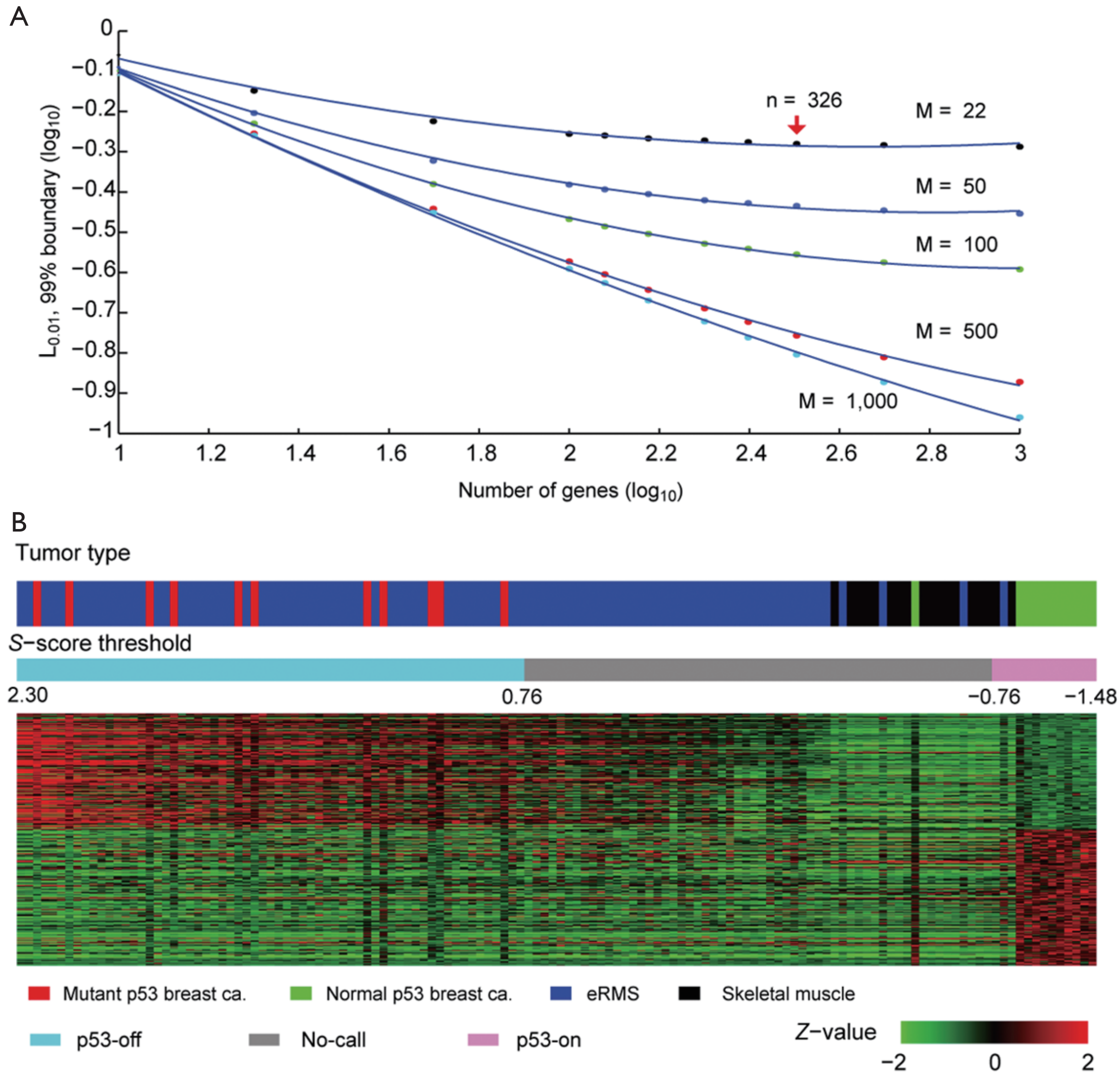

Figure S3 The qualitative analysis to define the p53-off status. A boundary of reliable region $L_{0.01}$ was determined to identify the status of p53-off through *S*-score. A. Relationship between number of genes, number of samples, and the no-call boundary $L_{0.01}$. The $L_{0.01}$ for the study, where n = 326 and M =22, was ±0.76 as the arrow indicates; B. Under the boundary of ±0.76, a total 52 of 97 eRMS samples (labeled as blue) were determined as p53-off status. Two normal muscles and one eRMS (labeled as purple) were determined as p53-on status. Others were in the no-call region and the status was not determined

S3. Qualtitative status of signature

Although the *S*-score of each tumor sample was evaluated, the status of the signature (*i.e.*, active or inactive) was not determined since simulations for the predictive interval need to be performed with current experiment setting. We have performed the simulation to cover a range of practical applications, and results are shown in *Figure S3A*. In the figure, we varied the number of genes in the signature set n=50 to 1,000, and the number of arrays M=22, 50, 100, 500, and 1,000. As expected, the boundaries are smaller with larger number of genes or more samples (chances of making "no call" shall be smaller with more genes in the gene set or more samples). For the parameters of p53off signature: n=326 and M=22, $L0.01$ was ±0.76, which is indicated by an arrow in *Figure S3A*. We will not make a status call for a given sample to be either "on" or "off" status of p53-off signature if

© Pioneer Bioscience Publishing Company. All rights reserved.

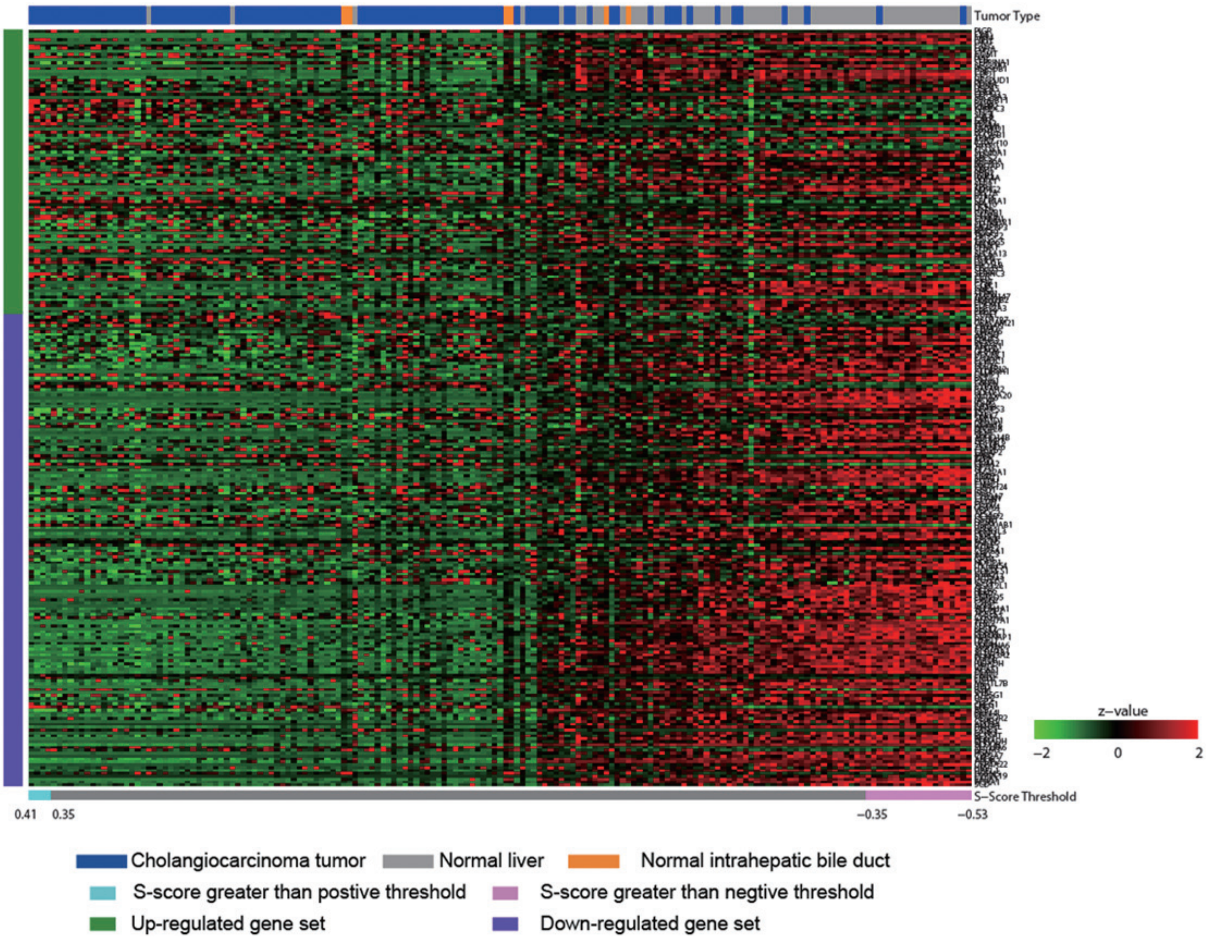

Figure S4 The heatmap of the c-Met signature set in the data set GSE26566. The expression of most genes in CAC samples was down-regulated for both up and down-regualted c-Met gene set. The expression profile lose the trend of expression direction in the original case-control study

the *S*-score is between (-0.76, 0.76) in order to guarantee no more than 1% of classification error. By applying the boundary value to the *S*-scores of test tumor samples, a total 52 of 97 eRMS samples (labeled as blue) were determined as p53-off status. Only two normal muscles and one eRMS (labeled as purple) were determined as p53-on status (*Figure S3B*).

References

1. Rubin BP, Nishijo K, Chen HI, et al. Evidence for an unanticipated relationship between undifferentiated pleomorphic sarcoma and embryonal rhabdomyosarcoma. Cancer Cell 2011;19:177-91.
2. Barbie DA, Tamayo P, Boehm JS, et al. Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. Nature 2009;462:108-12.
3. Ebi H, Tomida S, Takeuchi T, et al. Relationship of deregulated signaling converging onto mTOR with prognosis and classification of lung adenocarcinoma shown by two independent in silico analyses. Cancer Res 2009;69:4027-35.
4. Creighton CJ, Li X, Landis M, et al. Residual breast cancers after conventional therapy display mesenchymal as well as tumor-initiating features. Proc Natl Acad Sci U S A 2009;106:13820-5.
5. Gibbons DL, Lin W, Creighton CJ, et al. Expression signatures of metastatic capacity in a genetic mouse model of lung adenocarcinoma. PLoS One 2009;4:e5401.
6. Miller LD, Smeds J, George J, et al. An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. Proc Natl Acad Sci U S A 2005;102:13550-5.

© Pioneer Bioscience Publishing Company. All rights reserved.

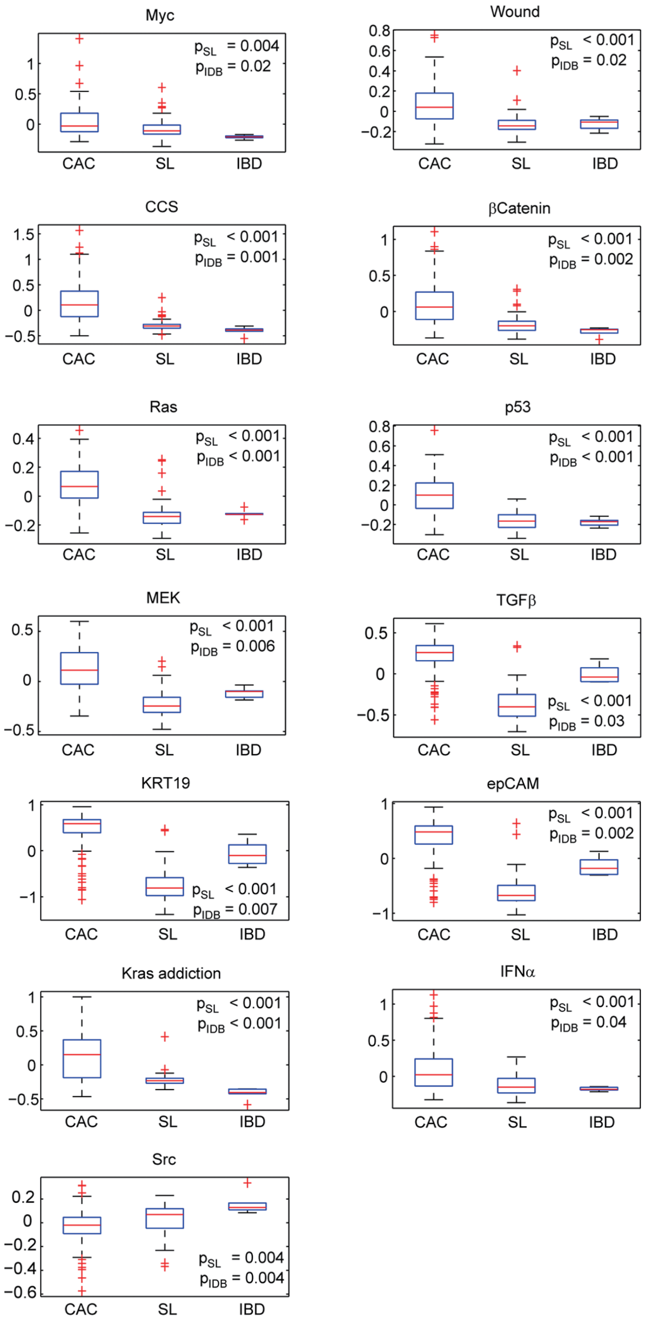

Figure S5 The boxplot of differential activities of signatures. 13 signature sets were showed differential level between cholangiocarcinoma (CAC) and other two normal tissues in data set GSE26566. P_{SL}: *P*-value of t-test between CAC and surrounding liver tissues (SL). P_{IDB}: *P*-value of t-test between CAC and bile ducts (IBD). After multi-test adjustment of Benjamini-Hocher, 12 out of 13 GSSs passed the criteria of adj. *P*<0.05

7. Laé M, Ahn EH, Mercado GE, et al. Global gene expression profiling of PAX-FKHR fusion-positive alveolar and PAX-FKHR fusion-negative embryonal rhabdomyosarcomas. J Pathol 2007;212:143-51.
8. Cam H, Balciunaite E, Blais A, et al. A common set of gene regulatory networks links metabolism and growth inhibition. Mol Cell 2004;16:399-411.
9. Cam H, Balciunaite E, Blais A, et al. A common set of gene regulatory networks links metabolism and growth inhibition. Mol Cell 2004;16:399-411.
10. Andersen JB, Loi R, Perra A, et al. Progenitor-derived hepatocellular carcinoma model in the rat. Hepatology 2010;51:1401-9.
11. Yamashita T, Forgues M, Wang W, et al. EpCAM and alpha-fetoprotein expression defines novel prognostic subtypes of hepatocellular carcinoma. Cancer Res 2008;68:1451-61.
12. Chang HY, Sneddon JB, Alizadeh AA, et al. Gene expression signature of fibroblast serum response predicts human cancer progression: similarities between tumors and wounds. PLoS Biol 2004;2:E7.
13. Kool M, Koster J, Bunt J, et al. Integrated genomics identifies five medulloblastoma subtypes with distinct genetic profiles, pathway signatures and clinicopathological features. PLoS One 2008;3:e3088.
14. Creighton CJ, Hilger AM, Murthy S, et al. Activation of mitogen-activated protein kinase in estrogen receptor alpha-positive breast cancer cells in vitro induces an in vivo molecular phenotype of estrogen receptor alpha-negative human breast tumors. Cancer Res 2006;66:3903-11.
15. Singh A, Greninger P, Rhodes D, et al. A gene expression signature associated with "K-Ras addiction" reveals regulators of EMT and tumor cell survival. Cancer Cell 2009;15:489-500.
16. Coulouarn C, Factor VM, Thorgeirsson SS. Transforming growth factor-beta gene expression signature in mouse hepatocytes predicts clinical outcome in human cancer. Hepatology 2008;47:2059-67.
17. Lamb JR, Zhang C, Xie T, et al. Predictive genes in adjacent normal tissue are preferentially altered by sCNV during tumorigenesis in liver cancer and may rate limiting. PLoS One 2011;6:e20090.
18. Lee JS, Heo J, Libbrecht L, et al. A novel prognostic subtype of human hepatocellular carcinoma derived from hepatic progenitor cells. Nat Med 2006;12:410-6.
19. Gatza ML, Lucas JE, Barry WT, et al. A pathway-based classification of human breast cancer. Proc Natl Acad Sci U S A 2010;107:6994-9.

© Pioneer Bioscience Publishing Company. All rights reserved.