

Figure S1 Study flow chart for our analysis.

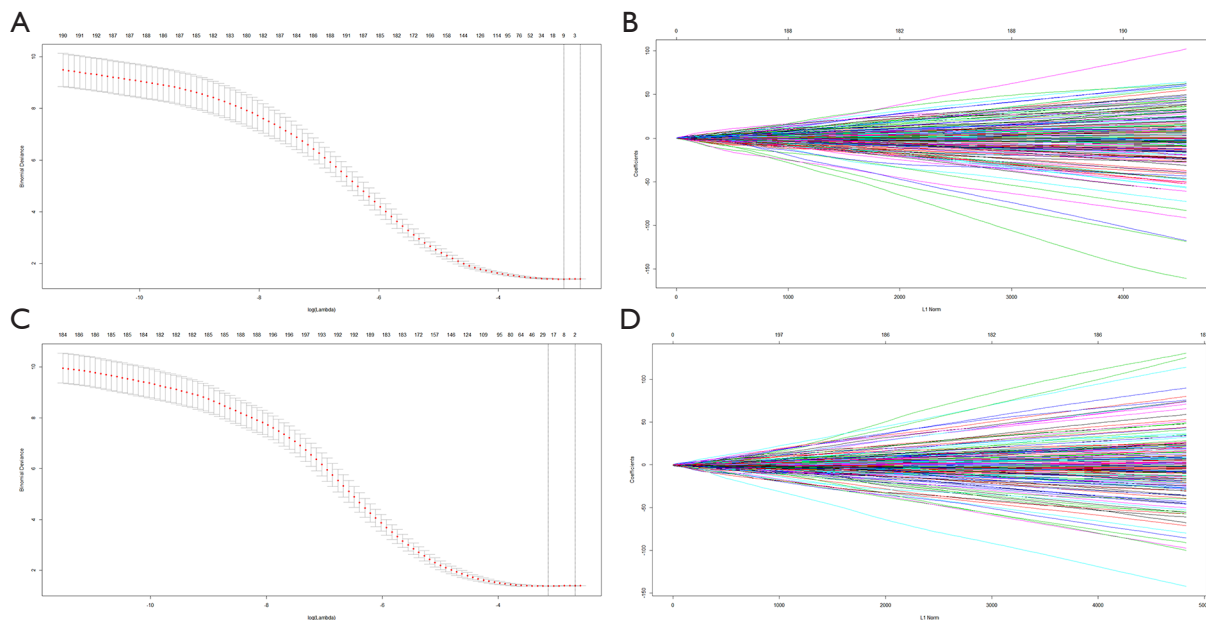


Figure S2 Establishment of the risk signature integrating the mutated genes. (A,B) OS; (C,D) PFS. The coefficients estimated by the Lasso regression method are presented. Each curve in the (A,C) represents the path of a lasso coefficient against the L1-norm (the penalty term for lasso) when λ changes. A coefficient that becomes non-zero when λ changes enters the LASSO regression model. (B,D) The coefficients estimated by the Lasso regression method are presented. OS, overall survival; PFS, progression free survival; LASSO, least absolute shrinkage and selection operator.

Table S1 Gene mutation type, frequency and risk score formula included in the analysis

Genes	Point mutation	Splice site	Total	Frequency (%)	Coefficient (OS risk score)	Coefficient (PFS risk score)
<i>ADCY8</i>	19	0	19	6.1	-0.35	-
<i>ALK</i>	5	0	5	1.6	0.07	0.27
<i>ARID1A</i>	12	4	16	5.1	0.13	-0.48
<i>CDK8</i>	6	0	6	1.9	0.26	0.58
<i>DICER1</i>	14	1	15	4.8	-0.25	-
<i>EPC1</i>	5	0	5	1.6	0.027	-
<i>ERBB2</i>	7	0	7	2.2	0.07	-
<i>MED12</i>	11	0	11	3.5	0.11	-
<i>TSC1</i>	8	0	8	2.6	0.43	-
<i>MAP3K13</i>	10	0	10	3.2	-	-0.91
<i>BAP1</i>	3	2	5	1.6	-	-0.75
<i>NCOR1</i>	12	0	12	3.8	-	-0.21
<i>TSHR</i>	9	0	9	2.9	-	-0.17
<i>CS</i>	4	0	4	1.3	-	-0.17
<i>PMS1</i>	3	0	3	1.0	-	0.036
<i>MSH2</i>	4	0	4	1.3	-	0.043
<i>BCR</i>	11	0	11	3.5	-	0.095
<i>MLL2</i>	58	14	72	23.1	-	0.14
<i>EPC1</i>	5	0	5	1.6	-	0.19
<i>PMS2</i>	5	0	5	1.6	-	0.20
<i>LAMA2</i>	22	0	22	7.1	-	0.22
<i>AP3B2</i>	6	0	6	1.9	-	0.30
<i>PTCH1</i>	11	6	17	5.4	-	0.34
<i>ERCC5</i>	7	0	7	2.2	-	0.42
<i>FANCF</i>	5	0	5	1.6	-	0.44
<i>PIK3CA</i>	21	1	22	7.1	-	0.47
<i>PDZRN4</i>	7	0	7	2.2	-	0.56

Risk score = sum up genes (coefficient*0/1); mutation =1, wild type =0. For example: OS risk score = -0.35*ADCY8+0.07*ALK+... +0.43*TSC1.

Table S2 Cutoff values based on X-title in OS and PFS analysis

	Overall survival (OS)			Progression free survival (PFS)		
	Pt No.	Events	Range	Pt No.	Events	Range
Low risk	38	13	-0.38 thru -0.13	207	117	-0.91 thru 0.00
High risk	274	155	0 thru 0.46	105	78	0.04 thru 1.19
Chi-Square		9.15			30.44	
P value		0.002			<0.001	

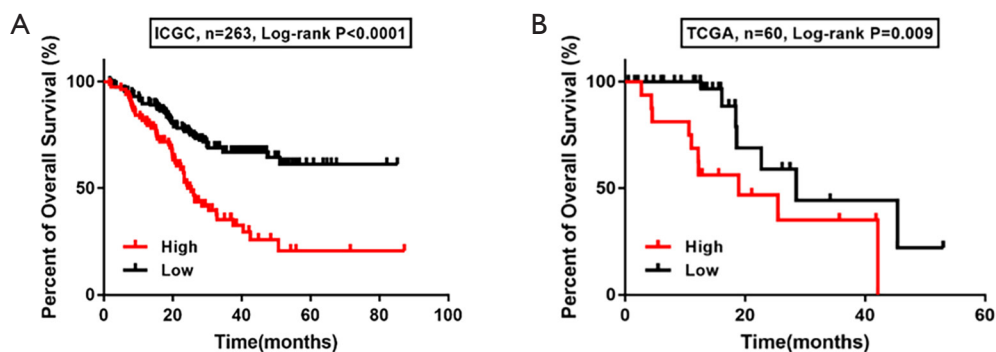


Figure S3 Verification of the risk model in the public database. (A) Comparison of survival curves of high-risk and low-risk groups in ICGC data. (B) Risk score model verified in TCGA data. ICGC, International Cancer Genome Consortium; TCGA, The Cancer Genome Atlas.

Table S3 Difference of gene expression in prediction model

Tag	logFold change	t	P.Value	adj.P.Value
MSH2	0.955421434	12.973787	3.55E-27	9.23E-26
PDZRN4	-2.797635282	-12.703337	2.29E-26	2.97E-25
PMS2	0.577708425	9.8023191	2.97E-18	2.58E-17
MAP3K13	0.683298236	8.47407183	1.10E-14	7.17E-14
PIK3CA	0.546798828	8.04385861	1.44E-13	7.49E-13
ERBB2	-0.940390979	-6.5502086	6.60E-10	2.86E-09
LAMA2	-0.921598582	-6.2652714	2.97E-09	1.10E-08
EPC1	-0.277881644	-4.8392592	2.91E-06	9.47E-06
PMS1	0.305496453	4.63025238	7.24E-06	1.93E-05
CDK8	-0.234367933	-4.6243013	7.43E-06	1.93E-05
ALK	-0.590100126	-3.9464695	0.000116	0.00027418
FANCF	0.3410801	3.89951657	0.00013869	0.00030049
AP3B2	0.705142481	3.65231723	0.00034605	0.00069211
ERCC5	-0.21241243	-3.4774067	0.00064328	0.00119466
PTCH1	0.383879766	2.837049	0.0051087	0.00874771
TSHR	0.419768959	2.81949044	0.00538321	0.00874771
CS	-0.141265076	-2.4226395	0.01646118	0.02517592
NCOR1	-0.077996044	-1.9835203	0.04892388	0.07066783
ARID1A	0.09669169	1.77274301	0.0780671	0.10682867
MED12	0.076218753	1.32440151	0.18715238	0.24329809
ADCY8	0.390736396	0.88031764	0.38176388	0.4511755
DICER1	-0.051965895	-1.1366515	0.25728809	0.31854716
BAP1	0.048765511	0.75285616	0.45258049	0.51161272
TSC1	-0.023436271	-0.5002885	0.61752071	0.66898077
BCR	-0.037372671	-0.4154559	0.67833352	0.70546686
MLL2	-0.018561492	-0.3088409	0.75782179	0.75782179

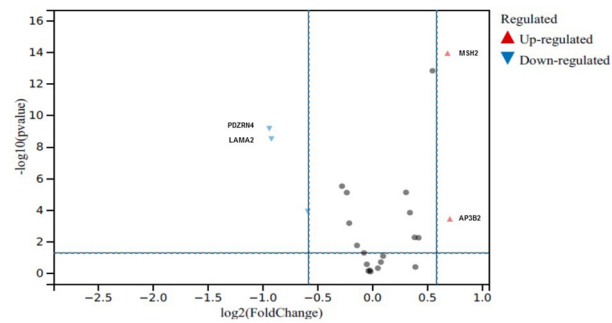


Figure S4 The expression of genes in PFS prediction model. PFS, progression free survival.

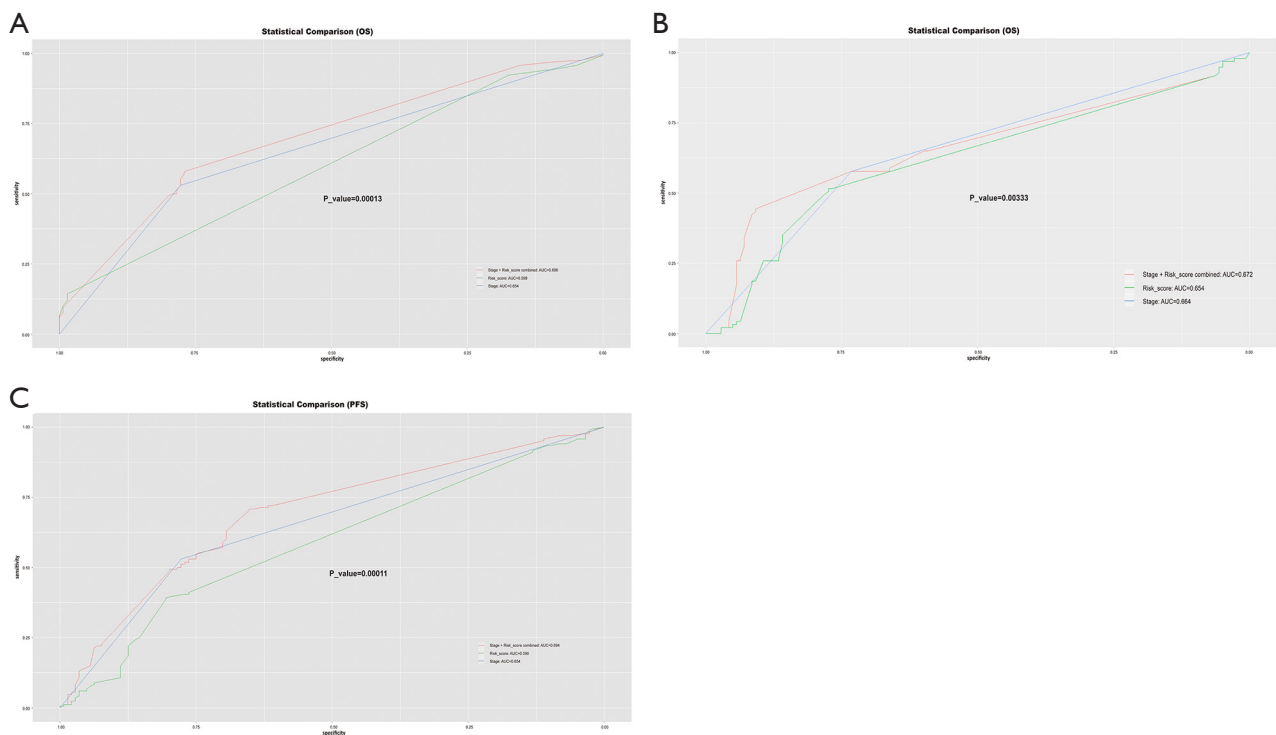


Figure S5 ROC plots of the Lasso regression model. (A) Comparison of the prognostic value for the gene-mutation-based model and TNM stage in OS; (B) comparison of the prognostic value for the gene-mutation-based model and TNM stage in OS for ICGC data; (C) comparison of the prognostic value for the gene-mutation-based model and TNM stage in PFS. The green dashed line represents the ROC curve of the gene mutation model. The blue line represents the ROC curve of TNM stage. The red line represents the combination of the gene mutation model and TNM stage. ROC, Receiver Operating Characteristic; OS, overall survival; PFS, progression free survival; ICGC, International Cancer Genome Consortium.