Supplementary Table 1. Information about datasets applied in this study

Data Source	Type	Platform	pSCC	pADC
GSE50081	frozen	Affy. Plus 2.0	43	127
GSE37745	frozen	Affy. Plus 2.0	24	40
GSE14814	frozen	Affy. U133A	26	32
GSE29016	frozen	Illu. HT	13	37
GSE42127	frozen	Illu. WG	33	94
Total	frozen		139	330
GSE58661	biopsy	Merck RSTA	36	42

Notes: frozen, frozen tissues; biopsy, small biopsy specimens; pADC, pathologically-determined ADC; pSCC, pathologically-determined SCC; Affy. Plus 2.0, Affymetrix Plus 2.0; Affy. U133A, Affymetrix U133A; Illu. WG, Illumina Human WG; Illu. HT, Illumina Human HT.

Supplementary Table 2. The protein expression of AGR2 and KRT5 evaluated by IHC for samples in training and validation sets

No.	Sample set	Pathological lable	AGR2		KRT5	
			expression results	IHC scores	expression results	IHC scores
1	Training set	ADC	+	4	-	1
2	Training set	ADC	+	7	-	0
3	Training set	ADC	+	4	-	0
4	Training set	ADC	+	4	-	0
5	Training set	ADC	+	7	-	2
6	Training set	ADC	+	4	-	0
7	Training set	ADC	-	1	-	0
8	Training set	ADC	+	3	-	1
9	Training set	ADC	+	3	-	0
10	Training set	ADC	+	4	-	0
11	Training set	ADC	+	3	-	0
12	Training set	ADC	+	4	-	0
13	Training set	ADC	+	4	-	1
14	Training set	ADC	+	5	-	0
15	Training set	ADC	+	3	-	0
16	Training set	ADC	+	6	-	2
17	Training set	ADC	+	4	-	0
18	Training set	ADC	+	4	-	0
19	Training set	ADC	+	5	-	0
20	Training set	ADC	+	4	-	1
21	Training set	ADC	+	4	-	0
22	Training set	ADC	+	4	-	0
23	Training set	ADC	-	0	-	0
24	Training set	ADC	+	5	-	2
25	Training set	ADC	+	4	-	0
26	Training set	ADC	+	6	-	0
27	Training set	ADC	+	4	-	0
28	Training set	ADC	+	4	-	0
29	Training set	ADC	+	6	-	0
30	Training set	ADC	+	5	-	0

31	Training set	ADC	+	6	-	0
32	Training set	ADC	+	6	-	0
33	Training set	ADC	+	4	-	0
34	Training set	ADC	+	7	-	0
35	Training set	ADC	+	7	-	2
36	Training set	ADC	+	7	-	2
37	Training set	ADC	+	7	-	0
38	Training set	ADC	-	1	-	0
39	Training set	ADC	+	6	-	0
40	Training set	ADC	-	0	-	0
41	Training set	ADC	+	7	-	0
42	Training set	ADC	+	7	-	0
43	Training set	ADC	+	6	-	0
44	Training set	ADC	+	4	-	0
45	Training set	ADC	+	6	-	2
46	Training set	ADC	+	4	-	0
47	Training set	ADC	+	5	-	0
48	Training set	ADC	+	7	-	1
49	Training set	ADC	+	6	-	1
50	Training set	ADC	+	7	-	0
51	Training set	ADC	+	4	-	0
52	Training set	ADC	+	4	-	0
53	Training set	ADC	+	4	-	1
54	Training set	ADC	+	5	-	0
55	Training set	ADC	+	6	-	0
56	Training set	ADC	+	4	-	0
57	Training set	ADC	+	6	-	1
58	Training set	ADC	+	4	-	0
59	Training set	ADC	+	4	-	0
60	Training set	ADC	+	7	+	4
61	Training set	ADC	+	7	-	0
62	Training set	ADC	+	7	-	0
63	Training set	ADC	+	6	-	2
64	Training set	ADC	+	7	-	1
65	Training set	ADC	+	4	-	0
66	Training set	ADC	+	4	-	2
67	Training set	ADC	+	4	-	2
68	Training set	ADC	+	4	-	0
69	Training set	ADC	+	6	-	0
70	Training set	ADC	+	4	-	1
71	Training set	ADC	+	4	-	0
72	Training set	ADC	+	3	-	0
73	Training set	ADC	+	4	-	0
74	Training set	ADC	+	6	-	0
75	Training set	ADC	+	5	-	0
76	Training set	ADC	+	4	-	0
77	Training set	ADC	-	2	-	0
78	Training set	ADC	+	4	-	0
79	Training set	ADC	+	5	-	0
80	Training set	ADC	+	4	-	0

81	Training set	ADC	+	3	-	0
82	Training set	ADC	+	6	-	1
83	Training set	ADC	+	7	-	0
84	Training set	ADC	+	6	-	0
85	Training set	ADC	+	3	-	1
86	Training set	ADC	+	7	-	0
87	Training set	ADC	+	4	-	0
88	Training set	ADC	+	5	-	1
89	Training set	ADC	+	4	-	0
90	Training set	SCC	-	0	+	7
91	Training set	SCC	-	0	+	5
92	Training set	SCC	-	1	+	5
93	Training set	SCC	-	0	+	7
94	Training set	SCC	-	0	+	7
95	Training set	SCC	-	0	+	6
96	Training set	SCC	-	1	+	4
97	Training set	SCC	-	0	+	7
98	Training set	SCC	-	1	-	1
99	Training set	SCC	-	0	-	1
100	Training set	SCC	-	0	-	0
101	Training set	SCC	-	0	+	7
102	Training set	SCC	-	0	+	7
103	Training set	SCC	-	1	+	7
104	Training set	SCC	-	0	+	6
105	Training set	SCC	-	0	+	7
106	Training set	SCC	-	0	+	7
107	Training set	SCC	-	0	+	3
108	Training set	SCC	-	0	+	5
109	Training set	SCC	-	0	+	6
110	Training set	SCC	+	3	-	2
111	Training set	SCC	-	0	+	6
112	Training set	SCC	-	0	+	6
113	Training set	SCC	-	0	+	6
114	Training set	SCC	-	0	+	7
115	Training set	SCC	-	1	+	7
116	Training set	SCC	-	0	+	3
117	Training set	SCC	-	1	+	6
118	Training set	SCC	-	0	+	7
119	Training set	SCC	-	0	+	7
120	Training set	SCC	-	1	-	2
121	Training set	SCC	-	0	+	6
122	Training set	SCC	-	0	+	5
123	Training set	SCC	-	0	+	5
124	Training set	SCC	-	1	+	5
125	Training set	SCC	-	0	+	6
126	Training set	SCC	-	2	+	5
127	Training set	SCC	-	0	+	6
128	Training set	SCC	-	0	+	5
129	Training set	SCC	-	0	+	3
130	Training set	SCC	-	0	+	3

131	Training set	SCC	-	2	+	7
132	Training set	SCC	-	2	+	5
133	Training set	SCC	-	0	+	6
134	Training set	SCC	-	0	+	5
135	Training set	SCC	-	0	+	7
136	Training set	SCC	-	1	+	7
137	Training set	SCC	-	0	+	6
138	Training set	SCC	-	0	+	7
139	Training set	SCC	-	1	+	7
140	Training set	SCC	-	0	+	6
141	Training set	SCC	-	2	+	5
142	Training set	SCC	-	0	+	6
143	Training set	SCC	-	0	+	3
144	Training set	SCC	-	0	+	6
145	Training set	SCC	-	0	+	6
146	Training set	SCC	-	0	+	3
147	Training set	SCC	-	0	+	4
148	Training set	SCC	-	2	+	7
149	Training set	SCC	-	0	+	7
150	Training set	SCC	-	2	+	6
151	Training set	SCC	-	0	+	6
152	Training set	SCC	-	0	+	3
153	Training set	SCC	-	0	-	2
154	Training set	SCC	-	0	+	7
155	Training set	SCC	-	0	+	6
156	Training set	SCC	-	0	+	7
157	Training set	SCC	-	1	+	7
158	Training set	SCC	-	0	+	5
159	Training set	SCC	-	1	+	5
160	Training set	SCC	-	0	+	4
161	Training set	SCC	-	0	+	3
162	Training set	SCC	-	0	+	3
163	Training set	SCC	-	0	+	7
164	Training set	SCC	-	0	+	6
165	Training set	SCC	-	0	+	6
166	Training set	SCC	-	0	+	7
167	Training set	SCC	+	3	+	7
168	Training set	SCC	-	0	+	3
169	Training set	SCC	-	0	+	4
170	Training set	SCC	-	0	+	4
171	Training set	SCC	-	0	+	6
172	Training set	SCC	-	0	+	7
173	Training set	SCC	-	0	+	5
174	Training set	SCC	-	0	+	5
175	Training set	SCC	-	0	+	3
176	Training set	SCC	-	2	+	5
177	Training set	SCC	-	0	+	4
178	Training set	SCC	-	0	+	7
179	Training set	SCC	-	0	+	4
180	Training set	SCC	+	4	+	4

181	Training set	SCC	-	0	+	3
182	Training set	SCC	-	0	+	5
183	Training set	SCC	-	0	+	6
184	Training set	SCC	-	2	+	7
185	Training set	SCC	-	0	+	6
186	Training set	SCC	-	0	+	4
187	Training set	SCC	-	1	+	4
188	Training set	SCC	-	1	+	7
189	Validation set	ADC	+	4	-	0
190	Validation set	SCC	-	0	+	4
191	Validation set	SCC	-	0	+	6
192	Validation set	ADC	+	3	-	0
193	Validation set	ADC	+	5	-	0
194	Validation set	SCC	-	1	+	7
195	Validation set	ADC	+	3	-	0
196	Validation set	SCC	-	0	+	4
197	Validation set	SCC	-	0	+	3
198	Validation set	SCC	-	0	-	1
199	Validation set	ADC	+	5	-	0
200	Validation set	SCC	-	1	+	4
201	Validation set	SCC	+	3	+	4
202	Validation set	ADC	+	6	-	0
203	Validation set	SCC	-	0	+	4
204	Validation set	ADC	+	4	-	0
205	Validation set	SCC	-	2	+	7
206	Validation set	SCC	-	2	+	6
207	Validation set	SCC	-	2	+	4
208	Validation set	SCC	-	0	+	3
209	Validation set	ADC	+	5	-	0
210	Validation set	SCC	-	0	+	5
211	Validation set	ADC	+	7	-	0
212	Validation set	SCC	-	0	+	5
213	Validation set	SCC	-	2	+	3
214	Validation set	ADC	+	6	-	0
215	Validation set	SCC	-	0	+	3
216	Validation set	SCC	-	0	+	4
217	Validation set	SCC	-	1	+	6
218	Validation set	SCC	-	2	+	4
219	Validation set	SCC	-	0	-	2
220	Validation set	SCC	-	1	+	4
221	Validation set	SCC	-	0	+	4
222	Validation set	ADC	+	5	-	0
223	Validation set	SCC	-	2	+	4
224	Validation set	SCC	-	2	+	3
225	Validation set	SCC	-	1	+	3
226	Validation set	ADC	+	5	-	0
227	Validation set	SCC	-	0	+	3
228	Validation set	SCC	-	2	+	6
229	Validation set	SCC	+	3	+	4
230	Validation set	SCC	-	0	+	4

Supplementary Table 3. Sensitivity, specificity, PPV and NPV of IHC markers in the training set (%)

Markers	Sensitivity	Specificity	PPV	NPV
AGR2	94.4% (84/89)	97.0% (96/99)	97.7% (84/87)	95.1% (96/101)
KRT5	93.9% (93/99)	98.9% (88/89)	98.9% (93/94)	93.6% (88/94)

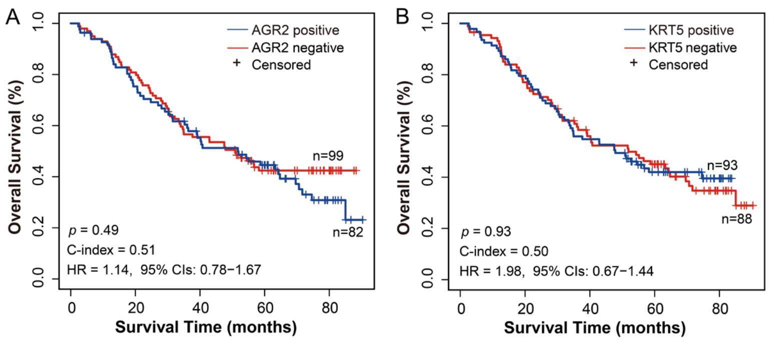
Notes: In this table, we took IHC staining score 3 as the cutoff value. Sensitivity = TP/TP+FN; Specificity = TN/TN+FP; Positive predictive value (PPV) = TP/TP+FP; Negative predictive value (NPV) = TN/TN+FN. FN indicates false negatives; FP, false positives; TN, true negatives; TP, true positives.

Supplementary Table 4. The comparison of two IHC marker combinations in the validation set

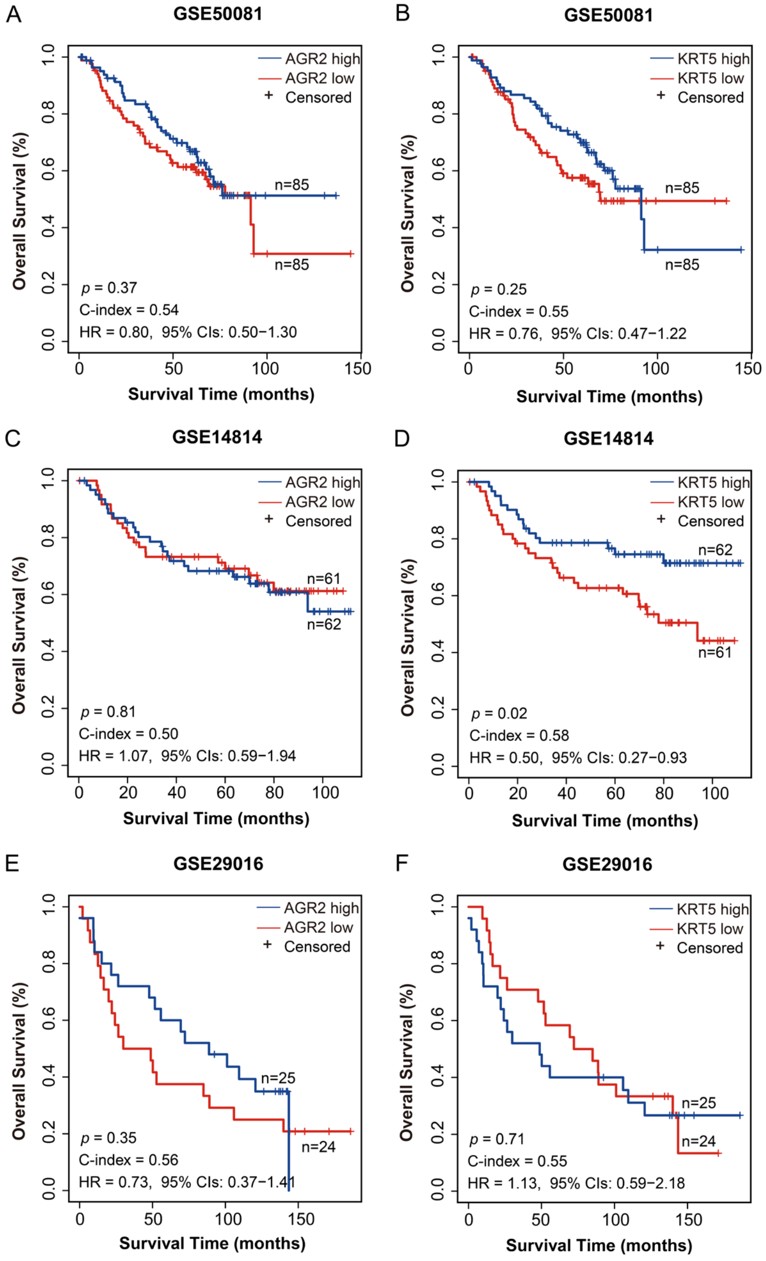
AGR2-KRT5	TTF1-P40		Total
	no. of correct diagnosis	no. of false diagnosis	
no. of correct diagnosis	37	1	38
no. of false diagnosis	1	3	4
Total	38	4	42

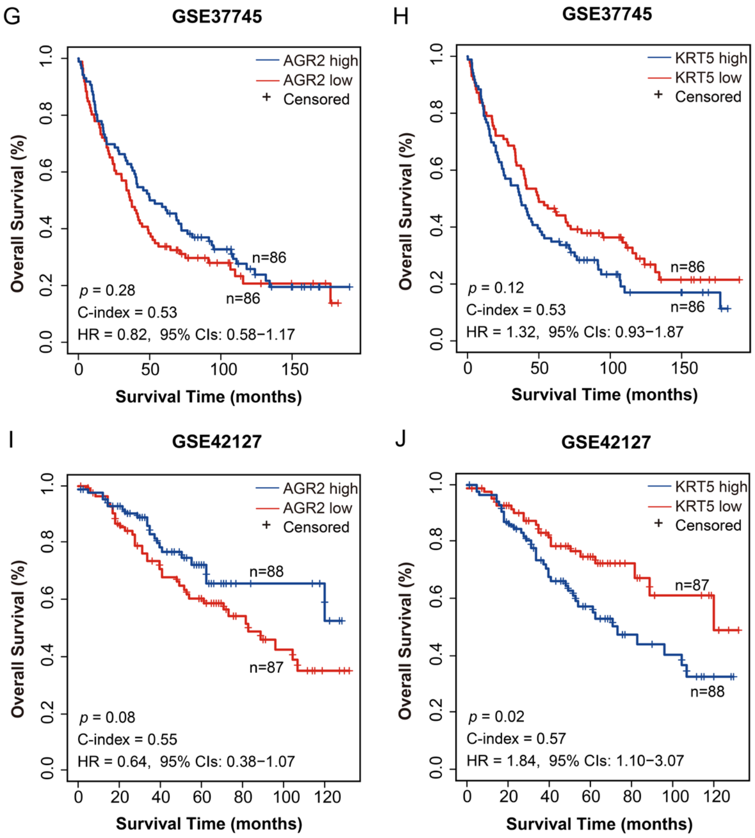
Supplementary Table 5. Associations between marker mRNA expression and clinicopathological parameters in GSE50081

Variable	No. of cases	AGR2		*P*	KRT5		*P*
		No. of pos. (%)	No. of neg. (%)		No. of pos. (%)	No. of neg. (%)	
Patient age							
<60 years	25	8 (32.00%)	17 (68.00%)	0.082	10 (40.00%)	15 (60.00%)	0.387
≥60 years	145	77 (53.10%)	68 (46.90%)		75 (51.72%)	70 (48.28%)	
Gender							
Male	90	51 (56.67%)	39 (43.33%)	0.091	44 (48.89%)	46 (51.11%)	0.878
Female	80	34 (42.50%)	46 (57.50%)		41 (51.25%)	39 (48.75%)	
Smoking history							
Yes	126	64 (50.79%)	62 (49.21%)	0.967	65 (51.59%)	61 (48.41%)	0.791
No	24	11 (45.83%)	13 (54.17%)		11 (45.83%)	13 (54.17%)	
Unable to determine	20	10 (50.00%)	10 (50.00%)		9 (45.00%)	11 (55.00%)	
Histology							
ADC	127	74 (58.27%)	53 (41.73%)	<0.001	46 (36.22%)	81 (63.78%)	<0.001
SCC	43	11 (25.58%)	32 (74.42%)		39 (90.70%)	9.30%	
pT-status							
pT1-2	168	85 (50.60%)	83 (49.40%)	0.497	85 (50.60%)	83 (49.40%)	0.497
pT3-4	2	0	2 (100%)		0	2 (100%)	
pN-status							
pN0-1	170	85 (50.00%)	85 (50.00%)	1	85 (50.00%)	85 (50.00%)	1
pN2-3	0	0	0		0	0	
TNM stage							
I-II	170	85 (50.00%)	85 (50.00%)	1	85 (50.00%)	85 (50.00%)	1
III-IV	0	0	0		0	0	



Supplementary Figure 1. Survival curves of training set based on AGR2 and KRT5 protein expression. Survival curves of positive and negative expressions for AGR2 (A) and KRT5 (B).





Supplementary Figure 2. Survival curves of patients with expression of AGR2 and KRT5 in GEO datasets. Survival curves of overall survival (OS) accordingly for the high and low AGR2 expression groups in GSE50081 (A), GSE14814 (C), GSE29016 (E), GSE37745 (G) and GSE42127 (I). Survival curves of OS respectively for the diverse expression groups of KRT5 in GSE50081 (B), GSE14814 (D), GSE29016 (F), GSE37745 (H) and GSE42127 (J). The high and low expression groups of marker genes were categorized according to the median of the gene expression.