Supplementary Materials

Supplementary methods

RNA extraction and microarray data analysis

Total RNA was extracted using the QIAGEN FFPE RNeasy kit (QIAGEN GmbH, Germany). RNA quality and quantity were analyzed using an Agilent Bioanalyzer 2100 (Agilent Technologies, Santa Clara, USA) and a Nanodrop 2000 spectrophotometer (ThermoFisher Scientific, Waltham, MA, USA). After amplification, fragmentation, and labeling with Ovation FFPE WTA System (NuGEN, San Carlos, CA, USA) and Encore Biotin Module (NuGEN), the total RNA was then hybridized to Affymetrix HTA 2.0 (Affymetrix, Santa Clara, CA, USA) using the GeneChip® Hybridization, Wash and Stain Kit (Affymetrix) following the manufacturer's instructions. All arrays were scanned with the GeneChip® Scanner 3000 (Affymetrix), and raw data were obtained using Command Console Software 4.0 (Affymetrix) with the default settings.

Affymetrix HTA 2.0 has high coverage of probes that can measure coding and noncoding transcripts, containing >6.0 million probes covering > 200,000 coding transcripts and > 40,000 non-coding transcripts. This array has ten probes per exon and four probes per exon-exon splice junction, thus having an average of 100 probes cover one transcript. The raw microarray data was normalized using the Transcriptome Analysis Console (version 4.0.1) with exon-level SST-RMA. After normalization, 893,193 of 913,035 probes remained based on the transcript accession numbers. Then, we used the "*ComBat*" algorithm in SVA package in R language to remove batch effects (1). Here, we only included 558,258 probes representing coding probes for the subsequent analysis. We then used empirical Bayes (eBayes) statistics in the '*limma* package' to determine which probes were significantly differentially expressed between patients with response and non-response to induction chemotherapy (IC) (2), and a total of 23,024 probes (eBayes $P<0.05$) were identified for further analysis.

As RNA extracted from paraffin-embedded samples is highly and randomly fragmented, we also performed Fisher's exact test to confirm whether a transcript was really significantly differentially expressed as described in our previous study (3). We first counted the number of probes that were designed for each transcript and matched the differentially expressed probes to their respective transcripts. Then, the rate of significantly differentially expressed probes of each transcript was evaluated by the Fisher's exact test, and the transcripts with probes that

were significantly enriched (Fisher *P* value <0.05) were screened out. Based on this, we obtained a list of 6,343 differentially expressed transcripts (Fisher *P* value <0.05).

Meanwhile, we calculated the median of the probe expression values of each transcript and defined it as the transcript expression value in each sample. We screened out transcripts which were significantly differentially expressed between patients with response and non-response to IC using the eBayes statistics. We obtained 385 differentially expressed transcripts (empirical fold-change ≥1.5 and eBayes *P* value <0.05). Analyzing together with the above Fisher's exact test, we obtained a total of 185 significantly differentially expressed transcripts (empirical fold-change ≥1.5, eBayes *P* value <0.05, and Fisher *P* value <0.05).

To obtain a unique gene list, we selected the representative transcript of a gene according to the following criteria: if a gene had only one significantly differentially expressed transcript, we selected this transcript; if a gene had two or more significantly differentially expressed transcripts, we selected the transcript with the highest fold-change. Then, we obtained a list of 85 unique genes for further analysis.

To further identify genes that were most strongly related to the efficacy of IC and narrow down the number of the 85 genes for further analysis, we used the least absolute shrinkage and selector operation (LASSO) and support vector machine-recursive feature elimination (SVM-RFE), which are two popular methods for regression with high dimensional predictors (4-6). We first performed LASSO algorithm with the '*glmnet* package' in R, with penalization parameter λ selected by a 10-fold cross-validation approach and minimum mean cross-validated error rule (4). Based on this, we identified 37 candidate genes from the 85 genes. We also used SVM-RFE for the candidate gene selection (5). We used SVM-RFE with the '*e1071* package' in R as follows: we trained the SVM classifier, computed the ranking score for each factor and then removed the factor with the smallest ranking score. Leave-one-out cross-validation was used to identify the number of best-ranked features of the SVM-RFE model. We then selected the top 31 candidate genes from the 85 genes were the first to appear with an accuracy of 95%. Finally, we obtained 43 candidate genes for further analysis through the incorporation of genes selected by LASSO and SVM-RFE.

NanoString data analysis

For the Nanostring nCounter assay, 5 housekeeping genes (B2M, RPLP0, RPL19, PGK1 and ACTB) were selected, as in previous study (3). 300 ng of RNA was hybridized to the NanoString custom codeset, and the subsequent reaction was performed with the nCounter™

Prep Station. The expression counts were obtained using the nCounter™ Digital Analyzer. We first calculated the sum counts of the five housekeeping genes and set a minimum count threshold of 1000 to filter out samples with low quality as previously described (7). We excluded 23 low-quality samples, and 769 samples that passed the quality control step were used for further analysis.

The Nanostring codeset reactions contained 6 positive and 8 negative spike-in controls, and these were used for hybridization and background correction. The counts of each sample were corrected for hybridization variability across samples by multiplying the mean sum of the positive spike-in controls across all of the good-quality samples and dividing by the sum of the positive spike-in controls for that sample. Then, the background correction was performed by subtracting the average of the negative spike-in controls for that sample.

We further normalized the measurable gene species that were loaded in each sample by dividing the counts by the geometric mean of 5 housekeeping genes in that sample and then multiplying by 1000. To exclude genes that were expressed at a level at or close to the background level, we selected genes with the following criterion: at least 20% of the samples had an expression level greater than the mean plus 2 standard deviations of the normalized negative spike-in controls. All 43 genes conformed to these criteria, and their normalized data were log2 transformed for further analyses.

Construction of a 6-gene signature

We performed LASSO logistic regression analysis and decision curve analysis (DCA) to select genes to construct a signature to predict the response to IC in the training cohort (3,4,8). We used the bootstrap method in the penalized logistic model with the R package *glmnet* to determine the robustness of the 43 candidate genes. A total of 1000 bootstrap samples with a sample size of 160 were randomly drawn, with the replacement of data from the training data cohort (80 responders and 80 non-responders). For each bootstrap sample, we built a penalized logistic model using the one-standard error (1se) value of the penalization parameter λ that training with a 10-fold cross-validation approach. We ranked genes by the frequencies at which they were included in the bootstrap LASSO models.

Then, we used DCA, a useful method to evaluate the optimal predictive model, with the R package *rmda* to determinate the net benefit derived from the inclusion of each of the genes in turn, according to the order of the bootstrap LASSO frequency. The DCA showed that the model with 6 genes was superior to the models with fewer than 6 genes across the 0% to 80%

threshold probabilities. Meanwhile, the model with 7 genes did not obtain much more net benefit than the models with 6 genes. In addition, we calculated the difference in the areas under the receiver operating characteristic curves (AUCs) between adjacent gene models. The results showed that the model with 6 genes had the most obvious increase in the AUC compared with the models with 1 to 5 genes, and adding genes to the 6-gene model did not further increase the AUC obviously. Hence, we included 6 genes in the final prediction model. Using the logistic regression method (9), we constructed a gene signature with these 6 genes to predict the response to IC in the training cohort.

References:

1. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*. 2012;28(6): 882-883.

2. Smyth GK. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol*. 2004;3: Article3.

3. Tang X, Li Y, Liang S, et al. Development and validation of a gene expression-based signature to predict distant metastasis in locoregionally advanced nasopharyngeal carcinoma: a retrospective, multicentre, cohort study. *Lancet Oncol*. 2018; 19(3): 382–393.

4. Tibshirani R. Regression Shrinkage and Selection via the Lasso. *J R Stat Soc Series B*. 1996;58: 267–88.

5. Huang ML, Hung YH, Lee WM, Li RK, Jiang BR. SVM-RFE based feature selection and Taguchi parameters optimization for multiclass SVM classifier. *Sci World J*. 2014;2014: 795624.

6. Qiu J, Peng B, Tang Y, et al. CpG methylation signature predicts recurrence in early-stage hepatocellular carcinoma: results from a multicenter study. *J Clin Oncol*. 2017;35(7):734-742.

7. Huet S, Tesson B, Jais JP, et al. A gene-expression profiling score for prediction of outcome in patients with follicular lymphoma: a retrospective training and validation analysis in three international cohorts. *Lancet Oncol*. 2018;19(4): 549–561.

8. Vickers AJ, Elkin EB. Decision Curve Analysis: A Novel Method for Evaluating Prediction Models. *Med Decis Making*. 2006;26(6): 565–574.

9. Xu R, Wei W, Krawczyk M, et al. Circulating tumour DNA methylation markers for diagnosis and prognosis of hepatocellular carcinoma. *Nat Mater*. 2017;16(11): 1155-1161.

Supplementary Table 1: Clinical characteristics of responders or non-responders to induction chemotherapy (IC) in the discovery stage

Variables	Responders (n=71)*	Non-responders (n=24)*
Age (y)		
<45	39 (54.9)	14 (58.3)
≥45	32 (45.1)	10 (41.7)
Sex		
Male	51 (71.8)	22 (91.7)
Female	20 (28.2)	2 (8.3)
T stage		
T2	4 (5.6)	1 (4.2)
T3	46 (64.8)	15 (62.5)
T4	21 (29.6)	8 (33.3)
N stage		
N0	1 (1.4)	1 (4.2)
N1	27 (38.0)	7 (29.2)
N2	29 (40.8)	12 (50.0)
N3	14 (19.7)	4 (16.7)
TNM stage		
III	41 (57.7)	15 (62.5)
IV	30 (42.3)	9 (37.5)
EBV DNA (copies/ml)		
<2000	25 (35.2)	6 (25.0)
≥2000	43 (60.6)	17 (70.8)
NA	3 (4.2)	1 (4.2)

*Data are n (%) unless otherwise stated. TNM: tumor-node-metastasis; EBV: Epstein-Barr virus

Supplementary Table 2: The NanoString Codeset

Gene name	Flags	Selected by LASSO	Selected by SVM-RFE	Accession	Position
ACTB	HK*			NM_001101.2	1011-1110
ADTRP		√	√	NM_032744.3	97-196
AL161418.1		√		ENST00000640280.1	23-122
APOO		√		NM_024122.4	801-900
ATMIN		√		NM_015251.2	417-516
B2M	HK			NM_004048.2	236-335
BCL11A		√	√	NM_018014.3	2585-2684
C19orf57		√	√	ENST00000585755.1	260-359
C7orf25		√	√	NM_001099858.1	1075-1174
CCDC32			√	NM_052849.3	546-645
CDK5R1		√	√	NM_003885.2	419-518
FAM96B			√	NR_024525.2	447-546
DCAF7		√	√	NM_005828.4	831-930
DOPEY1			√	NM_015018.3	1820-1919
DSC3		√		NM_001941.3	6376-6475
FAM72C		√	√	NM_001346071.1	1096-1195
FOXO1		√		NM_002015.3	1527-1626
GAPT		√		NM_152687.2	453-552
GOLGA2		√		NM_004486.4	68-167
HLA-DPA1		√	√	NM_033554.3	306-405
KLRD1			√	NM_007334.2	1253-1352
LOC102723532			√	XM_017030115.1	748-847
LOC401040		√		XM_011512295.1	284-383
LRRD1		√		NM_001161528.1	88-187
NDUFS3		√	√	NM_004551.2	248-347
NFATC2IP		√	√	NM_032815.3	1681-1780
NMNAT1		√	√	NM_022787.3	155-254
OGFRL1		√	√	NM_024576.3	1036-1135
PGK1	HK			NM_000291.2	1031-1130
PHLDA3		√	√	NM_012396.3	533-632
PJA1		√	√	NM_145119.3	952-1051
PLAC8		√	√	NM_016619.2	211-310
PRH1		√	√	NM_001291314.1	322-421
PRMT5		√	√	NM_006109.4	777-876
PTGS2		√		NM_000963.1	496-595
RNF138		√	√	NM_016271.4	463-562

Gene name	Flags	Selected by LASSO	Selected by SVM-RFE	Accession	Position
RPL19	HK			NM_000981.3	316-415
RPLP0	HK			NM_001002.3	251-350
RSRP1		√	√	NM_020317.4	362-461
SLC25A27		√	√	NM_004277.4	1481-1580
SUOX		√	√	NM_000456.2	603-702
TLR8			√	NM_016610.2	2311-2410
TMEM64		√	√	NM_001008495.3	998-1097
TUBA4A		√		NM_006000.2	645-744
USP14		√		NM_005151.3	239-338
ZNF155		√	√	NM_001260486.1	2209-2308
ZNF788P		√	√	NM_001348163.1	1991-2090
ZNF827		√	√	NM_001306215.1	4104-4203

*HK: housekeeping gene

Supplementary Table 3: Bootstrap LASSO analysis to rank 43 candidate genes in order of frequency

Rank	Gene Name	Frequency
1	*AL161418.1*	972
2	*RNF138*	961
3	*OGFRL1*	935
4	*PTGS2*	849
5	*PLAC8*	703
6	*LRRD1*	679
7	*PHLDA3*	674
8	*CDK5R1*	633
9	*KLRD1*	613
10	*GOLGA2*	594
11	*PRMT5*	585
12	*ZNF827*	579
13	*RSRP1*	567
14	*LOC401040*	556
15	*CCDC32*	553
16	*ZNF788P*	518
17	*C7orf25*	471
18	*ADTRP*	468
19	*ATMIN*	411
20	*TMEM64*	406
21	*C19orf57*	358
22	*DSC3*	358
23	*APOO*	345
24	*SLC25A27*	344
25	*NDUFS3*	321
26	*HLA-DPA1*	277
27	*GAPT*	268
28	*PRH1*	254
29	*FAM72C*	210
30	*USP14*	183
31	*NFATC2IP*	165
32	*DCAF7*	163
33	*LOC102723532*	149
34	*BCL11A*	136
35	*NMNAT1*	121
36	*PJA1*	118

Rank	Gene Name	Frequency
37	*TLR8*	103
38	*SUOX*	93
39	*TUBA4A*	84
40	*FAM96B*	46
41	*DOPEY1*	39
42	*ZNF155*	34
43	*FOXO1*	30

Supplementary Table 4: Determination of the number of genes included in the signature

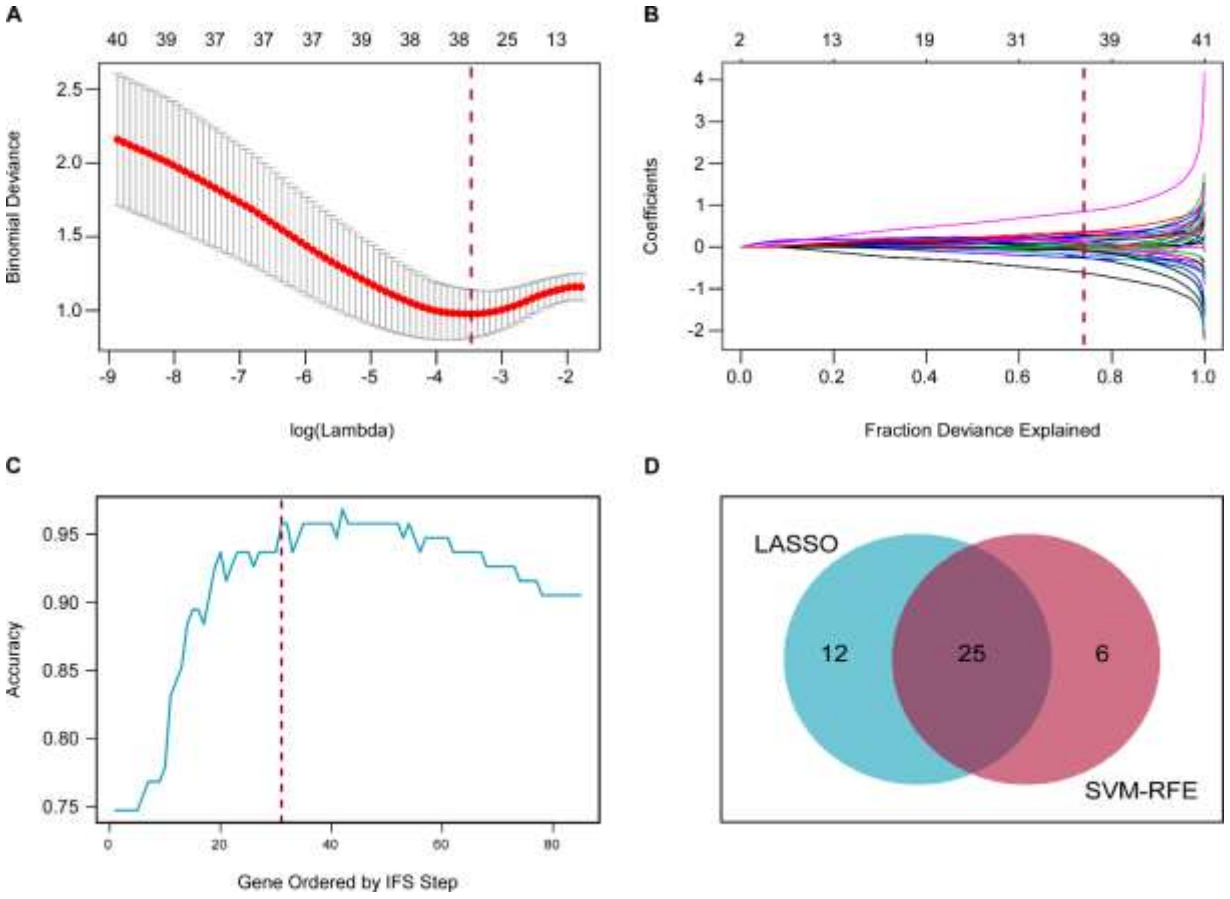
Number of genes included	AUC of the gene signature	ΔAUC*
1-gene model	0.64	
2-gene model	0.69	0.05
3-gene model	0.74	0.05
4-gene model	0.79	0.05
5-gene model	0.79	0.00
6-gene model	0.87	0.08
7-gene model	0.87	0.00
8-gene model	0.88	0.01
9-gene model	0.87	-0.01
10-gene model	0.89	0.02

*ΔAUC = the AUC of the (n+1)-gene model – the AUC of the n-gene model. We included genes in the model in the order of the bootstrap LASSO outcome, and the AUC of the 6-gene model increased most obviously compared with the models with fewer than 6 genes, and further adding genes to the 6-gene model did not dramatically increase the AUC. AUC, area under receiver the operating characteristic curves.

Supplementary Table 5: Multivariable logistic regression analysis of the 6-gene signature with the short-term tumor response to induction chemotherapy (IC) in the training, clinical trial, and external independent cohorts.

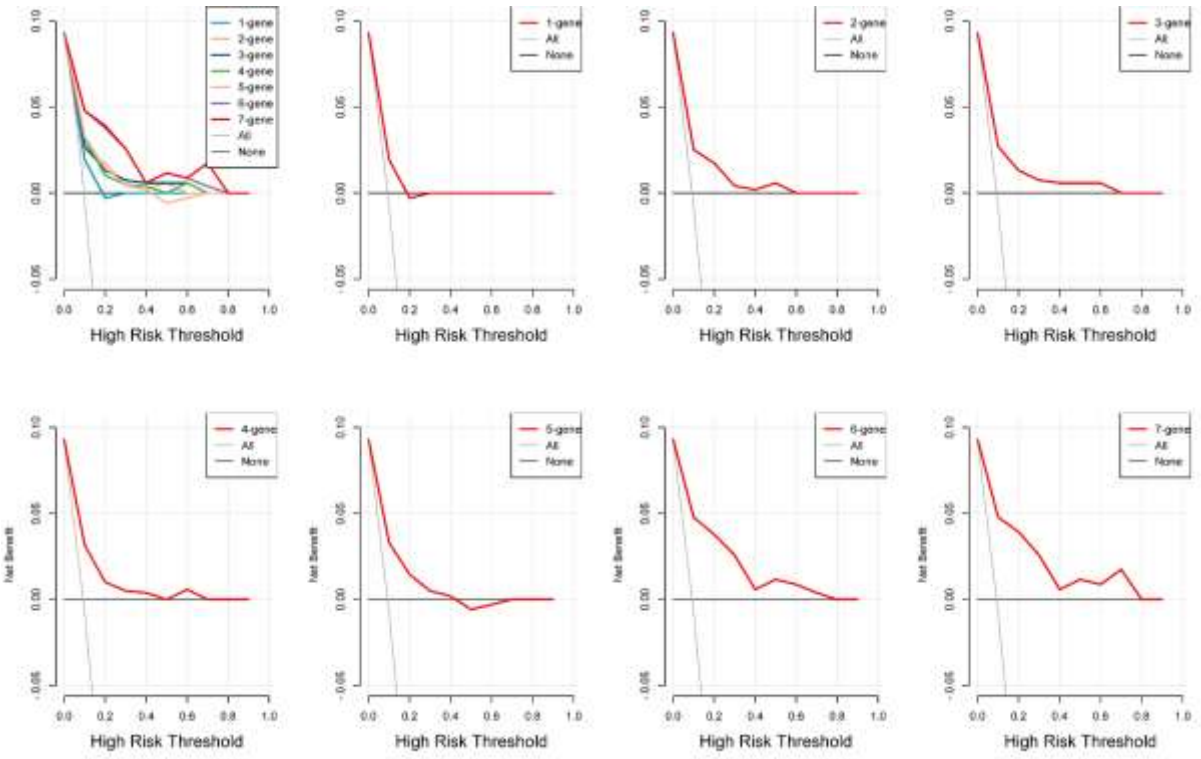
	Multivariable analysis	
Cohort	OR (95% CI)*	*P* value*
Training cohort	29.69 (6.11-144.23)	<0.001
Clinical trial cohort	22.68 (2.20-234.42)	0.009
External independent cohort	21.86 (4.31-111.02)	<0.001

*OR, 95% CI and *P* values were calculated using an adjusted multivariable logistic regression model, including 6-gene signature (no-benefit *vs*. benefit), age (≥ 45 years *vs*. < 45 years), sex (male *vs*. female), T stage (T3–4 *vs*. T1–2), N stage (N2-3 *vs*. N0-1), EBV DNA (≥ 2000 *vs*. < 2000) as covariables. OR, odds ratio; CI, confidence interval.

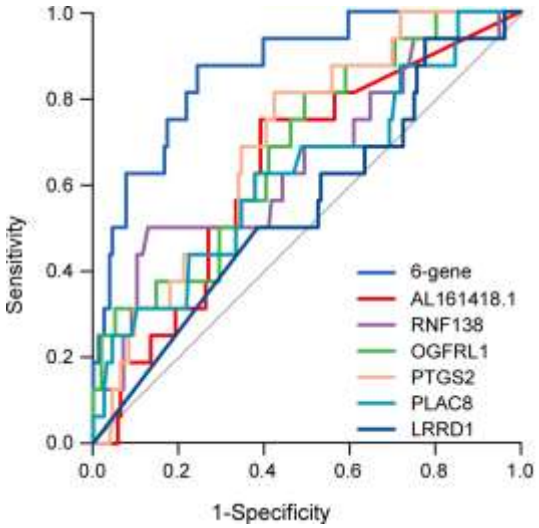


Supplementary Figure 1: Candidate gene selection by two algorithms.

(A) LASSO algorithm was performed with penalization parameter λ selected by a 10-fold cross-validation approach and minimum mean cross-validated error rule. The dotted vertical lines represented the optimal values by minimum criteria; (B) LASSO coefficient of the variables selected with the minimum criteria. The vertical line was the optimal value by minimum criteria and results in 37 non-zero coefficients; (C) SVM-RFE was used to rank genes and then leave-one-out cross-validation analyses were performed to calculate the accuracy of the SVM-RFE model. The dotted vertical lines represented the including gene number that first to appear with an accuracy of 95%; (D) Candidate genes were obtained through the incorporation of genes selected by LASSO and SVM-RFE. LASSO, least absolute shrinkage and selector operation; SVM-RFE, support vector machine-recursive feature elimination; IFS, incremental feature selection.

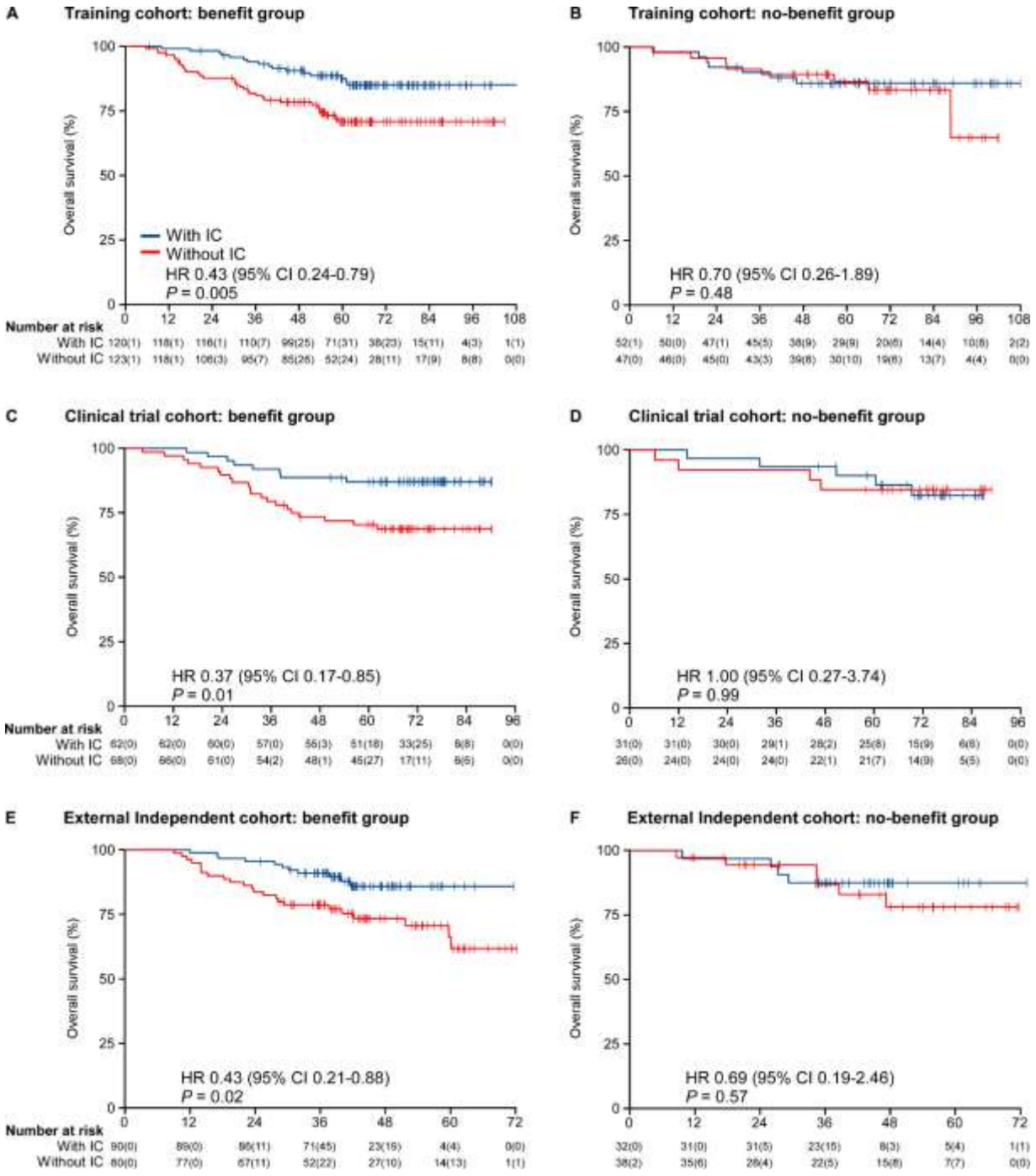


Supplementary Figure 2: Determination of the number of genes included in the signature. Decision curve analysis showed that the 6-gene model had the greatest net benefit across the 0% to 80% threshold probabilities compared with the models with 1 to 5 genes, and adding genes to the 6-gene model did not obviously increase the net benefit.



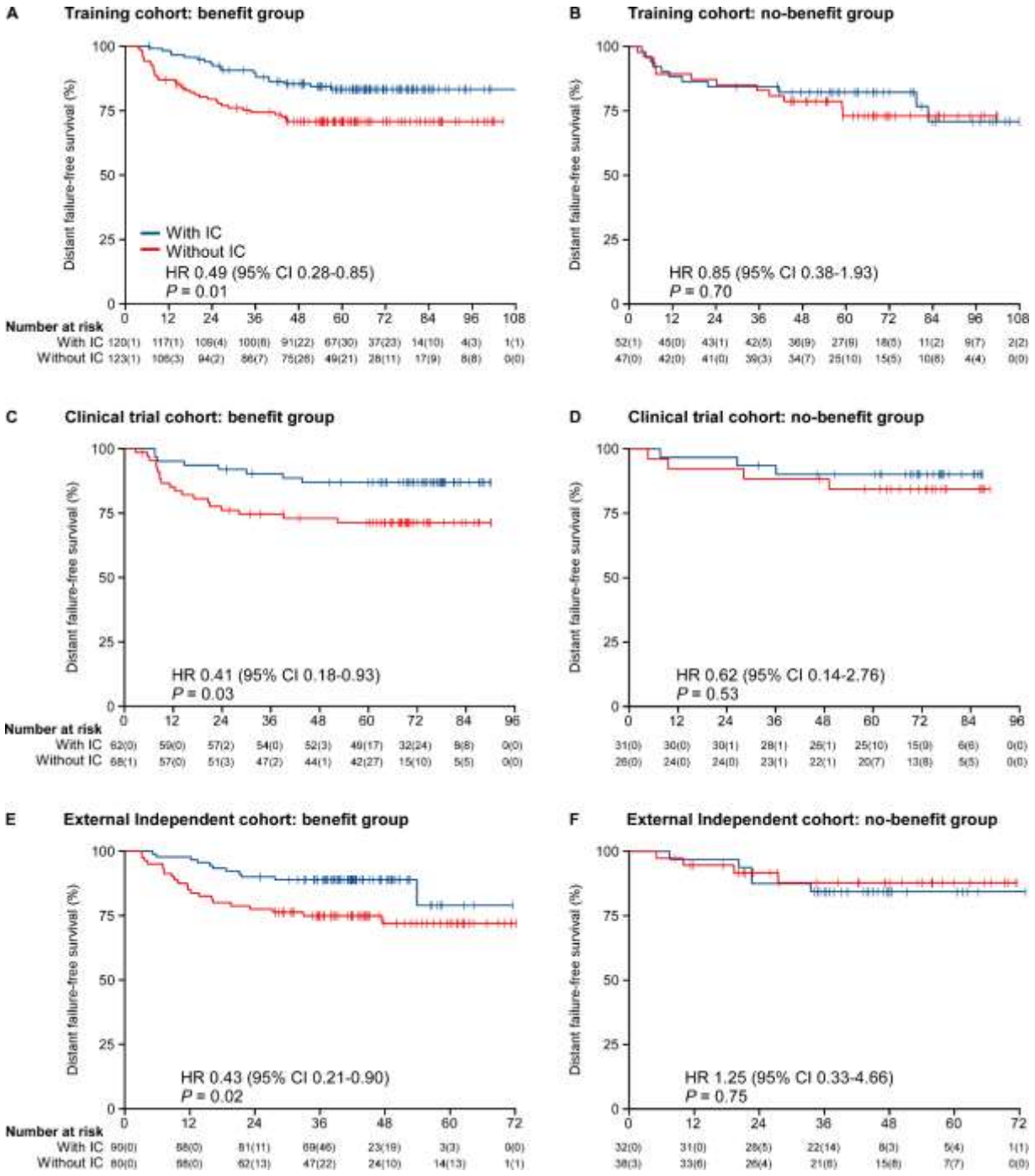
	ROC analysis		
Variable	AUC	95% CI	*P*-value
6-gene signature	0.87	0.78-0.94	Ref
AL161418.1	0.64	0.51-0.76	0.003
RNF138	0.65	0.49-0.80	0.013
OGFRL1	0.69	0.55-0.81	0.002
PTGS2	0.69	0.57-0.80	0.006
PLAC8	0.63	0.48-0.77	<0.001
LRRD1	0.55	0.40-0.69	<0.001

Supplementary Figure 3: Prediction of the short-term response to induction chemotherapy by the 6-gene signature and each single gene in the training cohort. The 95% CI of the AUC and *P* value were estimated using the bootstrap method. AUC: area under receiver operating characteristic curve; CI: confidence interval.



Supplementary Figure 4: Kaplan-Meier curves of overall survival according to treatment with or without induction chemotherapy (IC) in the benefit or no-benefit groups.

(A) Benefit group of the training cohort (n=243); (B) No-benefit group of the training cohort (n=99); (C) Benefit group of the clinical trial cohort (n=130); (D) No-benefit group of the clinical trial cohort (n=57); (E) Benefit group of the external independent cohort (n=170); (F) No-benefit group of the external independent cohort (n=70). HR, hazard ratio; CI, confidence interval. We calculated *P* values with the unadjusted log-rank test and hazard ratios (HRs) with univariable Cox regression analysis.



Supplementary Figure 5: Kaplan-Meier curves of distant failure-free survival according to treatment with or without induction chemotherapy (IC) in the benefit or no-benefit groups.

(A) Benefit group of the training cohort (n=243); (B) No-benefit group of the training cohort (n=99); (C) Benefit group of the clinical trial cohort (n=130); (D) No-benefit group of the clinical trial cohort (n=57); (E) Benefit group of the external independent cohort (n=170); (F) No-benefit group of the external independent cohort (n=70). HR, hazard ratio; CI, confidence interval. We calculated *P* values with the unadjusted log-rank test and hazard ratios (HRs) with univariable Cox regression analysis.