CHEST

Independent Validation of Early-Stage Non-Small Cell Lung Cancer Prognostic Scores Incorporating Epigenetic and Transcriptional Biomarkers With Gene-Gene Interactions and Main Effects

Ruyang Zhang, PhD; Chao Chen, BS; Xuesi Dong, MS; Sipeng Shen, PhD; Linjing Lai, MS; Jieyu He, MS; Dongfang You, BS; Lijuan Lin, MS; Ying Zhu, BS; Hui Huang, BS; Jiajin Chen, BS; Liangmin Wei, BS; Xin Chen, BS; Yi Li, PhD; Yichen Guo, PhD; Weiwei Duan, PhD; Liya Liu, PhD; Li Su, BS; Andrea Shafer, MPH; Thomas Fleischer, PhD; Maria Moksnes Bjaanæs, PhD; Anna Karlsson, PhD; Maria Planck, PhD; Rui Wang, PhD; Johan Staaf, PhD; Åslaug Helland, PhD; Manel Esteller, PhD; Yongyue Wei, PhD; Feng Chen, PhD; and David C. Christiani, MD, FCCP

BACKGROUND: DNA methylation and gene expression are promising biomarkers of various cancers, including non-small cell lung cancer (NSCLC). Besides the main effects of biomarkers, the progression of complex diseases is also influenced by gene-gene (G×G) interactions.

RESEARCH QUESTION: Would screening the functional capacity of biomarkers on the basis of main effects or interactions, using multiomics data, improve the accuracy of cancer prognosis?

STUDY DESIGN AND METHODS: Biomarker screening and model validation were used to construct and validate a prognostic prediction model. NSCLC prognosis-associated biomarkers were identified on the basis of either their main effects or interactions with two types of omics data. A prognostic score incorporating epigenetic and transcriptional biomarkers, as well as clinical information, was independently validated.

RESULTS: Twenty-six pairs of biomarkers with G×G interactions and two biomarkers with main effects were significantly associated with NSCLC survival. Compared with a model using clinical information only, the accuracy of the epigenetic and transcriptional biomarker-based prognostic model, measured by area under the receiver operating characteristic curve (AUC), increased by 35.38% (95% CI, 27.09%-42.17%; $P = 5.10 \times 10^{-17}$) and 34.85% (95% CI, 26.33%-41.87%; $P = 2.52 \times 10^{-18}$) for 3- and 5-year survival, respectively, which exhibited a superior predictive ability for NSCLC survival ($AUC_{3\ year}$, 0.88 [95% CI, 0.83-0.93]; and $AUC_{5\ year}$, 0.89 [95% CI, 0.83-0.93]) in an independent Cancer Genome Atlas population. G×G interactions contributed a 65.2% and 91.3% increase in prediction accuracy for 3- and 5-year survival, respectively.

INTERPRETATION: The integration of epigenetic and transcriptional biomarkers with main effects and G×G interactions significantly improves the accuracy of prognostic prediction of early-stage NSCLC survival.

CHEST 2020; 158(2):808-819

KEY WORDS: early stage; interaction; multiomics; non-small cell lung cancer; prognostic score

ABBREVIATIONS: AUC = area under the receiver operating characteristic curve; C_{index} = concordance index; CPTAC = Clinical Proteomic Tumor Analysis Consortium; GEO = Gene Expression Omnibus; G×G = gene-gene interactions; HR = hazard ratio; ISIS = iterative sure independence screening; KEGG = Kyoto Encyclopedia of Genes and Genomes; LUAD = lung adenocarcinoma; LUSC = lung squamous cell carcinoma; MGH = Massachusetts General Hospital; NSCLC = non-small cell lung cancer; QC = quality control; ROC = receiver operating characteristic; SIS = sure independence screening; TCGA = the Cancer Genome Atlas

AFFILIATIONS: From the Department of Biostatistics (Drs Zhang, Shen, Duan, Y. Wei, and F. Chen; Messrs C. Chen, You, Huang, J. Chen, L. Wei, and X. Chen; and Mss Dong, Lai, He, Lin, and Zhu),

Lung cancer is a leading cause of cancer-related death worldwide and was estimated to cause 1.76 million deaths in 2018.[1] The 5-year survival rate among patients with lung cancer remains relatively low, ranging from 4% to 17% depending on clinical characteristics.[2] Compared with patients diagnosed with late-stage disease, early-stage patients often have a considerably more favorable prognosis. However, significant heterogeneity in clinical prognosis is observed for patients with early-stage non-small cell lung cancer (NSCLC) with similar clinical characteristics, which indicates the importance of understanding molecular mechanisms.[3] Identifying molecular changes in oncogene and/or tumor suppressor genes that are associated with NSCLC survival is helpful for developing targeted therapies to prolong patients' survival time.

DNA methylation is a heritable, reversible, epigenetic modification that affects the spatial conformation of DNA and regulates gene expression.[4,5] DNA methylation is a molecular biomarker and may be a therapeutic target for the treatment of cancer.[6,7] In addition, gene-gene (G×G) interactions have long been recognized to regulate the progression of complex diseases, including NSCLC.[8] The development of cancer may be related to interactions between several key genes.[9] Lung cancer prognosis-associated biomarkers have been proposed on the basis of omics data, including DNA methylation,[10] gene expression,[11] microRNA,[12] and long noncoding RNA.[13] However, most studies are limited to a single type of omics data, which results in less accurate prognostic models.[14] For example, our previous integrative omics study of the *BTG2* gene showed that this gene could slightly improve the prediction accuracy of early-stage NSCLC survival.[6] However, a large-scale integrative analysis of multiomics data has identified genes with either important main effects or gene-gene (G×G) interactions, based on which a more accurate prognostic prediction model of NSCLC can be constructed.

Specifically, we used a two-stage study design and performed an integrative analysis of pan-cancer-related genes to identify prognostic biomarkers with either a main effect or G×G interactions using epigenome and transcriptome data from multiple study centers. We then built a prognostic prediction model for early-stage NSCLC by incorporating both selected epigenetic and transcriptional biomarkers.

Methods

Only patients with early-stage (stage I or II) lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC) were included in our study. DNA methylation data were harmonized from five international study centers, including Harvard, Spain, Norway,

Center for Global Health, School of Public Health, Nanjing Medical University, Nanjing, China; the Department of Environmental Health (Drs Zhang, Shen, Guo, Y. Wei, and Christiani; Mss Dong and Lin; and Messrs You and Su), Harvard T. H. Chan School of Public Health, Boston, MA; the China International Cooperation Center for Environment and Human Health (Drs Zhang, Shen, Y. Wei, and F. Chen), Nanjing Medical University, Nanjing, China; the Department of Medical Oncology (Drs Zhang and Wang), Jinling Hospital, School of Medicine, Nanjing University, Nanjing, China; the Department of Epidemiology and Biostatistics (Ms Dong), School of Public Health, Southeast University, Nanjing, China; the Department of Biostatistics (Dr Li), University of Michigan, Ann Arbor, MI; the Department of Biostatistics (Dr Guo), Harvard T. H. Chan School of Public Health, Boston, MA; the Department of Bioinformatics (Dr Duan), School of Biomedical Engineering and Informatics, Nanjing Medical University, Nanjing, China; the Department of Preventive Medicine (Dr Liu), Medical School of Ningbo University, Ningbo, China; the Pulmonary and Critical Care Division, Department of Medicine (Ms Shafer and Dr Christiani), Massachusetts General Hospital and Harvard Medical School, Boston, MA; the Department of Cancer Genetics (Drs Fleischer, Moksnes Bjaanæs, and Helland), Institute for Cancer Research, Oslo University Hospital, Oslo, Norway; the Division of Oncology and Pathology, Department of Clinical Sciences (Drs Karlsson, Planck, and Staaf), Lund and CREATE Health Strategic Center for Translational Cancer Research, Lund University, Lund, Sweden; the Institute of Clinical Medicine (Dr Helland), University of Oslo, Oslo, Norway; the Josep Carreras Leukemia Research Institute (Dr Esteller), Badalona, Barcelona, Spain; the Centro de Investigacion Biomedica en Red Cancer (Dr Esteller), Madrid, Spain; the Institucio Catalana de Recerca i Estudis Avançats (Dr Esteller), Barcelona, Spain; the Physiological Sciences Department (Dr Esteller), School of Medicine and Health Sciences, University of Barcelona, Barcelona, Spain; the State Key Laboratory of Reproductive Medicine (Dr F. Chen), Nanjing Medical University, Nanjing, China; and the Jiangsu Key Laboratory of Cancer Biomarkers, Prevention and Treatment (Dr F. Chen), Cancer Center, Collaborative Innovation Center for Cancer Personalized Medicine, Nanjing Medical University, Nanjing, China.

Dr Zhang, Mr C. Chen, and Ms Dong contributed equally to this work.

Drs Zhang, Y. Wei, and F. Chen contributed equally to this work as corresponding authors.

Dr Christiani is the senior author who supervised the work.

FUNDING/SUPPORT: This study was funded by the National Natural Science Foundation of China [81530088 to F. C., 81973142 to Y. W.], the National Key Research and Development Program of China [2016YFE0204900 to F. C.], the Natural Science Foundation of Jiangsu Province [BK20191354 to R. Z.], the Natural Science Foundation of the Jiangsu Higher Education Institutions of China [18KJB310011 to R. Z.], the China Postdoctoral Science Foundation [2018M633767 to R. Z.], the NIH [CA209414, CA092824, and ES000002 to D. C. C., AG056764 to Y. L.], and the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD). R. Z. was partially supported by the Outstanding Young Teachers Training Program of Nanjing Medical University.

CORRESPONDENCE TO: Feng Chen, PhD, SPH Bldg, Room 412, 101 Longmian Ave, Nanjing, Jiangsu 211166, China; e-mail: fengchen@njmu.edu.cn

DOI: https://doi.org/10.1016/j.chest.2020.01.048

Sweden, and the Cancer Genome Atlas (TCGA). Gene expression data were composed of four datasets from the Gene Expression Omnibus (GEO) and TCGA.

Harvard: The Harvard cohort consisted of patients seen at Massachusetts General Hospital (MGH), and histologically confirmed as having primary NSCLC, recruited since 1992.[15] We profiled 151 early-stage patients from this cohort. A lung pathologist at MGH evaluated each specimen for the amount (tumor cellularity, > 70%) and quality of tumor cells. The specimens were classified histologically according to World Health Organization criteria. The institutional review boards at the Harvard T. H. Chan School of Public Health and MGH approved the study. All patients provided written informed consent.

Spain: The Spanish cohort included 226 patients with early-stage NSCLC recruited from eight subcenters between 1991 and 2009.[10] Patients provided written consent and tumors were surgically collected. This study was approved by the Bellvitge Biomedical Research Institute institutional review boards.

Norway: The Norwegian cohort consisted of 133 patients with early-stage NSCLC from Oslo University Hospital, recruited between 2006 and 2011.[16] The project was developed with the approval of the Oslo University Institutional Review Board and regional ethics committee (S-05307). All patients provided informed consent. Tumor tissues were snap frozen in liquid nitrogen and stored at –80°C until DNA isolation.

Sweden: Tumor DNA was collected from 103 patients with early-stage NSCLC, including 80 patients with LUAD and 23 patients with LUSC, at the Skåne University Hospital in Lund, Sweden.[17] The study was developed under the approval of the Regional Ethical Review Board in Lund, Sweden (Registration nos. 2004/762 and 2008/702).

TCGA: A total of 332 LUAD and 285 LUSC with full DNA methylation, survival time, and covariates data were included. Level 1 HumanMethylation450 DNA methylation data from patients with early-stage NSCLC were downloaded on October 1, 2015.

GEO: Transcriptome information from 425 patients with early-stage NSCLC was profiled using the Affymetrix Human Genome U133A Plus 2.0 Array (e-Table 1). Only data from patients with available survival time, clinical stage, and tumor tissue expression values were analyzed.

Quality Control for DNA Methylation Data

DNA methylation was assessed with Illumina Infinium HumanMethylation450 BeadChips (Illumina Inc.). Raw image data were imported into GenomeStudio Methylation Module V1.8 (Illumina Inc.) to calculate methylation signals and to perform normalization, background subtraction, and quality control (QC). Unqualified probes were excluded if they fitted any one of the following quality control criteria: (1) failed detection ($P > .05$) in $\geq$ 5% samples; (2) coefficient of variance < 5%; (3) all samples were methylated or all were unmethylated; (4) common single-nucleotide polymorphisms located in probe sequence or in 10-bp flanking regions; (5) cross-reactive probes[18]; or (6) data did not pass QC in all centers. Samples with > 5% undetectable probes were excluded. Methylation signals were further processed for quantile normalization (*betaqn* function in R package *minfi*) as well as type I and II probe correction (*BMIQ* function in R package *lumi*). They were adjusted for batch effects (*ComBat* function in R package *sva*) according to the best pipeline by a comparative study.[19] Details of the QC process are described in e-Figure 1.

Quality Control for Gene Expression Data

The TCGA workgroup completed the mRNA sequencing data processing and QC. Raw counts were normalized using RNA-sequencing by expectation maximization. Level 3 gene quantification data were downloaded from the TCGA data portal and were further checked for quality. Gene probes were excluded if the missing rate > 80%, and the batch effect was corrected with *ComBat*. The expression value of each gene was transformed on a $\log_2$ scale and standardized before association analysis.

DNA methylation and gene expression of 719 pan-cancer-related genes were then used for subsequent association analysis. Gene symbols for the 719 pan-cancer-related genes were obtained from the Catalogue of Somatic Mutations in Cancer (COSMIC). After QC, there were 12,806 CpG probes identified for association analysis. CpG probes from five genes (*BTG2*,[6] *KDM*,[7] *EGLN2*,[8] *LRRC3B*,[15] and *SIPA1L3*[20]) reported in our previous study were also included.

Statistical Analysis

The flow of analysis is depicted in Figure 1. Epigenetic and transcriptional analyses were performed simultaneously, and a discovery phase and validation phase were used to identify NSCLC prognostic biomarkers. In each procedure, we conducted analysis of both the main effects and gene-gene interactions among biomarkers. Patients having DNA methylation data from Harvard, Spain, Norway, and Sweden, as well as patients having gene expression data from GEO, were assigned to the discovery phase for epigenetic analysis and transcriptional analysis, respectively. Patients having two types of omics data from TCGA were assigned to the validation phase.

For the main effect analysis, we used sure independence screening (SIS) and LASSO Cox penalized regression to screen biomarkers with main effects that were relevant to survival, using the R package *SIS*. SIS LASSO is a two-stage procedure. At the first stage, SIS selects the biomarkers with the strongest marginal associations with survival. At the second stage, LASSO was used to perform variable selection and parameter estimation simultaneously among the biomarkers selected at the first stage. During the LASSO procedure, tuning parameter selection was based on Bayesian information criteria. To capture biomarkers that might be missed at the first stage, we repeatedly applied the SIS LASSO algorithm to the remaining unselected biomarkers until no new biomarkers can be recruited.[21] This iterative procedure is termed iterative SIS (ISIS) LASSO. To account for the biologic heterogeneity between LUAD and LUSC, we used a histology-stratified multivariate Cox proportional hazards model to test these biomarkers, using the R package *survival*. The stratified model adjusted for the differences between LUAD and LUSC in baseline hazards. The other covariates adjusted in the model were age, sex, study center, clinical stage, and smoking status.

For the G×G interaction analysis, a histology-stratified multivariate Cox proportional hazards model adjusted for the aforementioned covariates was applied to identify biomarkers with G×G interactions. The *P* value thresholds for multiple testing were established by the Bonferroni method, which set the significance level to .05 divided by the number of tests. This way, the overall type I error would be controlled at the .05 level. In our study, the significance level of G×G interaction analysis of epigenetic and transcriptional biomarkers was defined as $6.10 \times 10^{-10} = 0.05/(12{,}806 \times 12{,}805/2)$ and $1.94 \times 10^{-7} = 0.05/(719 \times 718/2)$, respectively.

Significant biomarkers observed in the discovery phase were further confirmed in the validation phase and were retained if the *P* value was $\leq$ .05 and there was consistent direction of the effect across two phases. We also performed a test of proportional hazards assumption for each significant biomarker. The hazard ratio (HR) and 95% CI were described as per 1% level of DNA methylation or gene expression increment. Sensitivity analysis was performed to

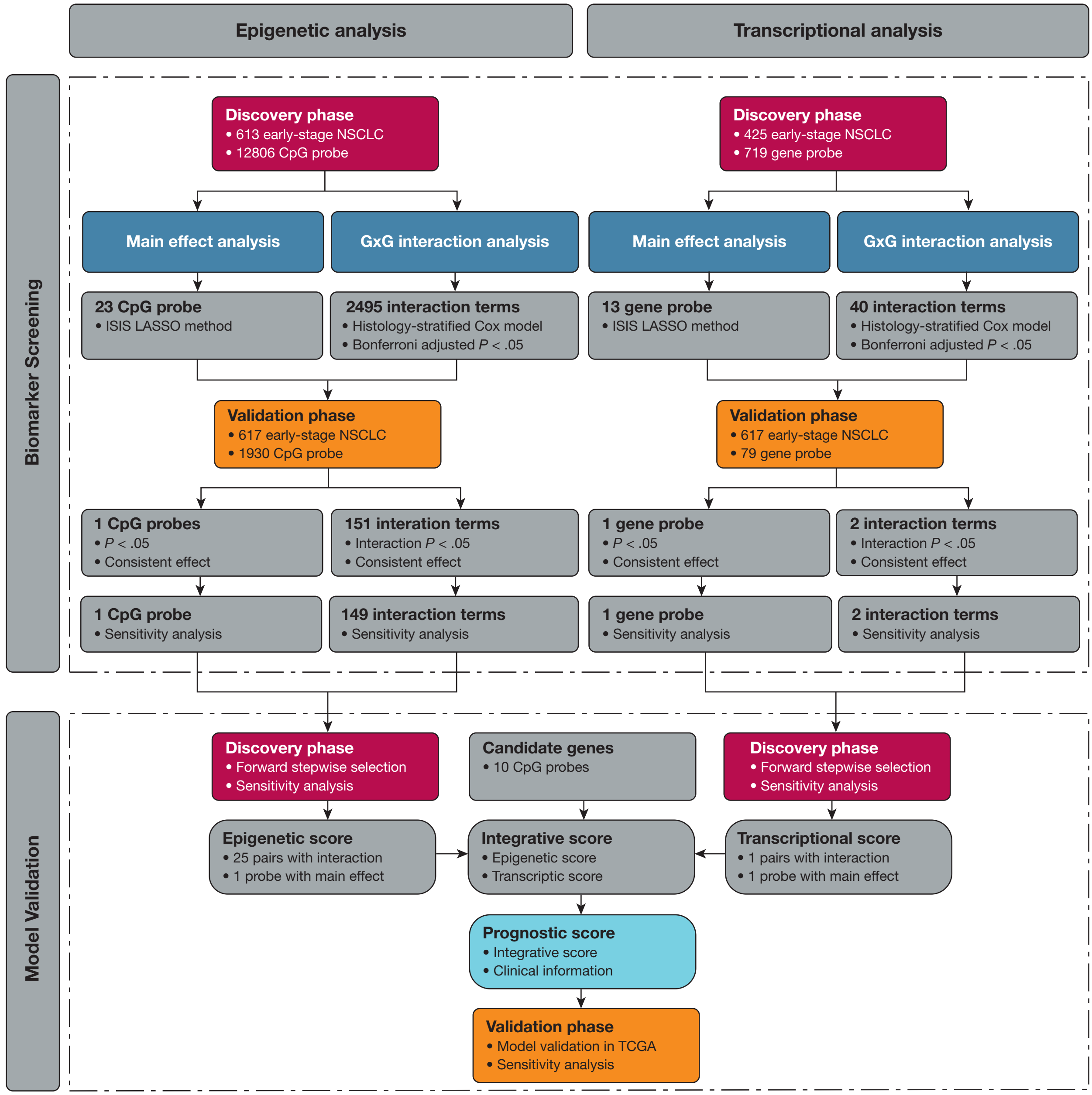


Figure 1 – *Flow chart of study design and statistical analyses. In the epigenetic analysis, patients with lung adenocarcinoma and lung squamous cell carcinoma from the Harvard, Spain, Norway, and Sweden cohorts were used in the discovery phase for screening, whereas data from the Cancer Genome Atlas (TCGA) was used for validation. In transcriptional analysis, gene expression data from Gene Expression Omnibus and TCGA were used in the discovery phase and the validation phase, respectively. Both main effect and G×G interaction analyses were performed. G×G = gene by gene; NSCLC = non-small cell lung cancer.*

confirm these robustly significant biomarkers. Patients were excluded if their DNA methylation (logit_2 transformed) or expression ($\log_2$ transformed) values were out of range, based on mean ± 3 × SD.

For those identified biomarkers, we applied a forward stepwise regression strategy to build up a multibiomarker Cox proportional hazards model in the discovery phase, which was then validated in TCGA samples. In the forward stepwise regression, a likelihood ratio test was applied to test the main effect or G×G interaction of biomarkers if $P_{\text{entry}} \leq .05$ and $P_{\text{elimination}} > .05$. Sensitivity analysis was also performed using two different thresholds: .10 and .15.

Epigenetic and transcriptional scores were calculated on the basis of a weighted linear combination of individual values of the DNA methylation and gene expression, with weights derived from the Cox model. Integrative scores were synthesized by epigenetic and transcriptional scores. Finally, the prognostic score was defined as the linear combination of clinical information and integrative score (see e-Appendix 1).

Kaplan-Meier survival curves adjusted for the covariates were drawn to represent the survival difference among patients with different scores. We predicted 3- and 5-year overall survival of patients, using the nearest neighbor method for time-to-event data.[22] The accuracy of the prediction is presented using a receiver operating characteristic (ROC) curve and was measured by area under the ROC curve (AUC), computed by the R package *survivalROC*. The prediction accuracy was confirmed with an independent TCGA population in the validation phase. The 95% CI and *P* value of the AUC improvement were calculated on the basis of 1,000-time bootstrap resampling. Stratification analysis of prognostic scores was carried out within subgroups stratified by age, sex, smoking status, clinical stage, and histology. The concordance index (C_{index}), an average accuracy of predictive survival across follow-up years, as well as the 95% CI, which ranges from 0.5 to 1.0, were calculated to estimate the predictive performance.[23] A nomogram was generated with R package *rms* to facilitate application of our model.

We assessed the potential functions of the identified genes at the protein level by taking advantage of limited public resources. First, we evaluated the association between protein expression and gene expression, using the reverse-phase protein array from the TCGA database. Second, we performed differential expression analysis between tumor and normal tissues, and further investigated the main effects of genes and G×G interactions between genes on LUAD survival, using the Clinical Proteomic Tumor Analysis Consortium (CPTAC) database. Differential protein expression analysis was performed with the R package *limma*, which generated a linear model to estimate fold changes and SEs prior to empirical Bayes smoothing.[24] Finally, Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis was carried out with Metascape. Gene network analysis was conducted with GeneMANIA,[25] a plugin of the Cytoscape application. The critical hubs, highly connected to nodes in a module, were defined as the highest connectivity degrees.

P values were two-sided. All statistical analyses were performed with R version 3.5.1 (R Foundation), unless otherwise specified.

Results

After QC, 1,230 ($N_{\text{discovery}} = 613$ and $N_{\text{validation}} = 617$) patients with 12,806 CpG probes and 719 gene probes were included in the association analysis. The demographic and clinical information are described in e-Tables 2, 3.

For the main effect analysis of DNA methylation and gene expression, 23 CpG probes (e-Tables 4-6) and 13 gene probes (e-Tables 7, 8) were selected by ISIS LASSO, respectively. However, only cg19286631$_{TRIM27}$ was significantly associated with survival in both phases ($HR_{\text{discovery}} = 1.03$ [95% CI, 1.01-1.05], $P = 1.43 \times 10^{-2}$; $HR_{\text{validation}} = 1.03$ [95% CI, 1.01-1.06], $P = 1.13 \times 10^{-3}$) and remained significant in sensitivity analysis. Also, only one gene probe located in the *NDRG1* gene remained significant in the validation phase ($HR_{\text{discovery}} = 1.41$ [95% CI, 1.05-1.89], $P = 2.16 \times 10^{-2}$; $HR_{\text{validation}} = 1.12$ [95% CI, 1.01-1.42], $P = 4.33 \times 10^{-2}$) and sensitivity analysis.

For the G×G interaction analysis, we observed 2,495 and 40 G×G interactions from epigenetic and transcriptional analysis, respectively, in the discovery phase. Finally, 149 and 2 G×G interactions were retained in the validation phase that were also significant in the sensitivity analysis (e-Tables 9-13).

By forward stepwise regression analysis in the discovery phase, we observed one CpG probe with a main effect and 25 pairs of CpG probes with G×G interactions in the multibiomarker model (e-Table 14), which was used to calculate the epigenetic score (e-Table 15) ($HR_{\text{discovery}} = 2.71$ [95% CI, 2.41-3.05]; $P = 1.15 \times 10^{-61}$). One gene probe with a main effect and one pair of gene probes with a G×G interaction were retained in the multibiomarker model and used to calculate the transcriptional score ($HR_{\text{discovery}} = 2.44$ [95% CI, 1.78-3.35]; $P = 2.79 \times 10^{-8}$). The associations between survival and each of these scores were independently confirmed in the validation phase when adjusted for covariates (epigenetic score: $HR_{\text{validation}} = 2.72$ [95% CI, 2.31-3.20], $P = 6.06 \times 10^{-33}$; transcriptional score: $HR_{\text{validation}} = 2.64$ [95% CI, 1.73-4.04], $P = 7.51 \times 10^{-6}$; integrative score: $HR_{\text{validation}} = 2.72$ [95% CI, 2.32-3.18], $P = 5.68 \times 10^{-35}$; prognostic score: $HR_{\text{validation}} = 2.72$ [95% CI, 2.34-3.17], $P = 5.04 \times 10^{-38}$).

To evaluate the discriminative ability of these scores, samples in the validation phase were categorized into low-, medium-, and high-score groups based on the tertiles of epigenetic, transcriptional, integrative, and prognostic scores, respectively. Compared with the epigenetic low-score group, the medium- and high-score groups had 4.39- and 21.24-fold mortality risk, respectively ($HR_{\text{Medium vs Low}} = 4.39$ [95% CI, 2.42-7.99], $P = 1.22 \times 10^{-6}$; $HR_{\text{High vs Low}} = 21.24$ [95% CI, 11.23-40.17], $P = 5.67 \times 10^{-21}$) (Fig 2A). Patients with a high transcriptional score had significantly worse survival ($HR_{\text{Medium vs Low}} = 1.46$ [95% CI, 0.92-2.33], $P = 1.04 \times 10^{-1}$; $HR_{\text{High vs Low}} = 2.26$ [95% CI, 1.41-3.60], $P = 6.52 \times 10^{-4}$) (Fig 2B). The significant survival difference was enhanced among patients with different integrative scores ($HR_{\text{Medium vs Low}} = 4.32$ [95% CI, 2.39-7.83], $P =1.33 \times 10^{-6}$; $HR_{\text{High vs Low}} = 24.32$ [95% CI, 12.71-46.56], $P = 5.76 \times 10^{-22}$) (Fig 2C). Moreover, when combined with clinical information, including age, sex, study center, clinical stage, and smoking status, the prognostic score significantly discriminated NSCLC survival ($HR_{\text{Medium vs Low}} = 7.32$ [95% CI, 3.50-15.33], $P = 1.29 \times 10^{-7}$; $HR_{\text{High vs Low}} = 28.85$ [95% CI, 13.13-63.43], $P = 5.83 \times 10^{-17}$) (Fig 2D). The discriminative ability of the prognostic score is

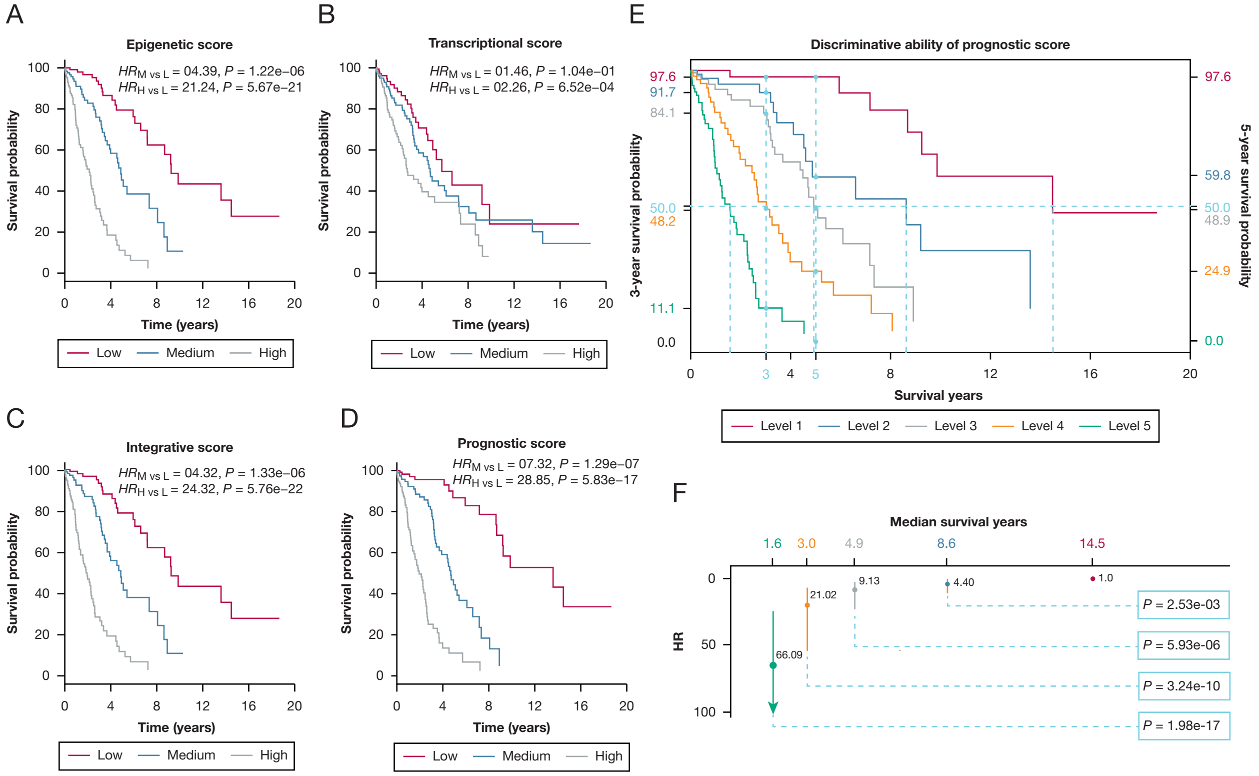


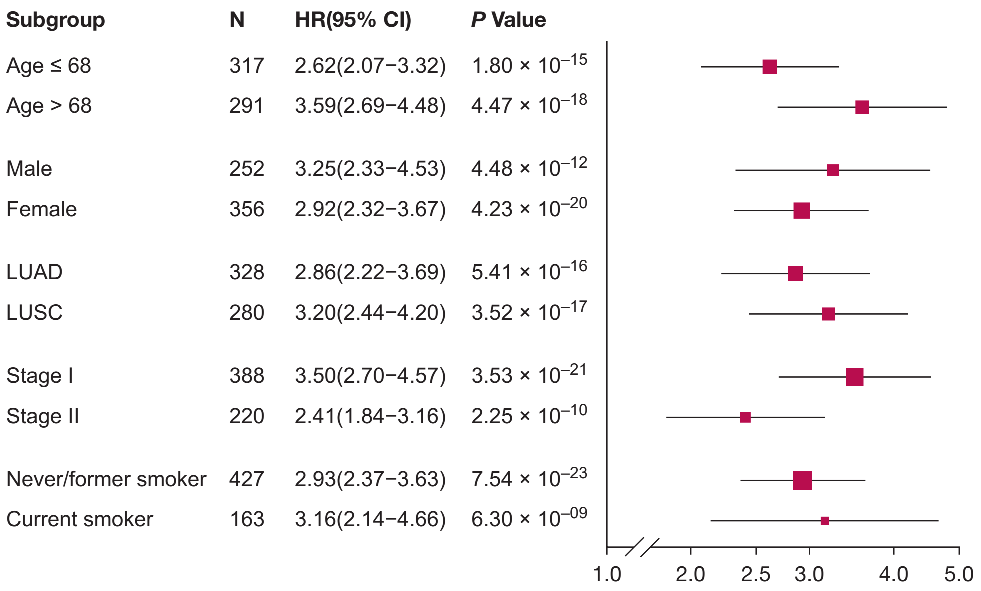
Figure 2 – *Estimated survival curves for patients grouped by various biomarker-based scores. A, Epigenetic score of DNA methylation. B, Transcriptional score of gene expression. C, Integrative score of DNA methylation and gene expression. D, Prognostic score of DNA methylation, gene expression, and clinical information. Patients were categorized into low-, medium-, and high-score groups by using the tertiles of each score as the cutoffs. E, Discriminative ability of the prognostic score. Results of 3- and 5-year survival rate, median survival time, and hazard ratio (HR) were compared across five groups, defined by using the quintiles of the prognostic score as the cutoffs. F, HR and* P *values were derived from the Cox proportional hazards model for patients with different quintile levels of the prognostic score.* $HR_{H\ vs\ L} = HR_{High\ vs\ Low}$; $HR_{M\ vs\ L} = HR_{Medium\ vs\ Low}$.

further illustrated by categorizing patients on the basis of the quintile level of the score. Figure 2E manifests an ordering relation: patients in higher-quintile groups had lower 3- and 5-year survival rates, as well as shorter median survival time. This indicates that patients with higher mortality risks can be detected by using our score system ($HR_{Level\ 5\ vs\ 1} = 66.09$ [95% CI, 25.13-173.80], $P = 1.98 \times 10^{-17}$; $HR_{Level\ 4\ vs\ 1} = 21.02$ [95% CI, 8.13-54.31], $P = 3.24 \times 10^{-10}$; $HR_{Level\ 3\ vs\ 1} = 9.13$ [95% CI, 3.51-23.78], $P = 5.93 \times 10^{-6}$; $HR_{Level\ 2\ vs\ 1} = 4.40$ [95% CI, 1.68-11.53], $P = 2.53 \times 10^{-3}$) (Fig 2F). The performance of the prognostic score was further confirmed in the analysis stratified by covariates (Fig 3).

We then independently validated the predictive ability of these biomarkers. The model with only clinical information, as aforementioned, had very limited prediction ability ($AUC_{3\ year} = 0.65$, $AUC_{5\ year} = 0.66$). However, by adding biomarkers with either main effects or G×G interactions, the AUCs significantly increased by 35.38% (95% CI, 27.09%-44.17%; $P = 5.10 \times 10^{-17}$) and 34.85% (95% CI, 26.33%-41.87%; $P = 2.52 \times 10^{-18}$) for 3- and 5-year survival, respectively, and exhibited a superior predictive ability for NSCLC survival ($AUC_{3\ year} = 0.88$ [95% CI, 0.83-0.93]; $AUC_{5\ year} = 0.89$ [95% CI, 0.83-0.93]) (Fig 4). G×G interactions contributed an additional 65.2% for 3-year and 91.3% for the 5-year prediction accuracy increase.

In the sensitivity analysis, we reanalyzed the stepwise regression using two different thresholds ($P = .10$ and .15) and found that the majority of the selected biomarkers were the same as those in the original regression model (e-Table 16). We then recalculated these scores, retested their associations with NSCLC survival, and obtained similar results (e-Table 17). Meanwhile, the AUCs of our prognostic model using different thresholds were comparable: $0.88_{P\ =\ .05}$ vs $0.85_{P\ =\ .10}$ vs $0.86_{P\ =\ .15}$ for 3-year survival; $0.89_{P\ =\ .05}$ vs $0.83_{P\ =\ .10}$ vs $0.86_{P\ =\ .15}$ for 5-year survival (e-Figs 2 and 3).

Moreover, we found that the effects of these four scores did not differ significantly between patients with LUAD and patients with LUSC ($P_{Epigenetic\ score} = .6572$;



Subgroup	N	HR(95% CI)	*P* Value
Age ≤ 68	317	2.62(2.07−3.32)	1.80×10^{-15}
Age > 68	291	3.59(2.69−4.48)	4.47×10^{-18}
Male	252	3.25(2.33−4.53)	4.48×10^{-12}
Female	356	2.92(2.32−3.67)	4.23×10^{-20}
LUAD	328	2.86(2.22−3.69)	5.41×10^{-16}
LUSC	280	3.20(2.44−4.20)	3.52×10^{-17}
Stage I	388	3.50(2.70−4.57)	3.53×10^{-21}
Stage II	220	2.41(1.84−3.16)	2.25×10^{-10}
Never/former smoker	427	2.93(2.37−3.63)	7.54×10^{-23}
Current smoker	163	3.16(2.14−4.66)	6.30×10^{-09}

Figure 3 – *Forest plots of results from stratification analysis of prognostic score. HR with 95% CI of the prognostic score on non-small cell lung cancer survival in various subgroups is stratified by clinical characteristics. LUAD = lung adenocarcinoma; LUSC = lung squamous cell carcinoma. See Figure 2 legend for expansion of other abbreviation.*

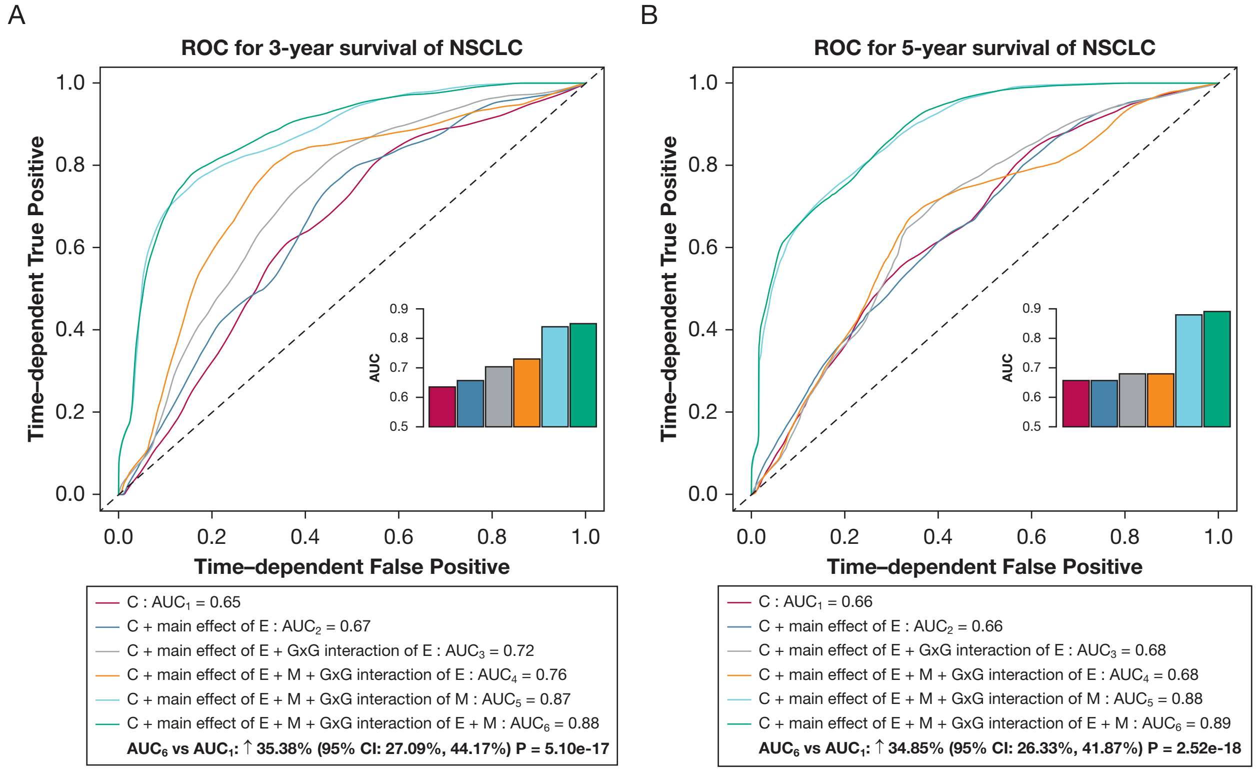


Figure 4 – *Receiver operating characteristic curves for various predictive models using the clinical information (C), the main and interaction effects of DNA methylation (M), and gene expression (E). A, Three-year survival prediction. B, Five-year survival prediction. The AUC increase (%) was evaluated by comparing the model with that with only the clinical information.* P *values and 95% CIs were calculated by using 1,000 bootstrap samples. AUC = area under the receiver operating characteristic curve; ROC = receiver operating characteristic. See Figure 1 legend for expansion of other abbreviations.*

$P_{\text{Transcriptional score}} = .1823$; $P_{\text{Integrative score}} = .5532$; $P_{\text{Prognostic score}} = .9653$) (e-Table 18). Our prognostic model retained similar prediction ability in both the LUAD ($\text{AUC}_{3\text{ year}} = 0.91$, $\text{AUC}_{5\text{ year}} = 0.89$, and $C_{\text{index}} = 0.82$; 95% CI, 0.76-0.87) and LUSC ($\text{AUC}_{3\text{ year}} = 0.85$, $\text{AUC}_{5\text{ year}} = 0.87$, and $C_{\text{index}} = 0.82$; 95% CI, 0.76-0.88) populations, indicating the usefulness of the selected biomarkers and their interactions in predicting the outcomes for patients with LUAD and patients with LUSC (e-Fig 4).

To facilitate application of our prognostic prediction model, we combined clinical information and scores of biomarkers and developed a nomogram, which estimated well a patient's 3- or 5-year survival (e-Fig 5). The C_{index} of the prognostic score indicated acceptable prediction accuracy ($C_{\text{index}} = 0.82$; 95% CI, 0.78-0.86) in an independent TCGA population. The calibration plots also showed good accordance between observed and predicted survival time (e-Fig 6).

In protein analysis, three of the four genes mapped in TCGA had significant correlation between gene expression and protein expression (e-Table 19). Most (77%) of the 47 genes mapped in CPTAC were differentially expressed between tumor and normal tissue, with statistical significance (e-Fig 7). In addition, one gene with main effect and four pairs of genes with G×G interaction had a significant effect on LUAD survival (e-Table 20). Among 49 genes identified in epigenetic analysis, five genes (*FOXP1*, *AFF3*, *BCL6*, *MAPK1*, and *STAT3*) were identified as hub genes with the highest connectivity degrees, greater than 25 (Fig 5). These 49 genes were enriched in cancer-related pathways including the non-small cell lung cancer pathway (e-Table 21).

Discussion

An accurate prognostic predictive model may aid physicians in making clinical decisions or guiding adjuvant therapy, especially for the vulnerable patients with high mortality risk. Although subject and tumor characteristics have been commonly used as valid predictors, increased evidence has indicated that molecular biomarkers may provide early warning signals. This is because tumor cells may metastasize even when the tumor size is undetectable (< 0.01 cm^3) and aberrations of biomarkers occur.[26] Thus, there is added value when a prognostic predictive model incorporates both genetic and nongenetic factors, whose effects can be captured using approaches that are both biologically stable and technically reproducible.

We conducted a two-stage integrative study of DNA methylation and gene expression data from multiple centers to propose a prognostic scoring method incorporating transomics biomarkers with main effects and G×G interactions. The prognostic score, which was validated in an independent population, effectively discriminates survival outcomes for patients with early-stage NSCLC and significantly improves prediction accuracy for their prognosis.

G×G interactions are of interest because they provide important clues regarding the biologic mechanisms of complex diseases.[27] It was suggested in previous studies that identification of G×G interactions would improve the predictive accuracy of statistical models.[28,29] However, interactions might not dramatically improve prediction if their effects are weak or there are few significant interactions, but might optimize statistical modeling.[30] Besides prediction, G×G interactions could increase the power to detect associations and then be leveraged for the identification of new biomarkers.[27] Our results showed that biomarkers with G×G interactions significantly and predominantly improved the prognostic prediction accuracy of early-stage NSCLC, which might be due to increased power.

To evaluate the prediction accuracy of our model, we conducted a literature search to compare our studies with others. The details of the literature screening process are summarized in e-Figure 8. The prediction accuracy of our model is superior (e-Table 22), as the one study with the best AUC (0.80) had a very small sample size. Another study with the largest sample size, without independent validation, had unsatisfactory prediction capacity ($C_{\text{index}} = 0.64$). Our study has a relatively large sample size and provides a satisfactory prediction model that performed well in an independent population regardless of AUC ($\text{AUC}_{3\text{ year}} = 0.88$ and $\text{AUC}_{5\text{ year}} = 0.89$) and C_{index} (0.82).

Among the genes identified in transcriptional analysis, *NDRG1*[31] *and RHOA*[32] have been reported to be associated with lung cancer. In this study, among 49 genes identified in epigenetic analysis, five (*AFF3*, *MAPK1*, *STAT3*, *FOXP1*, and *BCL6*) were identified as hub genes in the gene network. *AFF3* is associated with NSCLC prognosis.[33] *MAPK1* promotes NSCLC cell survival and is a therapeutic target for NSCLC chemotherapeutic resistance.[34] *STAT3*, one of the three major downstream pathways activated by EGFR phosphorylation,[35] is persistently activated in 22% to 65% of NSCLC.[36-38] It is a strong predictor of poor

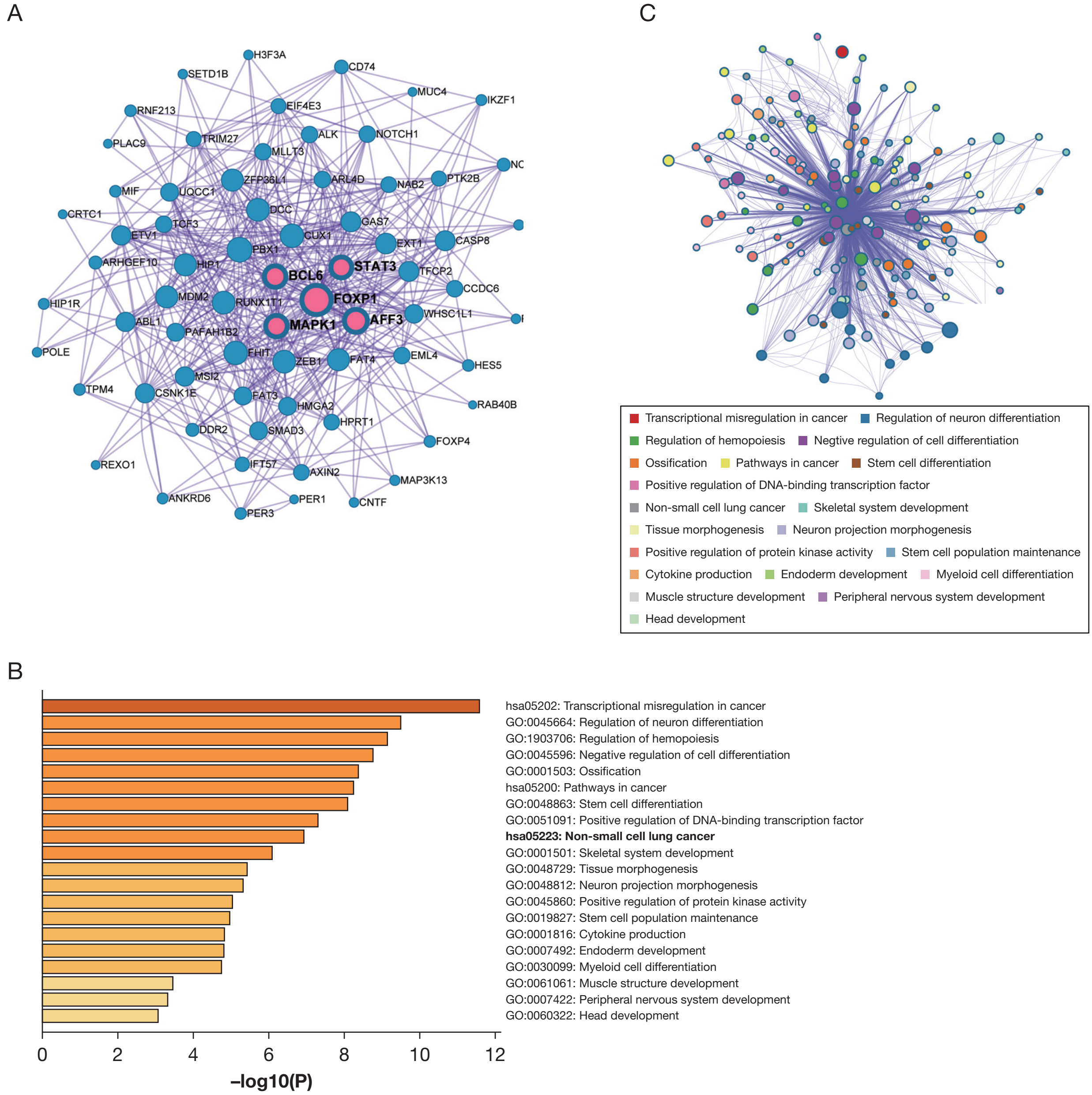


Figure 5 – *Gene network and gene enrichment analysis of 49 genes to which 25 pairs of CpG probes with interaction and one CpG probe with main effect are mapped. A, The gene network plot constructed by GeneMANIA. Central nodes with boldface outline represent hub genes, and the size represents the connectivity degree of each node. B, Barplot of gene pathways enriched with significant genes, and colored by* P *values. C, The pathway network plot of these pathways enriched with significant genes. Significant pathways with a similarity > 0.3 are connected by edges. Each node represents an enriched term and is colored by its cluster identification. The size of the node represents the number of genes in the pathway. The edge represents potential biologic relationships between two pathways. GO = Gene Ontology.*

NSCLC prognosis and related to cisplatin resistance in NSCLC cells.[39-41] *FOXP1* is an independent factor for predicting poor NSCLC prognosis.[42] *BCL6* could inhibit cell apoptosis in lung cancer[43] and plays a role in sustaining NSCLC genomic instability.[44]

In enrichment pathway analysis, 49 genes were significantly enriched in pathways or processes that are cancer related. Notably, the identified genes were also enriched in the KEGG non-small cell lung cancer pathway (hsa05223). The hub genes *MAPK1* and *STAT3* in the network were also involved in this pathway. The results indicated that, after functional confirmation, the identified CpG probes are potential epigenetic targets for NSCLC chemotherapy.

Our study has some strengths, as follows: (1) Most studies focus only on main effects of biomarkers, ignoring their G×G interactions that account for missing heritability of complex diseases like NSCLC.[15] Also, most studies focus on single omics data testing prognostic biomarkers.[10-13] Taking advantage of epigenomic and transcriptomic data and considering both G×G interactions and main effects,[45] we built up transomics prognostic scores, which could improve prognostic value; (2) to identify reliable prognostic biomarkers for the prediction of early-stage NSCLC overall survival, we used stringent statistical criteria. In the main effect analysis, candidate biomarkers, with effect sizes larger than a data-driven threshold in ISIS LASSO, must reach statistical significance to stay in the Cox regression model. For the G×G interaction analysis, we applied the most conservative Bonferroni correction to control for false positives. In addition, significant biomarkers observed in the discovery phase must be further validated in an independent population. However, one consequence was that only a few biomarkers were identified because of the limited sample size of gene expression data, which therefore contributed a small proportion of improved accuracy of our model; (3) we used ISIS LASSO and stepwise regression to screen biomarkers with main effect and interactions, respectively, and built multibiomarker models. These coefficients, used as weights to define scores, were derived from multibiomarker models rather than single-biomarker models. Single-biomarker models might result in biased estimates of effect sizes, whereas multibiomarker models are more beneficial to clarify the complex association and could improve prediction accuracy[46,47]; (4) the prediction accuracy of our prognostic model was robust toward different selection thresholds in stepwise regression as well as stratification by histology types; and (5) the genes we identified as enriched in the non-small cell lung cancer pathway and most of the hub genes have been reported to be associated with NSCLC, indicating the reliability of our prognostic biomarkers.

We also acknowledge some limitations of our study, as follows: (1) We focused only on pan-cancer genes, whereas most dysregulated genes represent the consequences rather than the causes of neoplastic process.[48] Moreover, few powerful statistical methods or excellent computer hardware can finish G×G interactions for time-to-even data on genome-wide scales within weeks. We exhaustively tested all pairs of pan-cancer-related genes; (2) limited clinical information was available for several cohorts that were initiated decades ago. However, in our study, a few easily accessible clinical predictors and dozens of biomarkers exhibited considerable accuracy, which indicated potentiality for real-world application; (3) the event rate of survival time for TCGA population is relatively low (23%), which considerably reduced the statistical power. However, through a conservative two-stage strategy this study showed the robustness of our findings; (4) our prognostic prediction model predicts survival outcome and distinguishes subgroups of patients with high mortality risk accurately, which provides a potential opportunity for the delivery of personalized medicine and interventions tailored to each individual's level of risk. However, it requires information on 54 biomarkers, which might restrict its clinical translatability without testing of specimens. Nevertheless, the history of cancer omics testing has taught us that, as technology improves and costs fall, the trend is toward more convenient and comprehensive approaches to quickly capture biomarker information.[49] In the coming years, advances in technology will facilitate our model's usefulness through a customized biochip to enable widespread clinical application and maximize the benefit to patients; and (5) further studies with a large-scale population and extension of other ethics are warranted to confirm the results of our association study and verify the underline biologic mechanisms of the genes and their interactions. Results of protein analysis in public resources and in gene network and enrichment analyses might provide insight into the functional mechanisms.

Conclusion

The prognostic score incorporating transomics biomarkers with both main effects and G×G interactions significantly improves prognostic prediction accuracy for early-stage NSCLC survival.

Acknowledgments

Author contributions: R. Z. takes responsibility for the content of the manuscript, including the data and analysis. R. Z., C. C., X. D., Y. W., F. C., and D. C. C contributed to the study design. R. Z., L. S., A. S., T. F., M. M. J., A. K., M. P., J. S., Å. H., M. E., and D. C. C. contributed to data collection. R. Z., C. C., X. D., S. S., L. Lai, Y. L., Y. W., F. C., and D. C. C. performed statistical analysis and interpretation and drafted the manuscript. J. H., D. Y., L. Lin, Y. Z., H. H., J. C., L. W., X. C., Y. G., W. D., L. Liu, L. S., and R. W. revised the manuscript. All authors contributed to critical revision of the manuscript and approved its final version. Financial support and study supervision were provided by F. C. and D. C. C.

Financial/nonfinancial disclosures: None declared.

Role of sponsors: The sponsor had no role in the design of the study, the collection and analysis of the data, or the preparation of the manuscript.

Availability of data and materials: COSMIC (https://cancer.sanger.ac.uk/cosmic/), which provided the symbol of 719 pan-cancer-related genes; TCGA (https://tcga-data.nci.nih.gov), now hosted at GDC (https://portal.gdc.cancer.gov/); GEO (https://www.ncbi.nlm.nih.gov/geo/), for which detailed information is provided in e-Table 1; CPTAC (https://cptac-data-portal.georgetown.edu/), which provided protein data; Metascape (http://metascape.org), which was used to perform pathway enrichment analysis.

Publisher's note: Springer Nature remains neutral regarding jurisdictional claims in published maps and institutional affiliations.

Other contributions: The authors thank TCGA, GEO, and CPTAC for contributing clinical, DNA methylation, RNA sequencing, and protein data; COSMIC for providing pan-cancer-related genes; as well as all of the study subjects who participated in this study.

Additional information: The e-Appendix, e-Figures, and e-Tables can be found in the Supplemental Materials section of the online article.

References

1. Bray F, Ferlay J, Soerjomataram I, et al. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. 2018;68:394-424.
2. Hirsch FR, Scagliotti GV, Mulshine JL, et al. Lung cancer: current therapies and new targeted treatments. *Lancet*. 2017;389(10066):299-311.
3. Tang S, Pan Y, Wang Y, et al. Genome-wide association study of survival in early-stage non-small cell lung cancer. *Ann Surg Oncol*. 2015;22(2):630-635.
4. Egger G, Liang G, Aparicio A, et al. Epigenetics in human disease and prospects for epigenetic therapy. *Nature*. 2004;429(6990):457-463.
5. Feinberg AP, Tycko B. The history of cancer epigenetics. *Nat Rev Cancer*. 2004;4(2):143-153.
6. Shen S, Zhang R, Guo Y, et al. A multi-omic study reveals *BTG2* as a reliable prognostic marker for early-stage non-small cell lung cancer. *Mol Oncol*. 2018;12: 913-924.
7. Wei Y, Liang J, Zhang R, et al. Epigenetic modifications in *KDM* lysine demethylases associate with survival of early-stage NSCLC. *Clin Epigenetics*. 2018;10(1):41.
8. Zhang R, Lai L, He J, et al. *EGLN2* DNA methylation and expression interact with *HIF1A* to affect survival of early-stage NSCLC. *Epigenetics*. 2019;14(2):118-129.
9. Lin Z, Hui L, Yufei H, et al. Cancer progression prediction using Gene Interaction Regularized Elastic Net. *IEEE/ACM Trans Comput Biol Bioinform*. 2017;14(1):145-154.
10. Sandoval J, Mendez-Gonzalez J, Nadal E, et al. A prognostic DNA methylation signature for stage I non-small-cell lung cancer. *J Clin Oncol*. 2013;31(32):4140-4147.
11. Shedden K, Taylor JM, Enkemann SA, et al. Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study. *Nat Med*. 2008;14(8):822-827.
12. Tan X, Qin W, Zhang L, et al. A 5-microRNA signature for lung squamous cell carcinoma diagnosis and hsa-miR-31 for prognosis. *Clin Cancer Res*. 2011;17(21):6802-6811.
13. Zhou M, Guo M, He D, et al. A potential signature of eight long non-coding RNAs predicts survival in patients with non-small cell lung cancer. *J Transl Med*. 2015;13(1):231.
14. Zhao Q, Shi X, Xie Y, et al. Combining multidimensional genomic measurements for predicting cancer prognosis: observations from TCGA. *Brief Bioinform*. 2015;16(2):291-303.
15. Guo Y, Zhang R, Shen S, et al. DNA Methylation of *LRRC3B*: a biomarker for survival of early-stage non-small cell lung cancer patients. *Cancer Epidemiol Biomarkers Prev*. 2018;27(12):1527-1535.
16. Bjaanæs MM, Fleischer T, Halvorsen AR, et al. Genome-wide DNA methylation analyses in lung adenocarcinomas: association with EGFR, KRAS and TP53 mutation status, gene expression and prognosis. *Mol Oncol*. 2016;10(2):330-343.
17. Karlsson A, Jonsson M, Lauss M, et al. Genome-wide DNA methylation analysis of lung carcinoma reveals one neuroendocrine and four adenocarcinoma epitypes associated with patient outcome. *Clin Cancer Res*. 2014;20(23):6127-6140.
18. Chen Y-a, Lemire M, Choufani S, et al. Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. *Epigenetics*. 2013;8(2):203-209.
19. Marabita F, Almgren M, Lindholm ME, et al. An evaluation of analysis pipelines for DNA methylation profiling using the Illumina HumanMethylation450 BeadChip platform. *Epigenetics*. 2013;8(3): 333-346.
20. Zhang R, Lai L, Dong X, et al. *SIPA1L3* methylation modifies the benefit of smoking cessation on lung adenocarcinoma survival: an epigenomic-smoking interaction analysis. *Mol Oncol*. 2019;13(5):1235-1248.
21. Saldana DF, Feng Y. SIS: an R package for sure independence screening in ultrahigh dimensional statistical models. *J Stat Software*. 2018;83(2):1-25.
22. Heagerty PJ, Lumley T, Pepe MS. Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics*. 2000;56(2):337-344.
23. Brentnall AR, Cuzick J. Use of the concordance index for predictors of censored survival data. *Stat Methods Med Res*. 2018;27(8):2359-2373.
24. Ritchie ME, Phipson B, Wu D, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*. 2015;43(7):e47.
25. Warde-Farley D, Donaldson SL, Comes O, et al. The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res*. 2010;38(Web Server issue):W214-W220.
26. Hu Z, Ding J, Ma Z, et al. Quantitative evidence for early metastatic seeding in colorectal cancer. *Nat Genet*. 2019;51(7): 1113-1122.
27. Cordell HJ. Detecting gene-gene interactions that underlie human diseases. *Nat Rev Genet*. 2009;10(6):392-404.
28. Pharoah PD, Antoniou AC, Easton DF, et al. Polygenes, risk prediction, and targeted prevention of breast cancer. *N Engl J Med*. 2008;358(26):2796-2803.
29. Khoury MJ, Yang Q, Gwinn M, Little J, Dana Flanders W. An epidemiologic assessment of genomic profiling for measuring susceptibility to common diseases and targeting interventions. *Genet Med*. 2004;6(1):38-47.
30. Aschard H, Chen J, Cornelis MC, Chibnik LB, Karlson EW, Kraft P. Inclusion of gene-gene and gene-environment interactions unlikely to dramatically improve risk prediction for complex diseases. *Am J Hum Genet*. 2012;90(6):962-972.
31. Azuma K, Kawahara A, Hattori S, et al. NDRG1/Cap43/Drg-1 may predict tumor angiogenesis and poor outcome in patients with lung cancer. *J Thorac Oncol*. 2012;7(5):779-789.
32. Konstantinidou G, Ramadori G, Torti F, et al. RHOA-FAK is a required signaling axis for the maintenance of KRAS-driven lung adenocarcinomas. *Cancer Discov*. 2013;3(4):444-457.

33. Zhang DL, Qu LW, Ma L, et al. Genome-wide identification of transcription factors that are critical to non-small cell lung cancer. *Cancer Lett*. 2018;434:132-143.

34. Vicent S, Lopez-Picazo JM, Toledo G, et al. ERK1/2 is activated in non-small-cell lung cancer and associated with advanced tumours. *Br J Cancer*. 2004;90(5):1047-1052.

35. Mitsudomi T, Yatabe Y. Mutations of the epidermal growth factor receptor gene and related genes as determinants of epidermal growth factor receptor tyrosine kinase inhibitors sensitivity in lung cancer. *Cancer Sci*. 2007;98(12):1817-1824.

36. Zimmer S, Kahl P, Buhl TM, et al. Epidermal growth factor receptor mutations in non-small cell lung cancer influence downstream Akt, MAPK and Stat3 signaling. *J Cancer Res Clin Oncol*. 2009;135(5):723-730.

37. Looyenga BD, Hutchings D, Cherni I, et al. STAT3 is activated by JAK2 independent of key oncogenic driver mutations in non-small cell lung carcinoma. *PLoS One*. 2012;7(2):e30820.

38. Jiang R, Jin Z, Liu Z, et al. Correlation of activated STAT3 expression with clinicopathologic features in lung adenocarcinoma and squamous cell carcinoma. *Mol Diagn Ther*. 2011;15(6):347-352.

39. Barre B, Vigneron A, Perkins N, et al. The *STAT3* oncogene as a predictive marker of drug resistance. *Trends Mol Med*. 2007;13(1):4-11.

40. Ikuta K, Takemura K, Kihara M, et al. Overexpression of constitutive signal transducer and activator of transcription 3 mRNA in cisplatin-resistant human non-small cell lung cancer cells. *Oncol Rep*. 2005;13(2):217-222.

41. Harada D, Takigawa N, Kiura K. The role of STAT3 in non-small cell lung cancer. *Cancers (Basel)*. 2014;6(2):708-722.

42. Feng J, Zhang X, Zhu H, et al. High expression of FoxP1 is associated with improved survival in patients with non-small cell lung cancer. *Am J Clin Pathol*. 2012;138(2):230-235.

43. Sun C, Li S, Yang C, et al. MicroRNA-187-3p mitigates non-small cell lung cancer (NSCLC) development through down-regulation of BCL6. *Biochem Biophys Res Commun*. 2016;471(1):82-88.

44. Marullo R, Ahn H, Cardenas M, Melnick A, Xue F, Cerchietti L. The transcription factor BCL6 is a rational target in non-small cell lung cancer (NSCLC) [abstract]. *Cancer Res*. 2016;76(14 suppl):Abstract 1271.

45. Ng SW, Mitchell A, Kennedy JA, et al. A 17-gene stemness score for rapid determination of risk in acute leukaemia. *Nature*. 2016;540(7633):433-437.

46. Kang J, Kugathasan S, Georges M, et al. Improved risk prediction for Crohn's disease with a multi-locus approach. *Hum Mol Genet*. 2011;20(12):2435-2442.

47. Segura V, Vilhjalmsson BJ, Platt A, et al. An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nat Genet*. 2012;44(7):825-830.

48. Vogelstein B, Papadopoulos N, Velculescu VE, et al. Cancer genome landscapes. *Science*. 2013;339(6127):1546-1558.

49. Kuo FC, Mar BG, Lindsley RC, et al. The relative utilities of genome-wide, gene panel, and individual gene sequencing in clinical practice. *Blood*. 2017;130(4):433-439.